


FUJITSU Software PRIMECLUSTER

A decorative horizontal band with a red-to-dark-red gradient, featuring abstract, glowing white and red lines that swirl and intersect, creating a sense of motion and technology.

Cluster Foundation Configuration and Administration Guide 4.6

Oracle Solaris

J2S2-1713-01ENZ0(00)
October 2021

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 4.6 and perform cluster system installation and operation management.

Configuration of This Documentation

This manual is organized as follows:

Chapter Title	Content
Chapter 1 Cluster Foundation	Describes the administration and configuration of the Cluster Foundation (CF).
Chapter 2 CF Registry and Integrity Monitor	Discusses the purpose and physical characteristics of the CF synchronized registry, and it discusses the purpose and implementation of Cluster Integrity Monitor (CIM).
Chapter 3 Cluster resource management	Discusses the database which is a synchronized clusterwide database holding information to multiple PRIMECLUSTER products.
Chapter 4 GUI administration	Describes the administrative features in the CF portion of the Cluster Admin graphical user interface (GUI).
Chapter 5 LEFTCLUSTER state	Discusses the LEFTCLUSTER state, describes this state in relation to the other states, and discusses the different ways a LEFTCLUSTER state is caused.
Chapter 6 CF topology table	Discusses the CF topology table as it relates to the CF portion of the Cluster Admin GUI.
Chapter 7 Shutdown Facility	Describes the components and advantages of PRIMECLUSTER SF and provides administrative information.
Chapter 8 CF over IP	Provides the overview of CF over IP and how CF over IP is configured.
Chapter 9 Diagnostics and troubleshooting	Provides help for troubleshooting and problem resolution for PRIMECLUSTER Cluster Foundation.

Related documentation

Refer to the following manuals as necessary when setting up the cluster:

- PRIMECLUSTER Concepts Guide
- PRIMECLUSTER Installation and Administration Guide
- PRIMECLUSTER Web-Based Admin View Operation Guide
- PRIMECLUSTER Reliant Monitor Services (RMS) with Wizard Tools Configuration and Administration 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 Global Link Services Configuration and Administration Guide: Redundant Line Control Function for Virtual NIC Mode
- PRIMECLUSTER Global Link Services Configuration and Administration Guide: Multipath Function
- PRIMECLUSTER DR/PCI Hot Plug User's Guide
- PRIMECLUSTER Messages
- FJQSS (Information Collection Tool) User's Guide

 Note

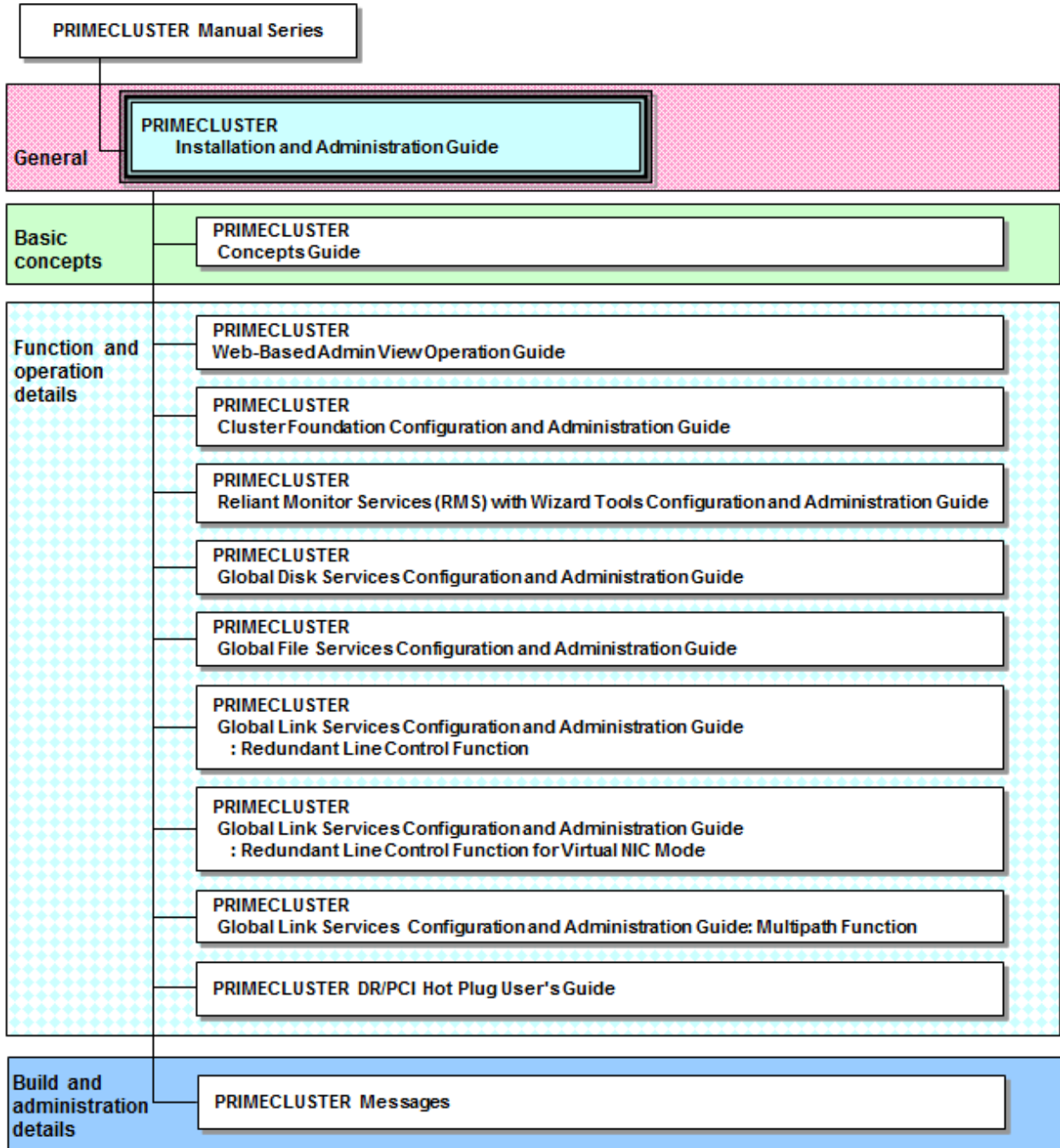
The PRIMECLUSTER documentation includes the following documentation in addition to those listed above:

- 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 the "DVD" for 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.

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 *fuji2#* would mean that the command was run as user root on the node named fuji2.

Manual page section numbers

In manuals, helps, and messages of PRIMECLUSTER, a section number in a manual page is shown in parentheses after a command name or a file name. Example: `cp(1)`

For Oracle Solaris 11.4 or later, replace the section numbers as follows:

- "(1M)" to "(8)"
- "(4)" to "(5)"
- "(5)" to "(7)"
- "(7)" to "(4)"

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	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> , <Italic>	Variables that you must replace with an actual entered value.
<Constant Width>	Variables that you must replace with an actual displayed value.
Bold	Items in a command line that you must type exactly as shown.
"Constant Width"	The title, documentation, screen, and etc of lookup destination.
[Constant Width]	Tool bar name, menu name, command name, button name, and icon names
Constant Width	Computer output and program listings; commands, file names, manual page names and other literal programming elements in the main body of text.

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.
()	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 the 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.

If "Solaris X" is indicated in the reference manual name of the Oracle Solaris manual, replace "Solaris X" with "Oracle Solaris 10 (Solaris 10)" or "Oracle Solaris 11 (Solaris 11)."

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

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

EMC, PowerPath and Symmetrix are registered trademarks of EMC Corporation.

TimeFinder and SRDF are registered trademarks of EMC Corporation.

Fujitsu SPARC M12 is sold as SPARC M12 by Fujitsu in Japan.

Fujitsu SPARC M12 and SPARC M12 are identical products.

Fujitsu M10 is sold as SPARC M10 by Fujitsu in Japan.

Fujitsu M10 and SPARC M10 are identical products.

Other product names are product names, trademarks, or registered trademarks of these 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

October 2021, First edition

Copyright notice

All Rights Reserved, Copyright (C) FUJITSU LIMITED 2021.

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.....	4
1.1.3 CF security.....	5
1.1.4 Example of creating a cluster.....	5
1.1.5 Adding a new node to CF.....	24
1.1.6 Example of CF setting by using CLI.....	24
1.2 CIP configuration file.....	26
1.3 Cluster Configuration Backup and Restore (CCBR).....	26
Chapter 2 CF Registry and Integrity Monitor.....	31
2.1 CF Registry (CFREG).....	31
2.2 Cluster Integrity Monitor (CIM).....	31
2.2.1 Configuring CIM	31
2.2.2 Query of the quorum state	32
2.2.3 Reconfiguring quorum	32
Chapter 3 Cluster resource management.....	34
3.1 Overview.....	34
3.2 Kernel parameters for Resource Database.....	34
3.3 Resource Database configuration.....	36
3.4 Registering hardware information.....	38
3.4.1 Setup exclusive device list.....	38
3.4.2 Exclusive device list for EMC Symmetrix.....	38
3.4.2.1 emcpower Devices and native Devices.....	39
3.4.2.2 BCV, R2, GateKeeper, CKD.....	39
3.4.2.3 VCMDB.....	39
3.4.2.4 Simplified setup for exclusive device list - clmakediskinfo, clmkdiskinfo.....	39
3.4.3 Automatic resource registration.....	40
3.4.4 Configuring the Resource Database by using the CRM main window.....	41
3.5 Startup synchronization.....	41
3.5.1 Startup synchronization and the new node.....	42
3.6 Adding a new node.....	42
3.6.1 Backing up the Resource Database.....	43
3.6.2 Reconfiguring the Resource Database.....	44
3.6.3 Configuring the Resource Database on the new node.....	45
3.6.4 Adjusting StartingWaitTime.....	45
3.6.5 Restoring the Resource Database.....	45
Chapter 4 GUI administration.....	47
4.1 Overview.....	47
4.2 Starting Cluster Admin GUI and logging in.....	47
4.3 Main CF table.....	49
4.4 CF route tracking.....	50
4.5 Node details.....	52
4.6 Displaying the topology table.....	53
4.7 Starting and stopping CF.....	56
4.7.1 Starting CF.....	57
4.7.2 Stopping CF.....	58
4.8 Marking nodes DOWN.....	59
4.9 Using PRIMECLUSTER log viewer.....	59
4.9.1 Search based on time filter	60
4.9.2 Search based on keyword	61
4.9.3 Search based on severity levels	62
4.10 Displaying statistics.....	63

4.11 Heartbeat monitor.....	65
4.12 Adding and removing a node from CIM.....	67
4.13 Unconfigure CF.....	68
4.14 CIM Override.....	69
Chapter 5 LEFTCLUSTER state.....	71
5.1 Description of the LEFTCLUSTER state.....	71
5.2 Recovering from LEFTCLUSTER.....	72
5.2.1 Caused by a panic/hung node.....	73
5.2.2 Caused by staying in the kernel debugger too long.....	73
5.2.3 Caused by a cluster partition.....	73
5.2.4 Caused by reboot.....	74
Chapter 6 CF topology table.....	75
6.1 Basic layout.....	75
6.2 Selecting devices.....	76
6.3 Examples.....	77
Chapter 7 Shutdown Facility.....	79
7.1 Overview.....	79
7.2 Configuring SF.....	79
7.2.1 Setting procedure before configuring SF.....	79
7.2.2 Configuration file of SF.....	80
7.3 Available SAs.....	80
7.3.1 RCI.....	81
7.3.2 XSCF.....	82
7.3.3 XSCF SNMP.....	84
7.3.4 ALOM.....	85
7.3.5 ILOM.....	85
7.3.6 KZONE.....	86
7.3.7 ICMP.....	87
7.4 SF split-brain handling.....	88
7.4.1 Administrative LAN.....	88
7.4.2 SF split-brain handling.....	88
7.4.3 Runtime processing.....	89
7.4.4 Configuration notes.....	89
7.5 Configuring the Shutdown Facility.....	90
7.6 SF administration.....	90
7.6.1 Starting and stopping SF.....	90
7.7 Logging.....	91
Chapter 8 CF over IP.....	92
8.1 Overview.....	92
8.2 Configuring CF over IP.....	93
Chapter 9 Diagnostics and troubleshooting.....	95
9.1 Beginning the process.....	95
9.2 Symptoms and solutions.....	97
9.2.1 Join-related problems.....	97
9.3 Collecting troubleshooting information.....	103
Glossary.....	104
Index.....	114

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 nodes in the cluster, and you can use the Cluster Admin SF Wizard to configure SF.

A CF configuration consists of the following main attributes:

- Cluster name
The name of cluster systems. This must be 31 characters or less per name and each character comes from the set of printable ASCII characters, excluding white space, newline, and tab characters. The cluster name is always mapped to upper case.
- Interconnect
Set of interfaces on each node in the cluster used for CF networking-For example, the interface of an IP address on the local node can be an Ethernet device.
- CF node name
By default, in Cluster Admin, the CF node name is the same as the Web-Based Admin View name; however, you can use the CF Wizard to change the name. Use up to 11 lower-case characters and symbols ("- and "_") for each node name. For the first letter of the CF node name, set a lowercase character.

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. There are a number of special requirements that these interconnects must meet if they are to be used for CF:

1. 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.

2. The interconnects should not be used on any network that might experience network outages for 5 seconds or more. A network outage of 10 seconds will, by default, cause a route to be marked as DOWN. (The state becomes DOWN if it is confirmed by the `cftool -r` command.) `cfset(1M)` can be used to change the 10 second default. See "1.1.2 `cfset`."

Since CF automatically attempts to activate interconnects, the problem with "split-brain" only occurs if all interconnects experience a 10-second outage simultaneously. Nevertheless, CF requires highly reliable interconnects.

CF can also be run over IP. Any IP interface on the node can be chosen as an IP device, and CF will treat this device as an interconnect in the same way as an Ethernet device. This is called the IP interconnect. However, all the IP addresses for all the cluster nodes on that interconnect must be on the same IP subnetwork, and their IP broadcast addresses must be the same.

The IP interfaces used by CF must be completely configured (the IP address must be assigned and activated) by the system administrator before they are used by CF so that communication can be performed by using the IP interfaces. You can run CF over both Ethernet devices and IP devices.

Higher level services, such as RMS, SF, Global File Services (hereinafter GFS), and so forth, will not notice any difference when CF is run over IP.

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 will need to bring down CF on each node to do the reconfiguration. Bringing down CF requires that higher level services, like RMS, SF and applications, be stopped on that node, so the reconfiguration process is neither trivial nor unobtrusive.

Note

To secure the reliability of the cluster system, the cluster interconnect redundancy is recommended.

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.

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 the CIP name, 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 nodenames and CIP names. In general, only one CIP interface per node is needed to be configured.

Note

A proper CIP configuration uses /etc/hosts to store CIP names. You should make sure that /etc/nsswitch.conf(4) is properly set up to use files criteria first in looking up its nodes.

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

1. CF/CIP, Web-Based Admin View, and Cluster Admin should be installed on all 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. If you are running CF over IP, then all interfaces used for CF over IP should be properly configured and be up and running. See "[Chapter 8 CF over IP](#)" for details.
4. Web-Based Admin View configuration must be configured. See "2.4.2 Management server configuration" in "PRIMECLUSTER Web-Based Admin View Operation Guide" for details.

To start the Cluster Admin screen, install the Java application ("PRIMECLUSTER Web-Based Admin View Startup") on a client.

Refer to "3.1.3.1 Installing Java application" and "3.1.3.2 Setting up Java application" in "PRIMECLUSTER Web-Based Admin View Operation Guide" to install and set up the Java application.

Log in to the Web-Based Admin View from the shortcut of the Java application (PRIMECLUSTER Web-Based Admin View Startup) and start the Cluster Admin screen from the menu.

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.

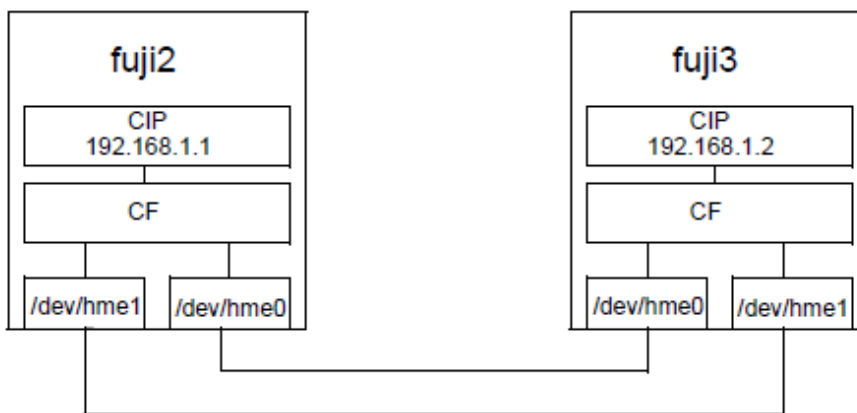
The CF/CIP Wizard is automatically started by selecting the node where CF has not been configured to start Cluster Admin.

1.1.1 Differences between CIP and CF over IP

Although the two terms CF over IP and CIP (also known as IP over CF) sound similar, they are two very distinct technologies.

CIP defines a reliable IP interface for applications on top of the cluster foundation (CF). CIP itself distributes the traffic generated by the application over the configured cluster interconnects (see Figure below).

Figure 1.1 CIP diagram



CF over IP uses an IPv4 interface, provided by the operating system, as a CF interconnect. This is not operated on IPv6. The IP interface should not run over the public network. It should only be on a private network, which is also the local network. The IP interface over the private interconnect can be configured by using an IP address designed for the private network. The IP address normally uses the following address:

```
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.

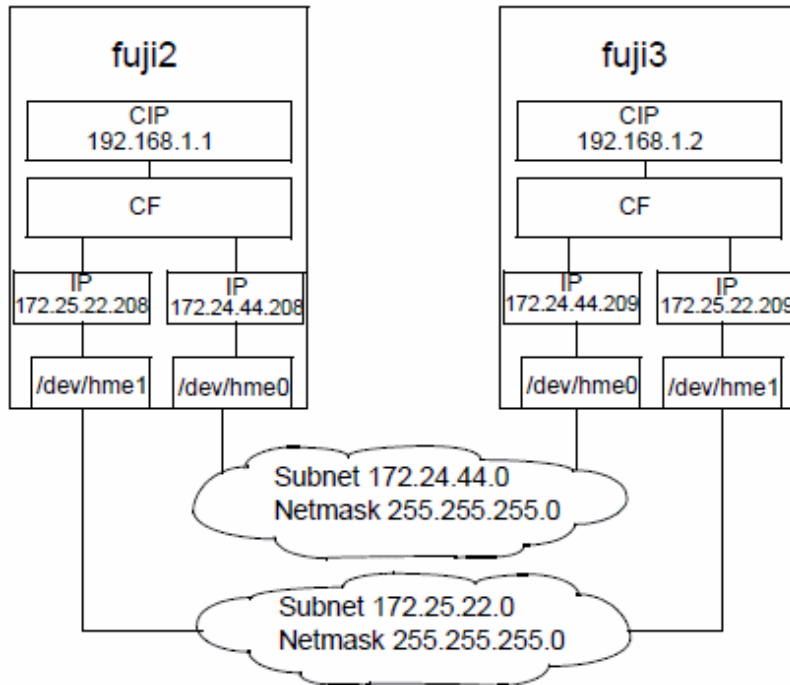
The following are possible scenarios for CF over IP:

- Where the cluster spans over two Ethernet segments of the same sub network.
Each sub-level Ethernet protocol is not forwarded across the router but does pass IP traffic.
- When you need to reach beyond the physical cable length.
Regular Ethernet is limited to the maximum physical length of the cable. Distances that are longer than the maximum cable length cannot be reached.
- If some of the network device cards that only support TCP/IP (for example, some Fiber channel) are not integrated into CF.

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 (see Figure below).

Figure 1.2 CF over IP diagram



- CF over IP is not supported in a Solaris 11 or later environment.

1.1.2 cfset

The cfset(1M) utility can be used to set certain tunable parameters in the CF driver. The values are stored in /etc/default/cluster.config. The cfset(1M) utility 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 cfcpl 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.
- Duplication of names will be detected and reported as an error when the entries are applied by cfconfig -l and by the cfset(1M) utility (cfset -r and -f option). This will log invalid and duplicate entries to /var/adm/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 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, treat 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 ]
```

The settable are as follows:

- CLUSTER_TIMEOUT (refer to the example that follows)
- CFSH (refer to the following Section "CF security")
- CFPC (refer to the following Section "CF security")

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

```
# cfset -r
```



Example

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.

1.1.3 CF security

In CF, cluster nodes execute commands on another node (cfsh), copy files from one node to another (cfcp) and there is the feature (CF Remote Services) to allow them. Because of this, these facilities are disabled by default.

The final step of the CF Configuration Wizard has two checkboxes. Checking one enables remote file copying and checking the other enables remote command execution.

PRIMECLUSTER has the exclusive feature for environment which does not support rhosts.

If the rhosts file is not used, it is necessary to enable the remote access by setting the parameters in cluster.config as below.

```
CFCP "cfcp"
CFSH "cfsh"
```

To deactivate, remove the settings from the /etc/default/cluster.config file and run cfset -r. cfsh does not support interactive commands. Therefore, some of the rsh functions are disabled.

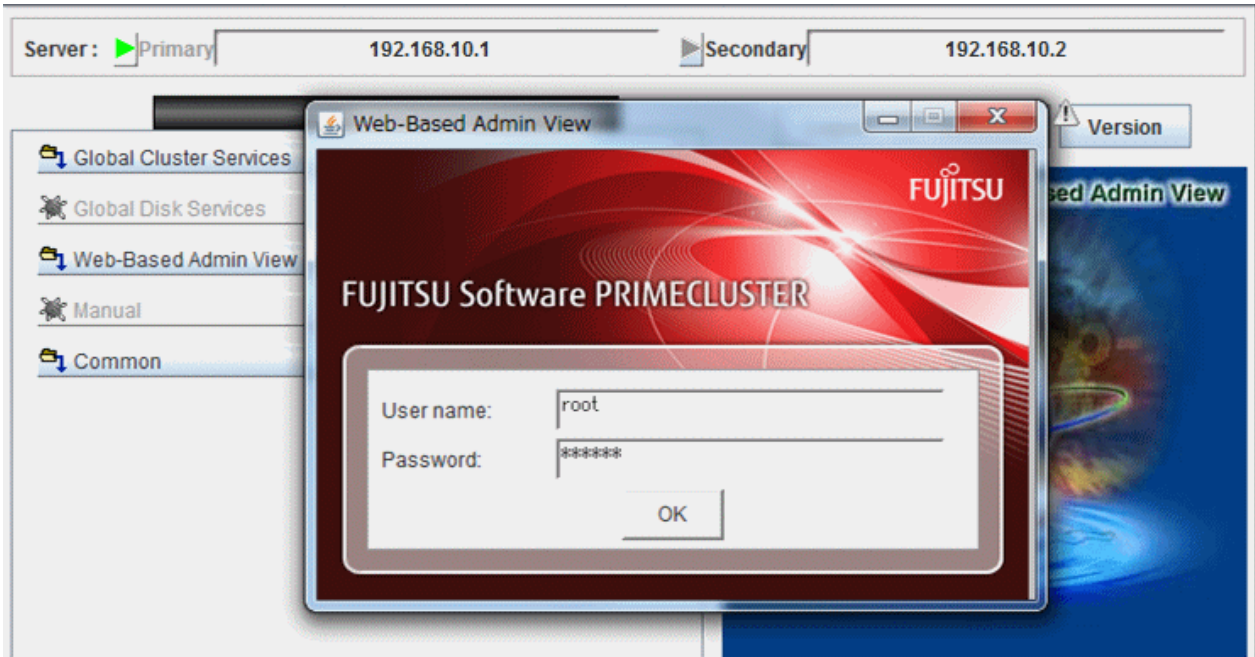
1.1.4 Example of creating a cluster

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.

Start the Web-Based Admin View screen from the shortcut of the Java application (PRIMECLUSTER Web-Based Admin View Startup). After a few moments, a login pop-up appears asking for a user name and password (similar to the screen below).

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.

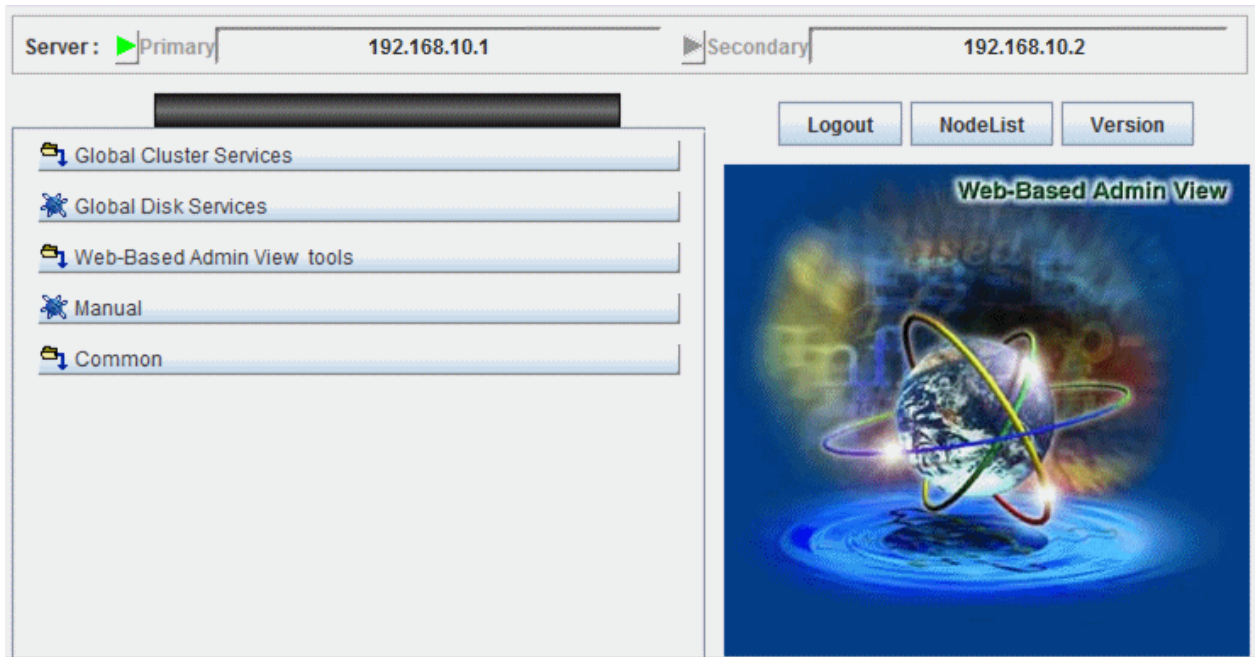
- A user in group wvroot

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, see "4.2.1 Assigning Users to Manage the Cluster" in "PRIMECLUSTER Installation and Administration Guide."

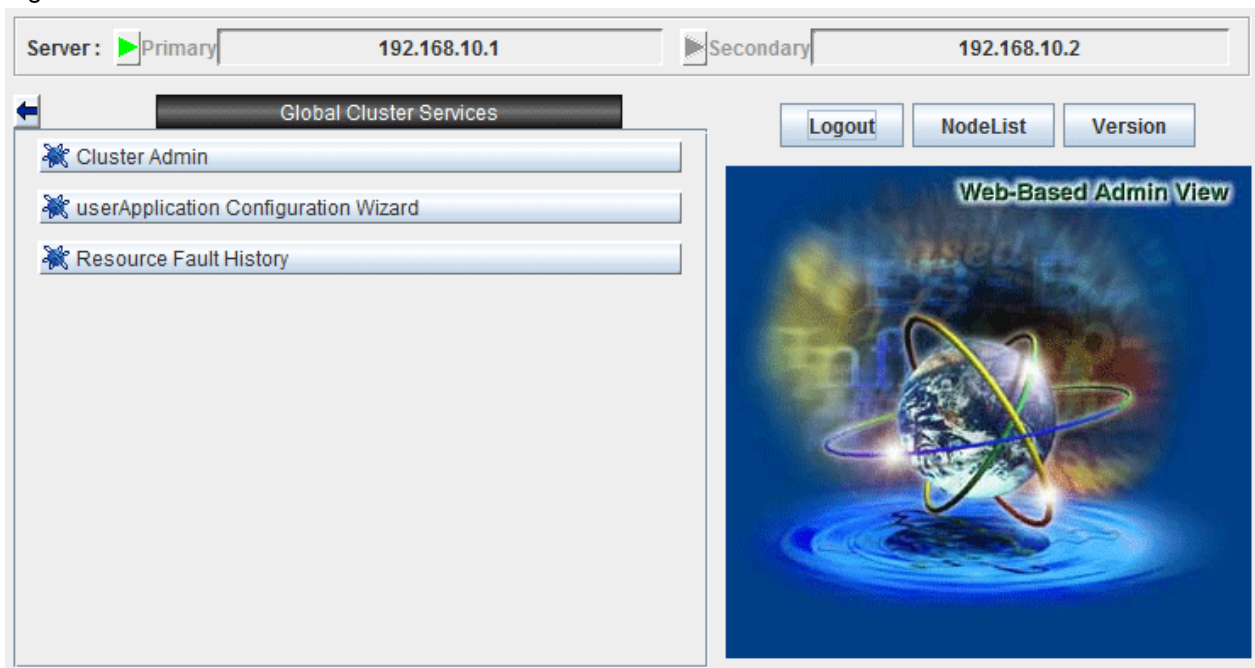
After clicking on the [OK] button, the top menu screen below appears. Click on the button labeled [Global Cluster Services].

Figure 1.4 Main Web-Based Admin View window after login



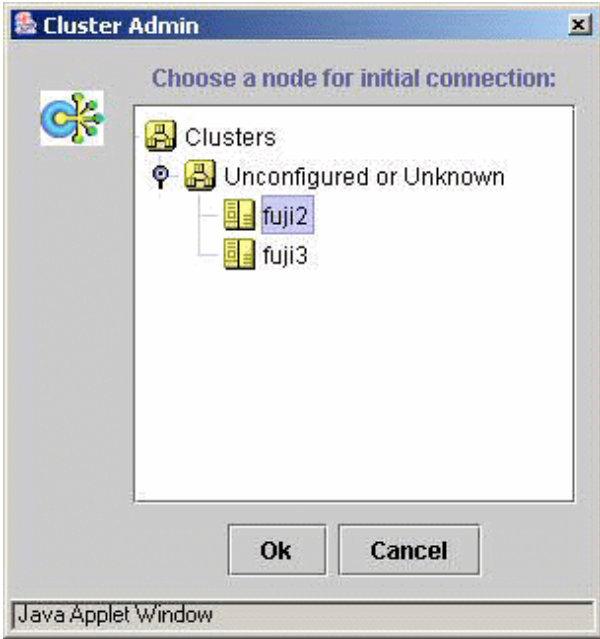
The Cluster Admin Selection screen below appears.

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. Choose a node for initial connection window appears below.

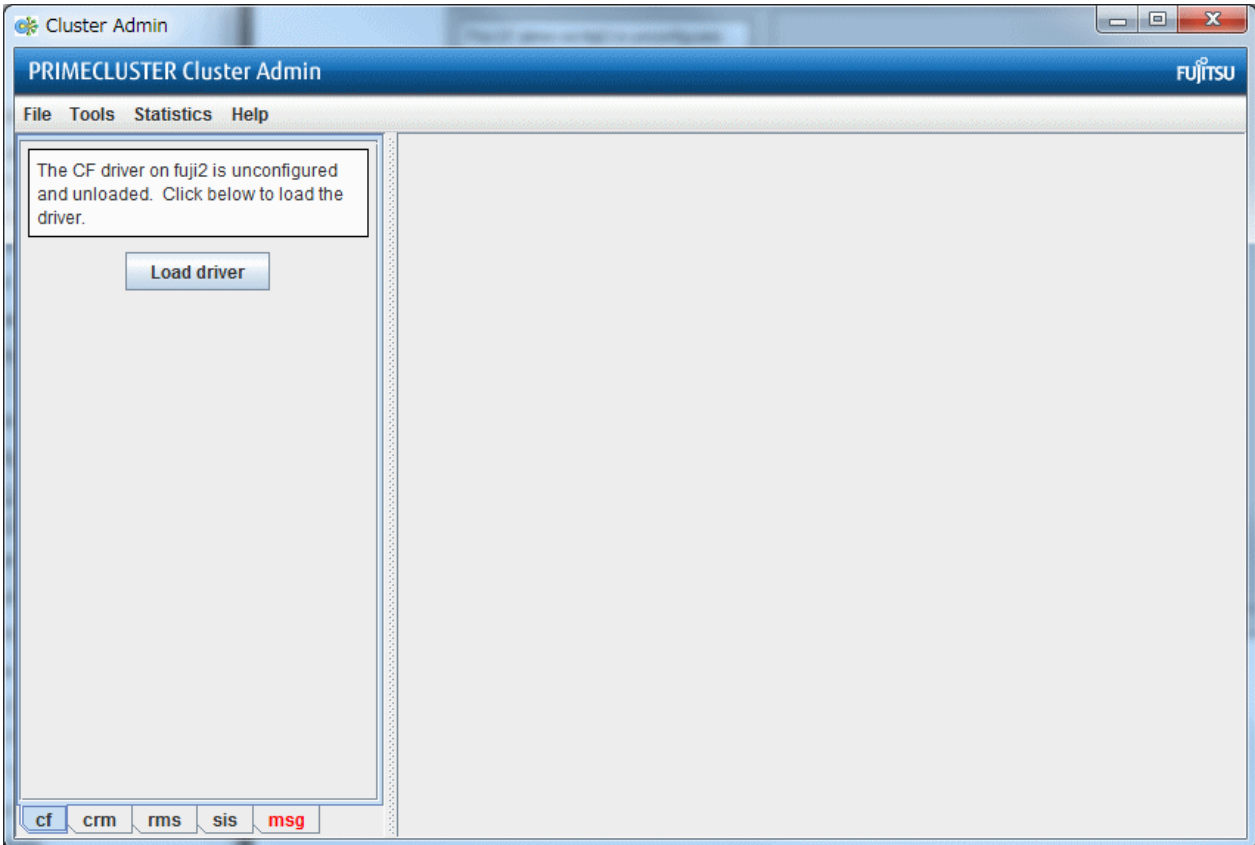
Figure 1.6 Initial connection pop-up



In the screen that selects the node for this initial connection, lists the nodes that are known to the Web-Based Admin View management station. However, if you select a node where CF has not yet been configured, the node is not displayed on the list in the [Node] tab. In this example, neither fuji2 nor fuji3 have had CF configured, so either would be acceptable as a choice. In the above screen, 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 one below appears.

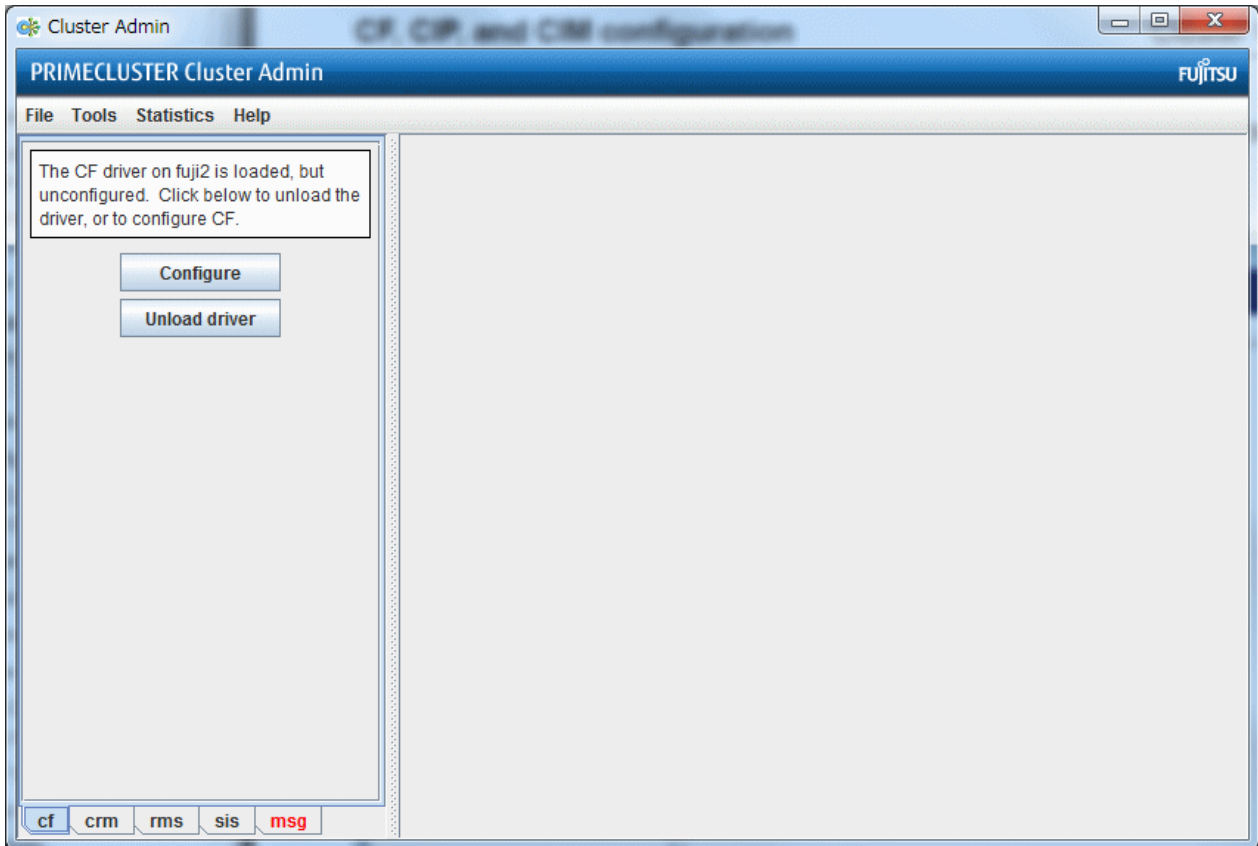
Figure 1.7 CF is unconfigured and unloaded



Click on the [Load driver] button to load the CF driver.

A window indicating that CF is loaded but not configured appears (similar below).

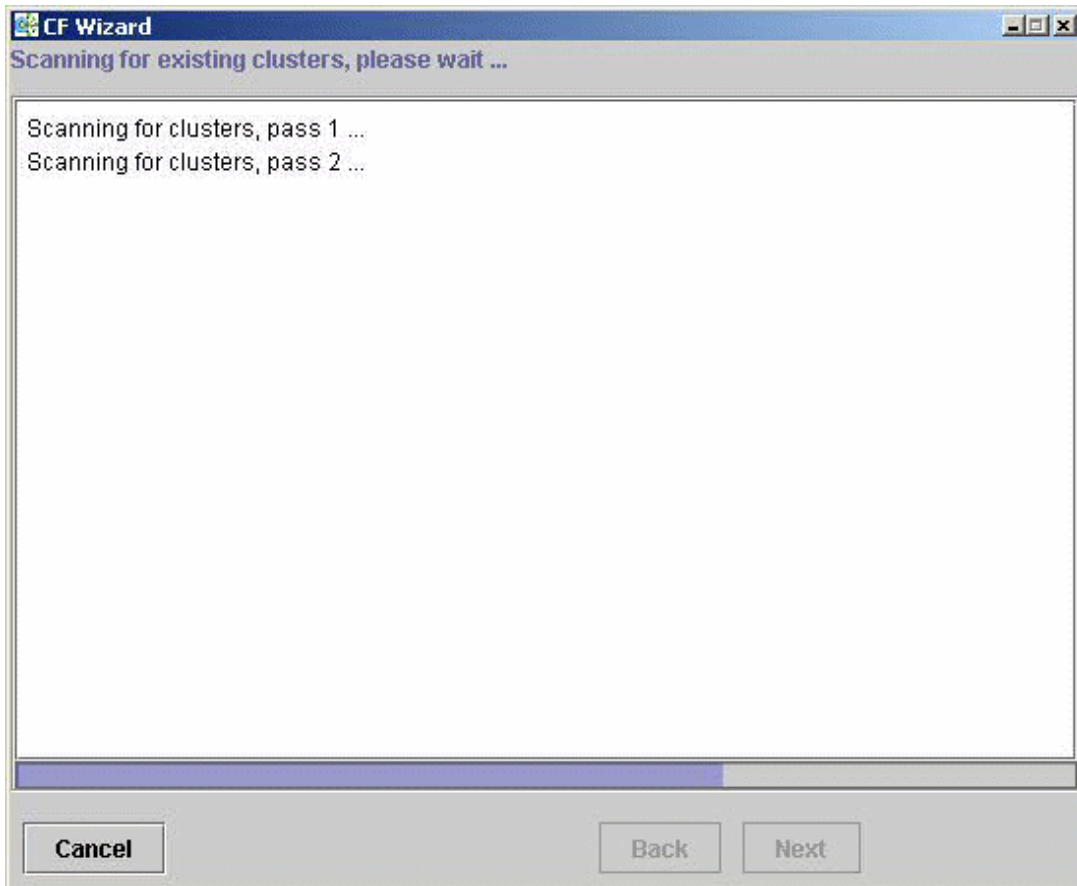
Figure 1.8 CF loaded but not configured



Click on the [Configure] button to bring up the CF Wizard.

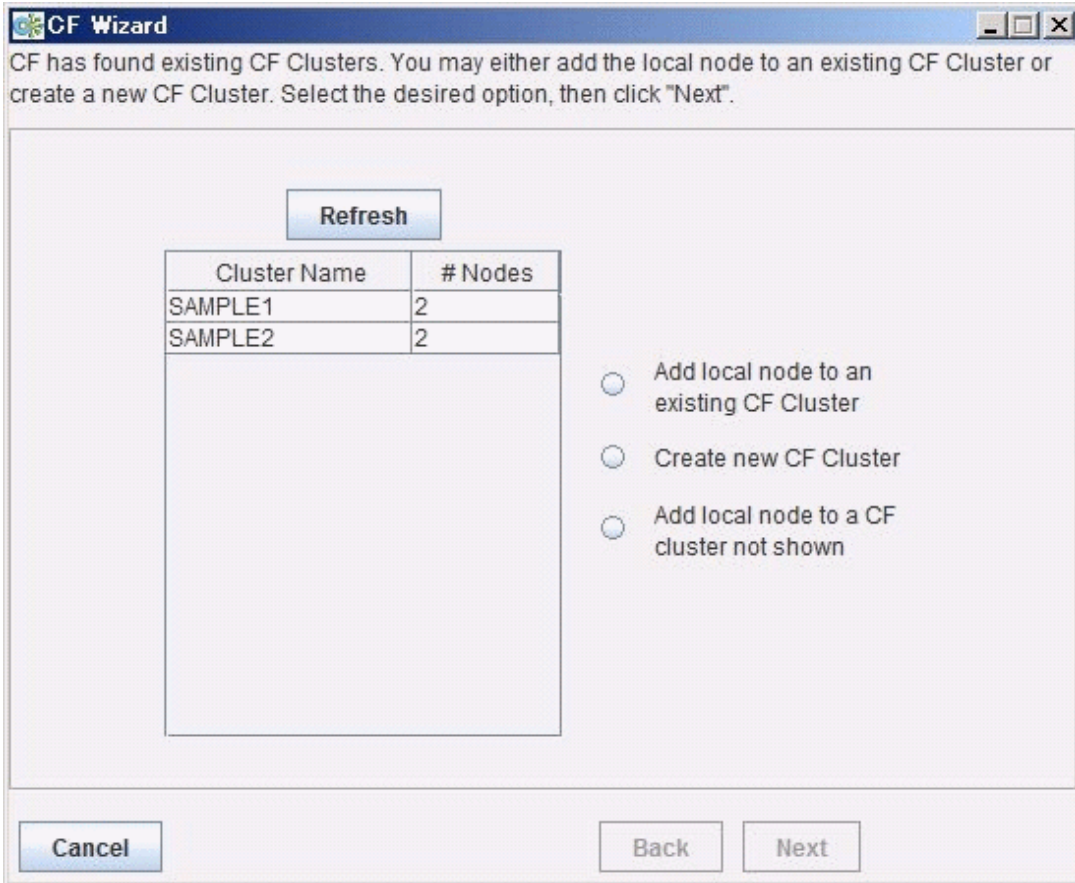
The CF Wizard begins by looking for existing clusters (screen below).

Figure 1.9 Scanning for clusters



After the CF Wizard finishes looking for clusters, the screen below is displayed.

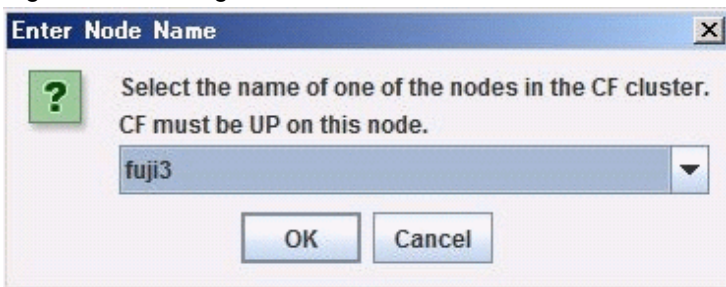
Figure 1.10 Creating or joining a cluster



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 1.11 Adding a local node to a CF cluster not shown

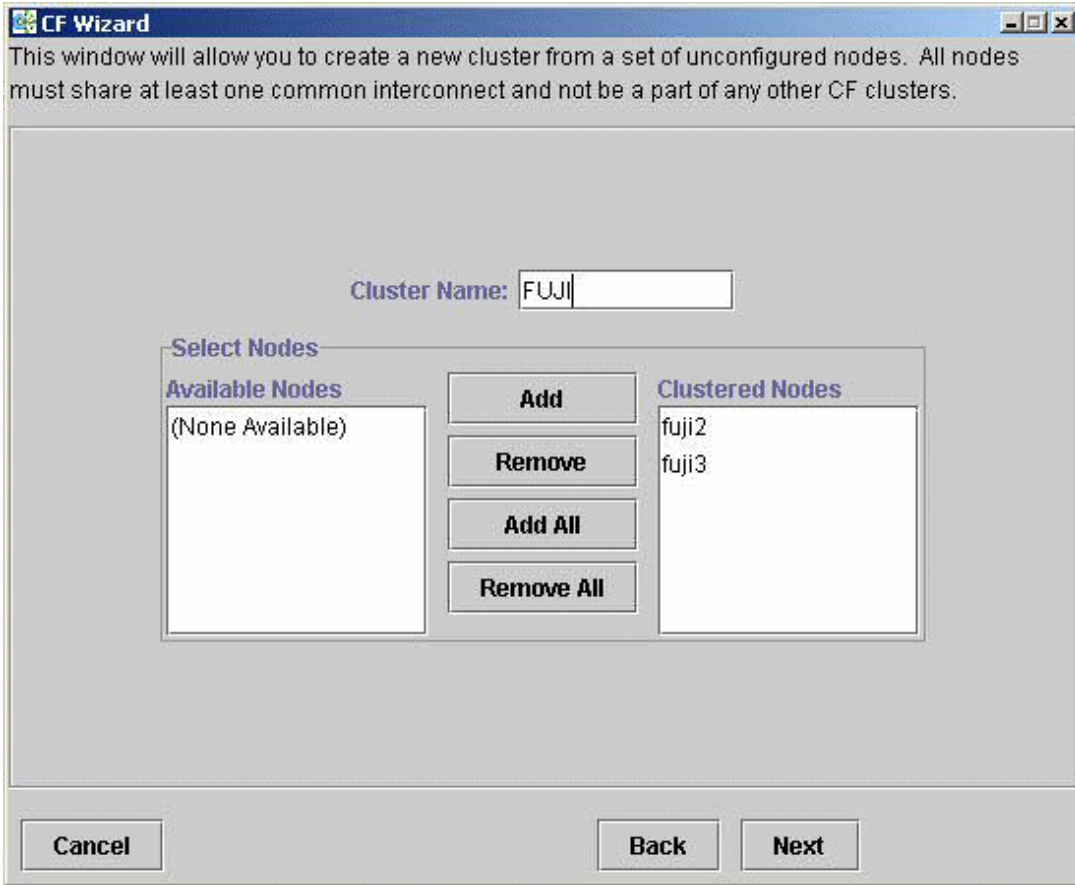


Enter the node name of the CF cluster that you want to join. Click [OK] to proceed.

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 the [Create new CF Cluster radio] button as shown in "Figure 1.10 Creating or joining a cluster." Then, click [Next]. Depending on your previous selection, a window for creating a new cluster or joining a cluster appears. The figure below is the screen for creating new cluster. The window for joining an existing cluster is very similar, except you cannot change the cluster name.

Figure 1.12 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 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.

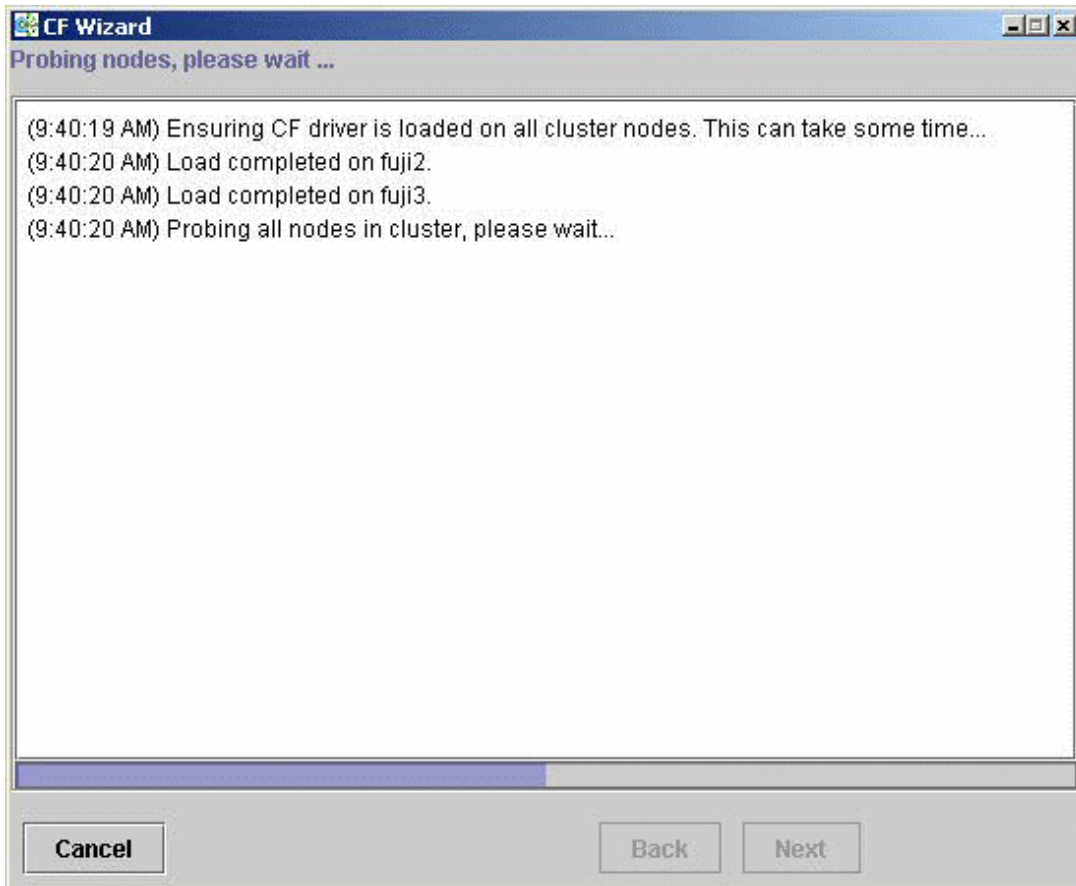
If you get to this window and you do not see all of the nodes that you want to be part of this cluster, then there is a very good chance that you have not configured Web-Based Admin View properly. When Web-Based Admin View is initially installed on the nodes in a potential cluster, it configures each 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 this screen 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.

See "4.2 Preparations for Starting the Web-Based Admin View Screen" in "PRIMECLUSTER Installation and Administration Guide" for more details on Web-Based Admin View configuration.

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 the one below appears.

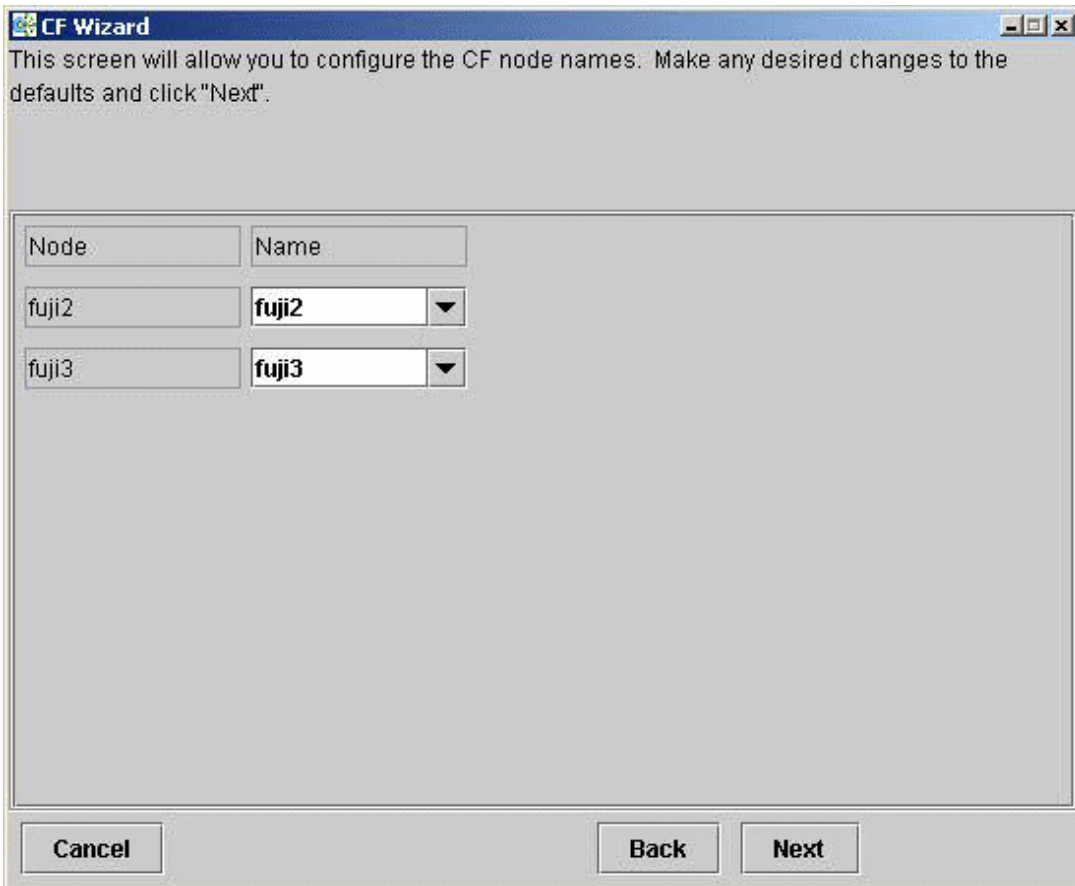
Figure 1.13 CF loads and pings



On most systems, loading the CF driver is a relatively quick process. However, on some systems that have certain types of large disk arrays, the first CF load can take up to 20 minutes or more.

The window that allows you to edit the CF node names for each node appears (see Figure 13). 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.

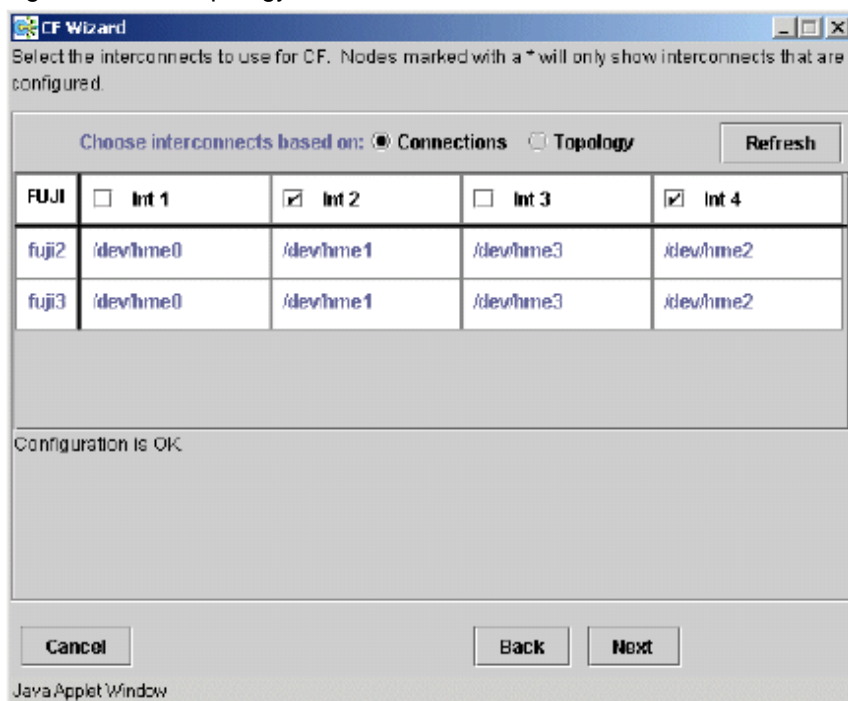
Figure 1.14 Editing CF node name



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 (similar below).

Figure 1.15 CF topology and connection table



Before using the CF topology and connection table of this screen, you should understand the following terms:

- Full interconnect

An interconnect where CF communication is possible to all nodes in the cluster.

- Partial interconnect

An interconnect where CF communication is possible between at least two nodes, but not to all 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 "[Figure 1.15 CF topology and connection table.](#)"

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 1.15 CF topology and connection table](#)", for example, Interconnect 1 (Int 1) has /dev/hme0 on fuji2 and fuji3 attached to it. The cluster name is FUJI.



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.

To configure CF using the connection table, click on the interconnects that have the devices that you wish to use. In "[Figure 1.15 CF topology and connection table](#)," Interconnects 2 and 4 have been selected. If you are satisfied with your choices, then you can click on [Next] to go to the CIP configuration window.

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 full interconnects, partial interconnects, and the devices that belong to the unconnected category and each category detected by the CF Wizard are displayed. A category where no target device exists is not displayed.

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.

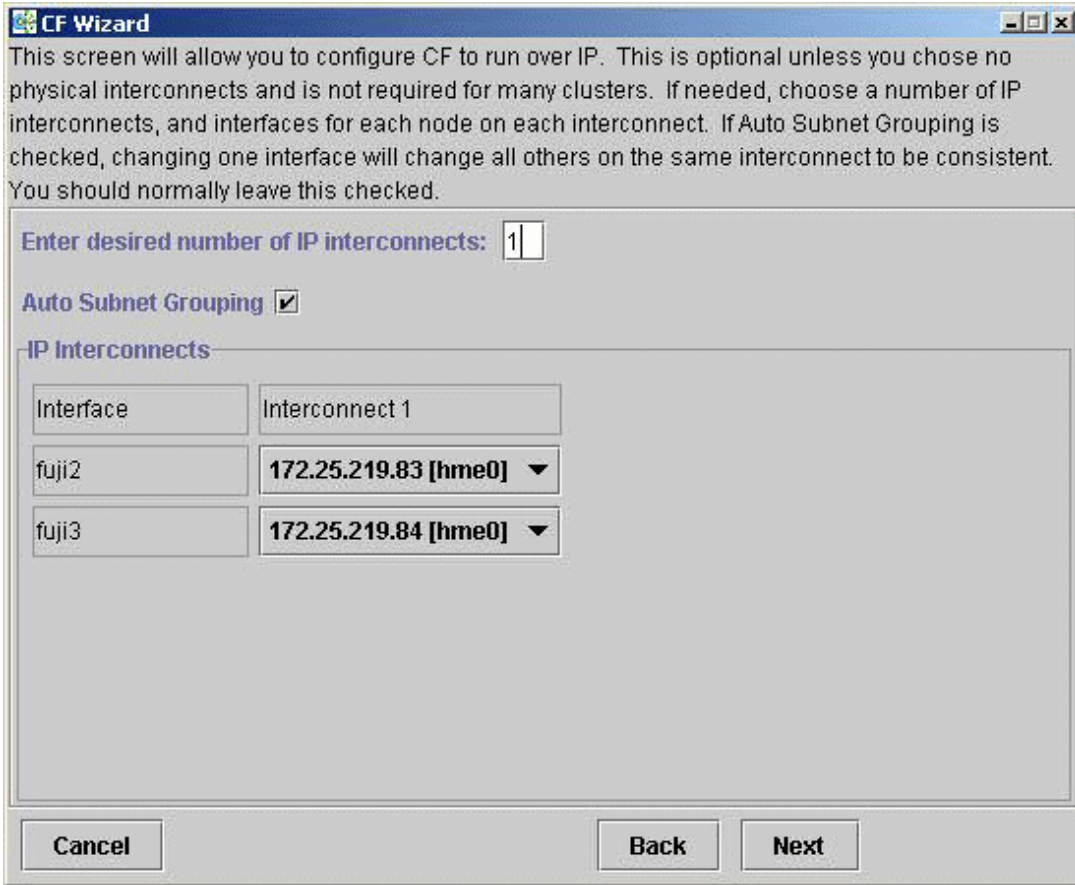
If the CF interconnect and the device are configured successfully, click [Next]. The "[Figure 1.16 "CF over IP" window](#)" is displayed.

Since CF over IP is not supported in a Solaris 11 or later environment, the CIP Wizard window as shown in the "[Figure 1.16 "CF over IP" window](#)" is not displayed. The "[Figure 1.17 CIP wizard \(IPv4\) window](#)" is displayed.

Select one or more full interconnects in the CF topology and connection table, and click [Next].

The "[Figure 1.18 CIP wizard \(IPv6\) window](#)" is displayed.

Figure 1.16 "CF over IP" window



This is optional. If desired, enter the desired number of IP interconnects and press [Return]. The CF Wizard then displays interconnects sorted according to the valid 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. CF over IP uses the IP broadcast address to find all the CF nodes during join process. So the dedicated network should be used for IP interconnects.

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 in this window, and click on [Next]. The "[Figure 1.17 CIP wizard \(IPv4\) window](#)" and "[Figure 1.18 CIP wizard \(IPv6\) window](#)" appear.

Figure 1.17 CIP wizard (IPv4) window

The screenshot shows a window titled "CF Wizard" with a standard Windows title bar. Below the title bar is a text area containing the following instructions: "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."

Below the text area are several configuration options:

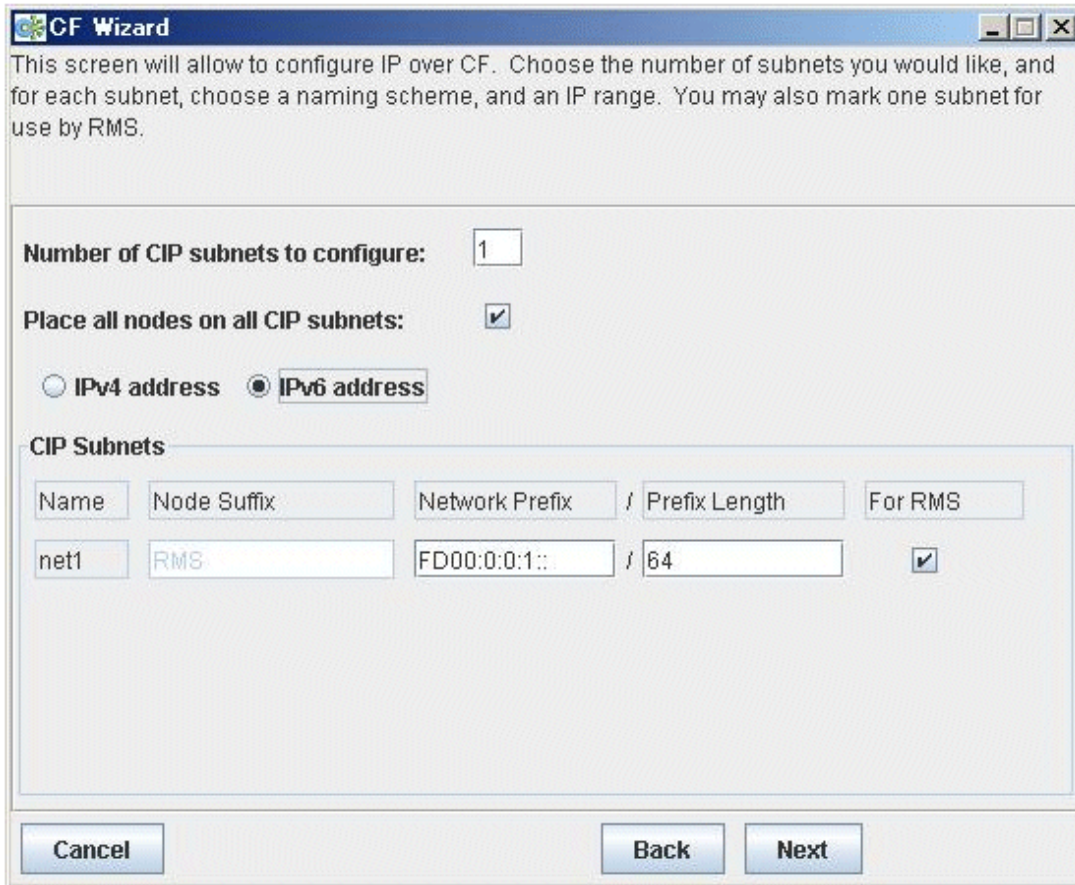
- "Number of CIP subnets to configure:" followed by a text input field containing the number "1".
- "Place all nodes on all CIP subnets:" followed by a checked checkbox.
- Two radio buttons: "IPv4 address" (which is selected) and "IPv6 address".

Below these options is a section titled "CIP Subnets" containing a table with five columns: "Name", "Node Suffix", "Subnet Number", "Subnet Mask", and "For RMS".

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

At the bottom of the window are three buttons: "Cancel", "Back", and "Next".

Figure 1.18 CIP wizard (IPv6) window



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 1.17 CIP wizard \(IPv4\) window](#)" and "[Figure 1.18 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. Refer to "[1.1.6 Example of CF setting by using CLI](#)" to set any IP address.

The CIP Wizard will automatically fill in a default value for the *Subnet Number* 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* 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 Host *Suffix*.

If the checkbox *For RMS* is selected, then the host 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 1.17 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:

```
IP 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:

```
IP 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 is a unique IP number on the network prefix specified in the [Network Prefix] field. The interface ID of the address starts from 1 and it is incremented by 1 for each additional node.

Refer to "1.1.6 Example of CF setting by using CLI" to set any IP address.

The CIP Wizard will automatically fill in a default value for the [Network 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 [Network 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 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 [Prefix Length] will be the value specified.

In "Figure 1.18 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 [Network 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:

```
IP nodename: fuji2RMS
IPv6 address: FD00:0:0:1::1
Prefix Length: 64
```

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

```
IP nodename: fuji3RMS
IPv6 address: FD00:0:0:1::2
Prefix Length: 64
```

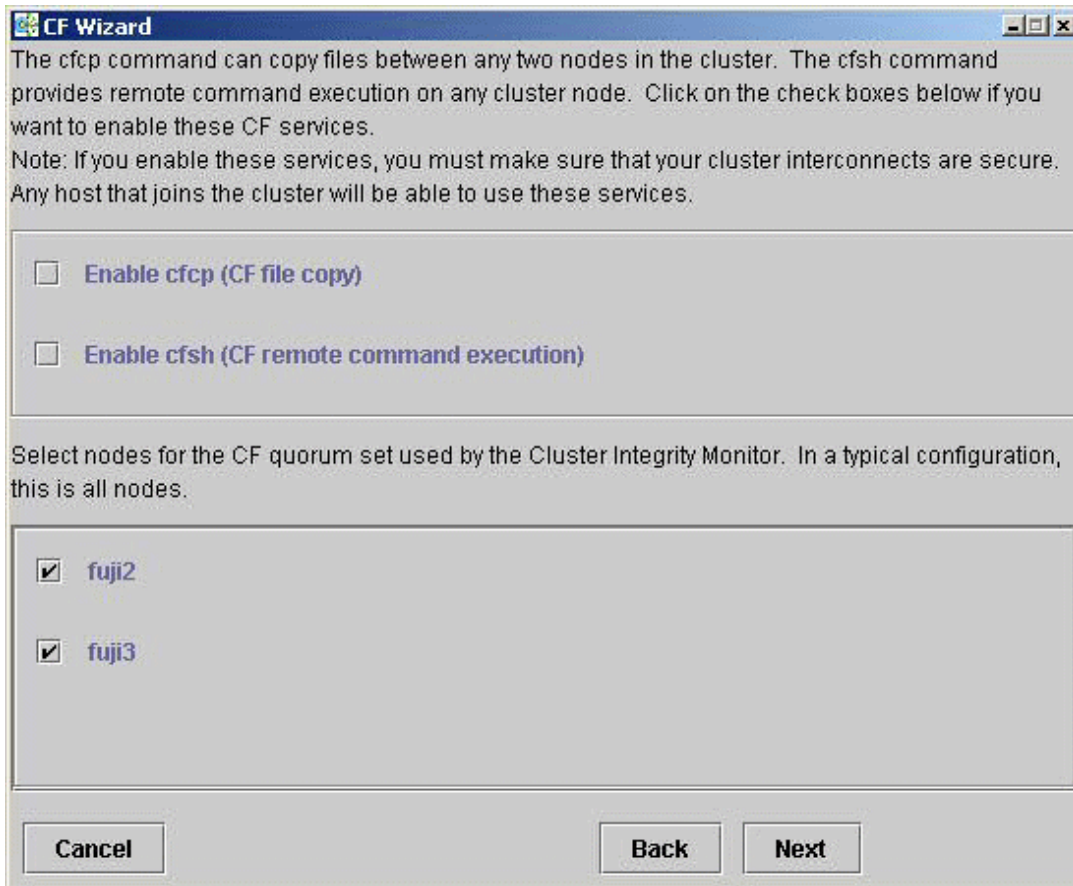
The CIP Wizard stores the configuration information in the file `/etc/cip.cf` on each node in the cluster. This is the default CIP configuration file. The Wizard will also automatically update `/etc/hosts` and `/etc/inet/ipnodes` on each node in the cluster to add the new IP nodenames. The cluster console will not be updated.



The CIP Wizard always follows an orderly naming convention when configuring CIP names. If you have done some CIP configuration by hand before running the CIP Wizard, then you should consult the Wizard documentation to see how the Wizard handles irregular names.

When you click on the [Next] button, the window below appears.

Figure 1.19 CIM configuration window



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

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

cfcf 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 similarly 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

Do not manually add the CIP node name to the */etc/hosts*, */etc/inet/ipnodes* because, each node of */etc/hosts*, */etc/inet/ipnodes* files with in the cluster are automatically renewed.

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.

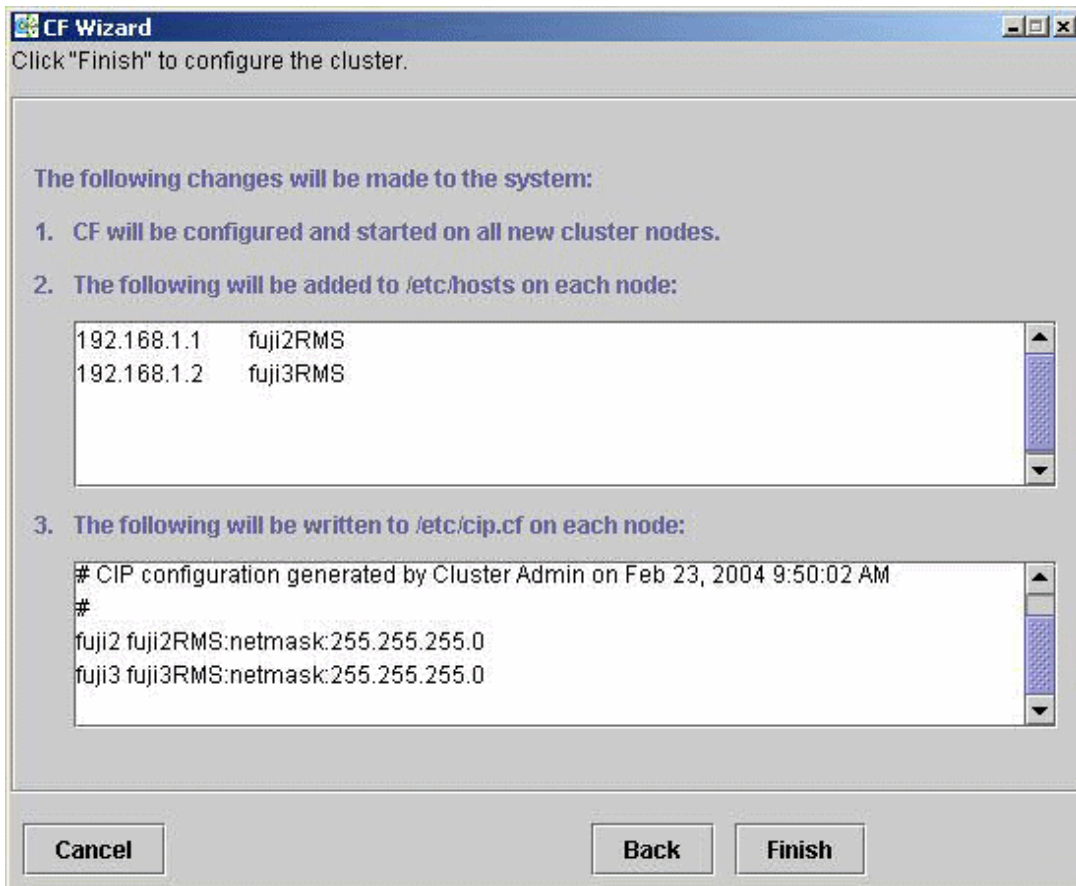
This window will also allow you to configure CF Remote Services. You can enable either remote command execution, remote file copying, or both.

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 *cfsh*.

Click on the [Next] button to go to the summary window below.

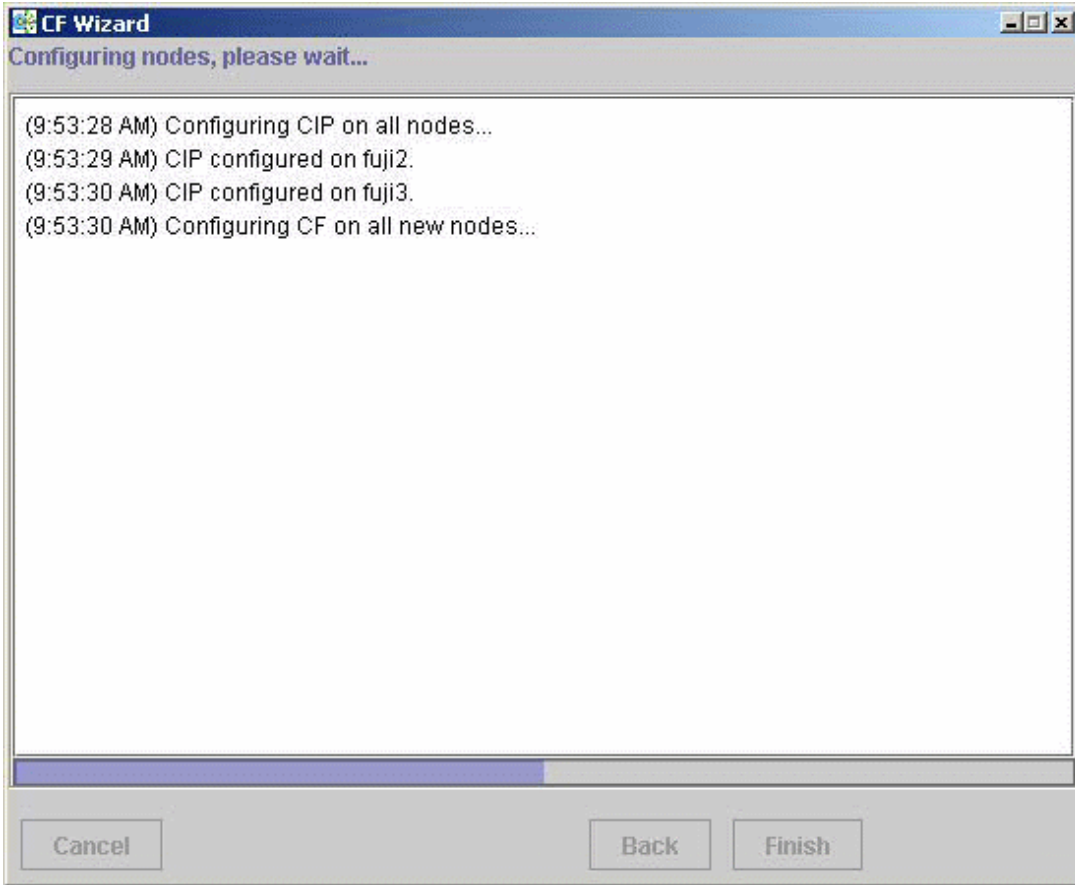
Figure 1.20 Summary window



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.

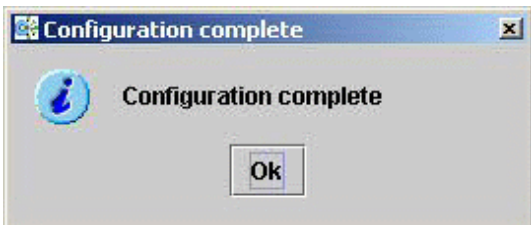
The window below is displayed while the configuration is being done.

Figure 1.21 Configuration processing window



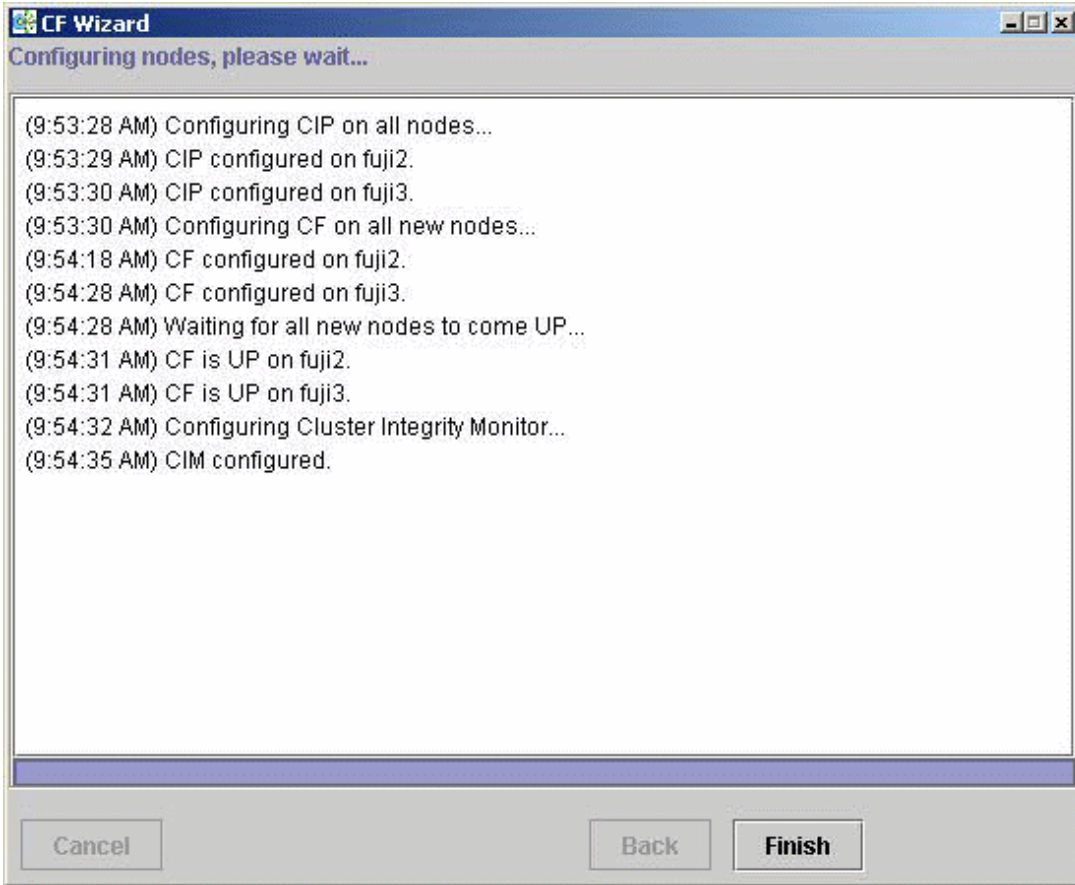
This window is updated after each configuration step. When the configuration successfully completes, a completion pop-up window similar to the one below appears.

Figure 1.22 Configuration completion pop-up



Click on the [OK] button to close the pop-up window. A [Finish] button is being displayed, for the configuration processing window similar to the one below.

Figure 1.23 Configuration window after completion



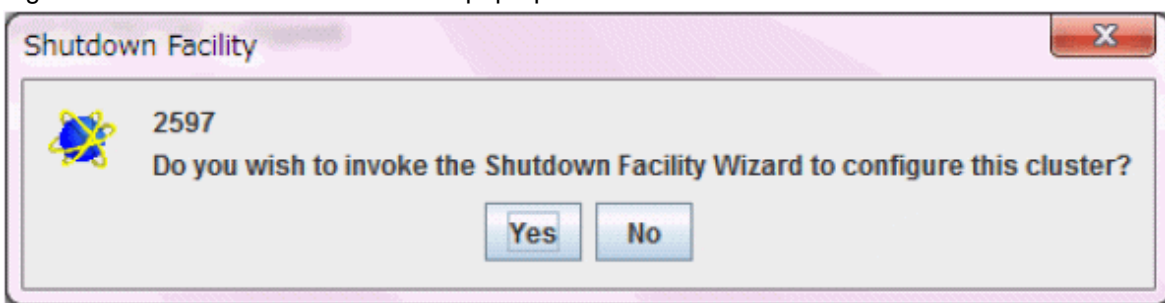
You might see the following error message in the window shown in "Figure 1.23 Configuration window after completion."

```
cf:cfconfig    OSDU_stop: failed to unload cf_drv
```

You can safely ignore this message.

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

Figure 1.24 SF wizard start notification pop-up

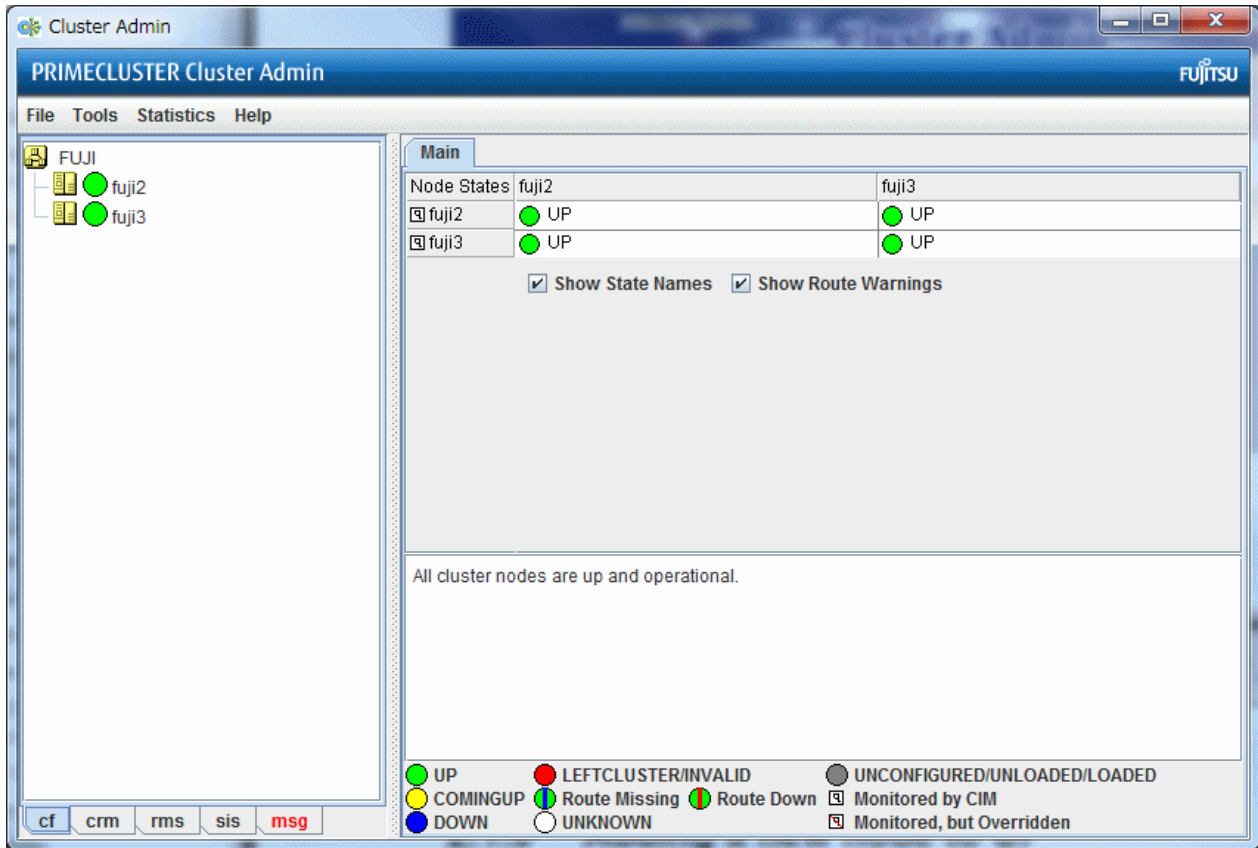


If you have already decided which shutdown agent to use, click the <Yes> button and start the SF wizard.

If you have not decided which shutdown agent to use yet, click the <No> button, and follow the procedure described in "5.1.2 Configuring the Shutdown Facility" in "PRIMECLUSTER Installation and Administration Guide" to decide the shutdown agent to use.

After the CF (and optionally the SF) Wizards are done, you see the main CF window as follows. After several moments, the window will be updated with new configuration and status information.

Figure 1.25 Main CF window



1.1.5 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 initial connection pop-up window (see "Figure 1.6 Initial connection pop-up"). 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 1.7 CF is unconfigured and unloaded"). 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.

1.1.6 Example of CF setting by using CLI

Follow the procedures below if setting CF by using CLI.

The following example shows the cluster system configured by two nodes with the CF node names "fuji2" and "fuji3."

1. Create the CIP configuration file.

Describe /etc/cip.cf as below on all the nodes that configure the cluster system.

Example:

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

2. Set the IP address.

Describe /etc/inet/hosts as below on all the nodes that configure the cluster system.

```
<cip address1> <CIP/Sysnode name1>
<cip address2> <CIP/Sysnode name2>
```

Example: If setting the CIP address of fuji2 to 192.168.1.1 and the CIP address of fuji3 to 192.168.1.2.

```
192.168.1.1 fuji2RMS
192.168.1.2 fuji3RMS
```

3. Use cfcf/cfsh to enable the remote access.

Describe /etc/default/cluster.config as below on all the nodes that configure the cluster system.

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

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

a. Edit /etc/default/cluster and create the file with the following content.

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

Example:

```
nodename fuji2
clustername FUJI
device /dev/hme1
device /dev/hme2
```



Make sure to specify the CF node name, not the OS node name, to nodename.

b. Set owners, groups and access rights.

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

c. Restart the node.

5. Execute the following command on any one node that configures the cluster system, and set up the Cluster Integrity Monitor (CIM).

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

nodename: CF node name

Example:

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

If this command returns an error, check if the cluster name and the CF node name set in /etc/default/cluster in step 4 are set correctly.

6. Check that transmission is possible with the CIP/Sysnode name.

Example: If checking from fuji2 to fuji3

```
# ping fuji3RMS
```

If transmission is not possible, check if the CF node name, .CIP/Sysnode name, and CIP address set in /etc/cip.cf and /etc/inet/hosts in step 1 and step 2 are set correctly.

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[:Option[:Option...]]
```

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 "1.1.4 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 "4.7 Starting and stopping CF".

1.3 Cluster Configuration Backup and Restore (CCBR)

Note

CCBR only saves PRIMECLUSTER configuration information.

CCBR provides a simple method to save the current PRIMECLUSTER configuration information of a cluster node. It also provides a method to restore the configuration information whenever a node update has caused severe trouble or failure, and the update (and any side-effects) must be removed. CCBR provides a node-focused backup and restore capability. Multiple cluster nodes must each be handled separately.

CCBR provides the following commands:

- `cfbackup(1M)`
Saves all information into a directory that is converted to a compressed tar archive file.
- `cfrestore(1M)`
Extracts and installs the saved configuration information from one of the `cfbackup(1M)` compressed tar archives.

After `cfrestore(1M)` is executed, you must reactivate the RMS configuration in order to start RMS. Once the reactivation of the RMS configuration is done, RMS will have performed the following tasks:

- Checked the consistency of the RMS configuration
- Established the detector links for RMS to be able to monitor resources
- Ensured proper communication between cluster nodes
- Created the necessary aliases for the shell commands used in the Wizard Tools. This is done automatically during RMS activation.

See "3.4 Activating a configuration" in "PRIMECLUSTER Reliant Monitor Services (RMS) with Wizard Tools Configuration and Administration Guide."

Note

- To guarantee that the `cfrestore(1M)` command will restore a functional PRIMECLUSTER configuration, it is recommended that there be no hardware or operating system changes since the backup was taken, and that the same versions of the PRIMECLUSTER products are installed.
- Because the installation or reinstallation of some PRIMECLUSTER products and kernel drivers, device reconfiguration may occur. This is usually not a problem. However, if Network Interface Cards (NICs) have been installed, removed, replaced, or moved, the device instance numbers (for example, the number 2 in `/dev/hme2`) can change. Any changes of this nature can, in turn, cause a restored PRIMECLUSTER configuration to be invalid.

`cfbackup(1M)` and `cfrestore(1M)` consist of a framework and plug-ins. The framework and plug-ins function as follows:

1. The framework calls the plug-in for the SMAWcf package.
2. This plug-in creates and updates the saved-files list, the log files, and error log files.
3. All the other plug-ins for installed PRIMECLUSTER products are called in name sequence.
4. Once all plug-ins have been successfully processed, the backup directory is archived by means of `tar(1M)` and compressed.
5. The backup is logged as complete and the file lock on the log file is released.

The `cfbackup(1M)` command runs on a PRIMECLUSTER node to save all the cluster configuration information. To avoid any problem, this command should be concurrently executed on every cluster node to save all relevant PRIMECLUSTER configuration information. This command must be executed as root. If a backup operation is aborted, no tar archive is created. If the backup operation is not successful for one plug-in, the command processing will abort rather than continue with the next plug-in. `cfbackup(1M)` exits with a status of zero on success and non-zero on failure.

The `cfrestore(1M)` command runs on a PRIMECLUSTER node to restore all previously saved PRIMECLUSTER configuration information from a compressed tar archive. The node must be in single-user mode with CF not loaded. The node must not be an active

member of a cluster. The node must not be an active member of a cluster. cfrestore(1M) exits with a status of zero on success and non-zero on failure.

It is recommended to reboot once cfrestore(1M) returns successfully. If cfrestore(1M) aborts, the reason for this failure should be examined carefully since the configuration update may be incomplete.

Note

- You cannot run cfbackup(1M) and cfrestore(1M) at the same time on the same node. However, cfbackup(1M) command can be run on multi-user mode, but cfrestore(1M) command cannot be run on single-user mode.
- Some PRIMECLUSTER information is given to a node when it joins the cluster. The information restored is not used. To restore and to use this PRIMECLUSTER information, the entire cluster needs to be DOWN, and the first node to create the cluster must be the node with the restored data. When a node joins an existing, running cluster, the restored configuration is gone because it is the first node in the cluster that determines which restored configuration to use.

The following files and directories that are fundamental to the operation of the cfbackup(1M) and cfrestore(1M) commands:

- The /opt/SMAW/ccbr/plugins directory contains executable CCBR plug-ins. The installed PRIMECLUSTER products supply them.
- The /opt/SMAW/ccbr/ccbr.conf file must exist and specifies the value for CCBRHOME, the pathname of the directory to be used for saving CCBR archive files.
A default ccbr.conf file, with CCBRHOME set to /var/spool/SMAW/SMAWccbr is supplied as part of the SMAWccbr package. The system administrator can change the CCBRHOME pathname at any time. The system administrator might need to change the CCBRHOME pathname to a file system with sufficient disk space.

Note

It is important to remember that re-installing the SMAWccbr package will reset the contents of the /opt/SMAW/ccbr/ccbr.conf file to the default package settings.

The following is an example of ccbr.conf:

```
#!/bin/ksh -
#ident "@(#)ccbr.conf  Revision: 12.1  02/05/08 14:45:57"
#
# CCBR CONFIGURATION FILE
#
# set CCBR home directory
#
CCBRHOME=/var/spool/SMAW/SMAWccbr
export CCBRHOME
```

- The /opt/SMAW/ccbr/ccbr.gen (generation number) file is used to form the name of the CCBR archive to be saved into (or restored from) the CCBRHOME directory.

This file contains the next backup sequence number. The generation number is appended to the archive name.

If this file is ever deleted, cfbackup(1M) or cfrestore(1M) will create a new file containing the value string of 1. Both commands will use either the generation number specified as a command argument, or the file value if no command argument is supplied. The cfbackup(1M) command additionally checks that the command argument is not less than the value of the /opt/SMAW/ccbr/ccbr.gen file. If the command argument is less than the value of the /opt/SMAW/ccbr/ccbr.gen file, the cfbackup(1M) command will use the file value instead.

Upon successful execution, the cfbackup(1M) command updates the value in this file to the next sequential generation number. The system administrator can update this file at any time.

- If cfbackup(1M) backs up successfully, a compressed tar archive file with the following name will be generated in the CCBRHOME directory as follows:

```
hostname_ccbrN.tar.Z
```

hostname is the nodename and *N* is the number suffix for the generation number.

For example, in the cluster node `fuji2`, with the generation number 5, the archive file name is as follows:

```
fuji2_ccbr5.tar.Z
```

- Each backup request creates a backup tree directory.
The directory is as follows:

This directory will be deleted after completing the execution of command.

```
CCBRHOME/nodename_ccbrN.
```

nodename is the node name and *N* is the number suffix for the generation number.

CCBROOT is set to this directory.

For example, enter the following on the node `fuji2`:

```
fuji2# cfbackup 5
```

Using the default setting for CCBRHOME, the following directory will be created:

```
/var/spool/SMAW/SMAWccbr/fuji2_ccbr5
```

This backup directory tree name is passed as an environment variable to each plug-in.

- The CCBRHOME/ccbr.log log file
Contains startup, completion messages, and error messages. All the messages are time stamped.
- CCBROOT/errlog log file
Contains specific error information when a plug-in fails. All the messages are time stamped.
- CCBROOT/plugin.blog or CCBROOT/plugin.rlog log files
Contain startup and completion messages from each backup/restore attempt for each plug-in. These messages are time stamped.

Example

Example 1: Backup

The following command backs up and validates the configuration files for all CCBR plug-ins that exist on the system `fuji2`.

```
fuji2# cfbackup
```

CCBR performs the backup automatically and does not require user interaction. Processing has proceeded normally when a message similar to the following appears at the end of the output: `cfbackup(1M)` command will output the following:

```
04/30/04 09:16:20 cfbackup 11 ended
```

This completes the backup of PRIMECLUSTER.

In the case of an error, the subdirectory `/var/spool/SMAW/SMAWccbr/fuji2_ccbr11` is created.

Refer to "[Chapter 9 Diagnostics and troubleshooting](#)" for more details on troubleshooting CCBR.

Example

Example 2: Restore

Before doing `cfrestore(1M)`, CF needs to be unloaded, the system needs to be in single-user mode, and the disks need to be mounted.

The following files are handled differently during `cfrestore(1M)`:

- root files

These are the files under the CCBROOT/root directory. They are copied from the CCBROOT/root file tree to their corresponding places in the system file tree.

- OS files

These files are the operating system files that are saved in the archive but not restored. The system administrator might need to merge the new OS files and the restored OS files to get the necessary changes.

For example, on fuji2 we entered the following command to restore the configuration to backup 11.

```
fuji2# cfrestore 11
```

The restore process asks you to confirm the restoration and then carries out the process automatically. Processing has proceeded normally when a message similar to the following appears at the end of the output:

```
05/05/04 13:49:19 cfrestore 11 ended
```

This completes the PRIMECLUSTER restore.



Chapter 2 CF Registry and Integrity Monitor

This chapter discusses the purpose and physical characteristics of the CF registry (CFREG), and it discusses the purpose and implementation of 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.

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 "[4.12 Adding and removing a node from CIM](#)" for the GUI procedures.

- CLI

For details on the CLI options and arguments, see the manual pages of each command. 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, after the node joins the cluster, it checks if any configuration is present in the CFReg.database. This is done as part of the GUI 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. For details on the CLI options and arguments, see the manual pages of each command.

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 "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

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

```
fuji2# cftool -n
Node   Number  State  Os      Cpu
fuji2  1       UP     Solaris Sparc
fuji3  2       UP     Solaris Sparc
```

- Display the current quorum configuration as follows:

```
fuji2# rcqconfig -g
```

Nothing is returned, since all nodes have been deleted from the quorum.

- 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, fuji10 (which is not in the cluster), in a quorum set of nodes as follows:

```
fuji2# rcqconfig -a fuji2 fuji3 fuji10
Cannot add node fuji10 that is not up.
```

The quorum cannot be set because fuji10 is specified that is excluded from the cluster. The quorum set remains empty.

```
fuji2# rcqconfig -g
```

Nothing is returned, since no quorum configuration has been done.



Chapter 3 Cluster resource management

This chapter discusses the Resource Database, which is a synchronized clusterwide database, holding information specific to several PRIMECLUSTER products.

3.1 Overview

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.

3.2 Kernel parameters for Resource Database

The default values of the Solaris kernel have to be modified when the Resource Database is used. This section lists the kernel parameters that have to be changed. In the case of kernel parameters that have already been set in the file `/etc/system`, the values recommended here should be added. In the case of kernel parameters that have not been defined in the file `/etc/system`, the values recommended here must be added to the default values.



Note

The values in the `/etc/system` file do not take effect until the system is rebooted. If an additional node is added to the cluster, or if more disks are added after your cluster has been up and running, it is necessary to recalculate using the new number of nodes and/or disks after the expansion, change the values in `/etc/system`, and then reboot each node in the cluster.

Refer to the OS manual for details on meanings and methods of changing kernel parameters.



Note

The values used for product and user applications operated under the cluster system must also be reflected in kernel parameter values.

The table below shows the value of a kernel parameter required to use the resource database.

Table 3.1 Kernel parameter

Kernel parameter	Value required for Resource Database
<code>semsys:semnfo_semmni</code>	20
<code>shmsys:shminfo_shmmni</code>	30
<code>shmsys:shminfo_shmmax</code>	Refer to the section that follows.

The value of `shminfo_shmmax` is calculated in the following way:

1. Remote resources:

$$DISKS \times NODES + 1) \times 2$$

DISKS is the number of shared disks. For disk array units, use the number of logical units (LUN). For devices other than disk array units, use the number of physical disks.

NODES is the number of nodes connected to the shared disks.

2. Local resources:

LOCAL_DISKS: Add up the number of local disks of all nodes in the cluster.

3. Total resources:

$$\text{Total resources} = (\text{remote resources} + \text{local resources}) \times 2776 + 1048576.$$

4. Selecting the value:

If `shminfo_shmmax` has already been changed for the other products, which means that `/etc/system` has a `shminfo_shmmax` entry, set the largest value among the following three values:

- Current value of `shminfo_shmmax`
- Value in Step 3
- 4194394

If `shminfo_shmmax` has not been altered from the default (meaning, there is no entry for `shminfo_shmmax` in `/etc/system`) and the result from Step 3 is greater than 8388608 (default value of Solaris OS), set `shminfo_shmmax` to the result of Step 3, otherwise `shminfo_shmmax` is not edited.

In summary, the formula to calculate the total resources is as follows:

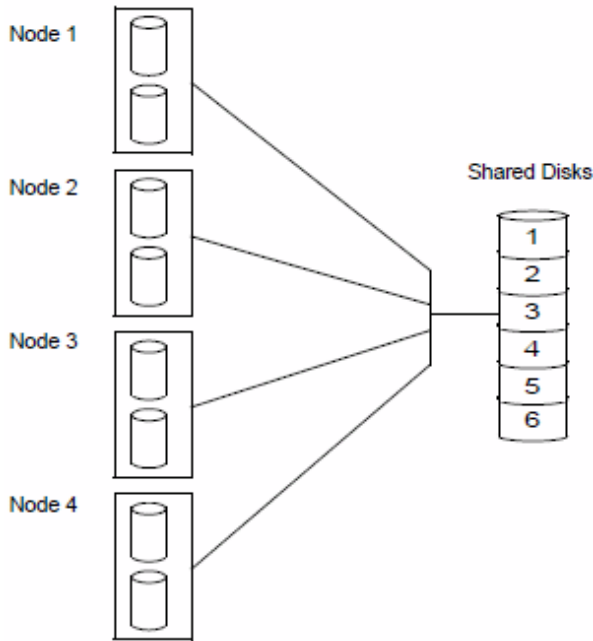
$$TotalResources = \{ DISKS \times (NODES + 1) \times 2 + LOCAL_DISKS \} \times 2776 + 1048576$$

- `shminfo_shmmax` is defined in `/etc/system`
 - If the current value is greater than *TotalResources* and less than 4194394:
You do not need to change `shminfo_shmmax`.
 - If the current value is greater than *TotalResources* and less than 4194394:
You need to change `shminfo_shmmax` to 4194394.
 - If neither of the above:
You need to change `shminfo_shmmax` to *TotalResources*.
- `shminfo_shmmax` is not defined in `/etc/system`
 - If *TotalResources* is greater than the default value (8388608) of Solaris OS:
You need to change `shminfo_shmmax` to *TotalResources*.
 - If *TotalResources* is the default value (8388608) of Solaris OS or less:
You do not need to change `shminfo_shmmax`.

Example

Take the diagram below as an example, the following article describes how to calculate the total resources.

Figure 3.1 Cluster resource diagram



Referring to the diagram above, calculate the total resources as follows:

- 1.Remote resources:

$$DISKS=6 \text{ NODES}=4$$

$$\text{Remote Resources} = 6 \times (4+1) \times 2 = 60$$

- 2.Local resources:

$$LOCAL_DISKS = 2 \times 4 = 8$$

- 3.Total resources:

$$1048576 + 2776 \times (60+8) = 1237344$$

Since 1237344 is less than 4194394, it is necessary to set 4194394 for `shminfo_shmmax`. Since 1237344 is less than 4194394, it is necessary to set 4194394 for `shminfo_shmmax`.

3.3 Resource Database configuration

You must configure the Resource Database after configuring CF, CIP, and CIM.

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.

If you need to add a new node to the cluster, and the existing nodes are already running the Resource Database, then a slightly different procedure needs to be followed. Refer to "3.6 Adding a new node" for details.

You can easily configure the Resource Database by using the CRM main window of Cluster Admin.

The following describes how to configure the Resource Database by using the commands. For how to configure the Resource Database by using the CRM main window, see "3.4.4 Configuring the Resource Database by using the CRM main window."

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 "[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:

```
CF node name RMS (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 nodes in the cluster.
3. Reboot each node.

4. Rerun the `clsetup` command on each node.
5. Use the `clgettree` command to verify the configuration

3.4 Registering hardware information

This section explains how to register hardware information in the Resource Database.

You can register the following hardware in the Resource Database by executing the `clautoconfig` command. For details, see the manual page for `clautoconfig`.

- Shared disk unit
- Network interface card
- Line switching unit (Only in Oracle Solaris 10 environment)

3.4.1 Setup exclusive device list

If you have any disk devices that needs to be excluded from automatic resource registration, describe the devices in the

```
/etc/opt/FJSVcluster/etc/diskinfo
```

File (exclusive device list) on all nodes. List all the disks in this exclusive device list that meet the following conditions:

- Disks that should not be used for cluster services
- Disks that should be registered in the resource database in other cluster system

An example of the `/etc/opt/FJSVcluster/etc/diskinfo` file that is setup is as follows:

```
# cat /etc/opt/FJSVcluster/etc/diskinfo <RETURN>
c1t0d16
c1t0d17
c1t0d18
c1t0d19
. . . . .
emcpower63
emcpower64
emcpower65
emcpower66
```

Refer to "[3.4.2 Exclusive device list for EMC Symmetrix](#)" if you use the EMC Symmetrix series of RAID devices (Symmetrix) in a SPARC Enterprise M-series/PRIMECLUSTER environment.

3.4.2 Exclusive device list for EMC Symmetrix

This section describes how to set up an exclusive device list (disk devices that should be excluded from automatic resource registration) when the EMC Symmetrix series of RAID devices (Symmetrix) is used in a SPARC Enterprise M-series/PRIMECLUSTER environment (refer to "[3.4.1 Setup exclusive device list](#)").

You must exclude the following EMC Symmetrix devices from automatic resource registration:

- BCV (Business Continuance Volume) devices
- R2 (SRDF target) devices
- GateKeeper devices
- CKD (Count Key Data) devices
- VC MDB (Volume Configuration Management Data Base) devices used by EMC SAN management software (Volume Logix, ESN Manager, SAN Manager)

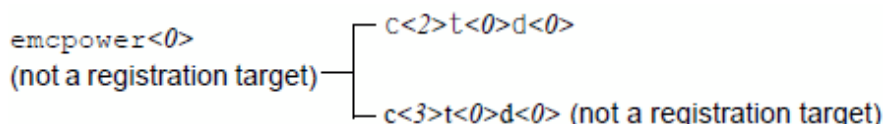
Add these devices in an exclusive device list after completing the settings for BCV, GateKeeper and EMC PowerPath. Then, you can perform automatic resource registration.

3.4.2.1 emcpower Devices and native Devices

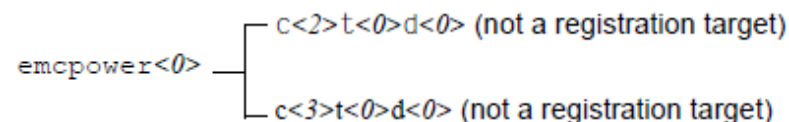
You can set emcpower devices or native devices that compose emcpower devices as the targets of automatic resource registration.

You can set emcpower devices or native devices that compose emcpower devices as the targets of automatic resource registration. When you use native devices, there is the benefit of not having to reexecute automatic resource registration when you change a storage device to a higher model. However, for systems in which emcpower devices are already set as the targets of automatic resource registration, continue to use the emcpower devices.

When setting native devices as the targets of automatic resource registration, specify all emcpower devices (emcpower<N>) and the native devices to be excluded from registration (c<C>t<T>d<D>) in the exclusive device list.



When setting emcpower devices as the targets of automatic resource registration, do not specify either emcpower devices (emcpower<N>) or native devices (c<C>t<T>d<D>) in the exception device list.



Where <C> is the controller number, <T> is the target ID, <D> is the disk number, and <N> is the emcpower device number.

3.4.2.2 BCV, R2, GateKeeper, CKD

You can differentiate which disk is BCV, R2, GateKeeper, or CKD by executing the syminq command provided in SYMCLI. Execute the syminq command, and describe all the devices (c<C>t<T>d<D>, emcpower<N>), indicated as BCV, R2, GK, or CKD in the excluded device list. Where <C> is the controller number, <T> is the target ID, <D> is the disk number, and <N> is the emcpower device number.

3.4.2.3 VCMDB

VCMDB is not output by executing syminq. If you use EMS SAN management software such as Volume Logix, ESN Manager or SAN Manager, check the VCMDB device name with EMC customer support engineers or a system administrator who set up the management software before adding the VCMDB to an exclusive device list.

3.4.2.4 Simplified setup for exclusive device list - clmakediskinfo, clmkdiskinfo

PRIMECLUSTER provides the following sample scripts for simplified setup of an exclusive device list: /etc/opt/FJSVcluster/sys/clmakediskinfo.sample /etc/opt/FJSVcluster/sys/clmkdiskinfo.sample

To set native devices as targets of automatic resource registration, use clmakediskinfo. Executing the command shown below creates an exclusive device list that contains emcpower devices, native devices to be excluded from automatic resource registration, as well as the BCV, R2, GateKeeper, and CKD devices.

```
# cp /etc/opt/FJSVcluster/sys/clmakediskinfo.sample
  /mydir/clmakediskinfo
# chmod u+x /mydir/clmakediskinfo
```

```
# /mydir/clmakediskinfo -M >
/etc/opt/FJSVcluster/etc/diskinfo <RETURN>
```

To use this script, use the vi command and modify the following two parameters (syminq and powermt command paths) in the script so that they match the execution environment.

```
SYMINQ=/usr/symcli/bin/syminq
POWERMT=/etc/powermt
```

To set emcpower devices as targets of automatic resource registration, use `clmkdiskinfo`. Executing the command shown below creates an exclusive device list that includes the BCV and GateKeeper devices.

```
# cp /etc/opt/FJSVcluster/sys/clmkdiskinfo.sample
   /mydir/clmkdiskinfo
```

```
# syminq | nawk -f /mydir/clmkdiskinfo >
/etc/opt/FJSVcluster/etc/diskinfo <RETURN>
```

If there are other devices to be included in the exclusive device list besides those listed automatically by the executed script, use the `vi` command and add those devices to the list.

If you do not know the path of the `syminq` command, check the SYMCLI installation settings. Normally the path is `/usr/symcli/bin/syminq`.

If you do not know the path of the `powermt` command, check the PowerPath installation settings. Normally the path is `/etc/powermt`.



Note

- PowerPath is required to use EMC Symmetrix.

Set the BCV and R2 devices to be used in the GDS Snapshot proxy configuration as targets of automatic device registration.

When setting the native devices that configure the BCV and R2 devices as targets of automatic resource registration, specify the emcpower devices (`emcpower<N>`) and the native devices (`c<C>t<T>d<D>`) to be excluded from registration in the exclusive device list.

When setting the BCV and R2 devices themselves as targets of automatic resource registration, do not include the BCV and R2 devices (`emcpower<N>`) or the native devices (`c<C>t<T>d<D>`) in the exclusive device list.

For details of GDS Snapshot, see "PRIMECLUSTER Global Disk Services Configuration and Administration Guide."

- If BCV is not added to an exclusive device list, you need to cancel or split the BCV pair before working on automatic resource registration.
- If the R2 device of the SRDF pair is not added to an exclusive device list, split the SRDF pair before working on automatic resource registration.

3.4.3 Automatic resource registration

This section explains how to register the detected hardware in the Resource Database

The registered network interface card should be displayed in the plumb-up state as a result of executing the `ifconfig(1M)` command.

Do not modify the volume name registered in VTOC using the `format(1M)` command after automatic resource registration. The volume name is required when the shared disk units are automatically detected.

The following prerequisites should be met:

- The Resource Database setup is done.
- Hardware is connected to each node.
- All nodes are started in the multi-user mode.

Take the following steps to register hardware in the Resource Database. This should be done on an arbitrary node in a cluster system.

1. Log in with system administrator access privileges.
2. Execute the `clautoconfig` command, using the following full path:

```
# /etc/opt/FJSVcluster/bin/clautoconfig -r
```

3. Confirm registration.

Execute the `clgettree` command for confirmation as follows:

```
# /etc/opt/FJSVcluster/bin/clgettree <RETURN>
Cluster 1 cluster0
      Domain 2 domain0
```



```

Shared 7 SHD_domain0
    SHD_DISK 9 shd001 UNKNOWN
        DISK 11 c1t1d0 UNKNOWN node0
        DISK 12 c2t2d0 UNKNOWN node1
    SHD_DISK 10 shd002 UNKNOWN
        DISK 13 c1t1d1 UNKNOWN node0
        DISK 14 c2t2d1 UNKNOWN node1
Node 3 node0 ON
    Ethernet 20 hme0 UNKNOWN
    DISK 11 c1t1d0 UNKNOWN
    DISK 13 c1t1d1 UNKNOWN node0
Node 5 node1 ON
    Ethernet 21 hme0 UNKNOWN
    DISK 12 c2t2d0 UNKNOWN
    DISK 14 c2t2d1 UNKNOWN

```

Reference

When deleting the resource of hardware registered by automatic registration, the following commands are used. Refer to the manual page for details of each command.

- `cldeldevice` Deletes the shared disk resource
- `cldelrsc` Deletes the network interface card resource
- `cldelswursc` Deletes the line switching unit resource (Only in Oracle Solaris 10 environment)

3.4.4 Configuring the Resource Database by using the CRM main window

This section explains how to configure the Resource Database that the cluster resource management facility (hereinafter CRM) manages. Configure CRM as follows.

- Initial installation

Configure the Resource Database that CRM manages.

- Automatic resource registration of units

Register the hardware units that are connected to the system (shared disk, network interface card, line switching unit) to the Resource Database that CRM manages.



Information

For details on the configuration, see "5.1.3 Initial Setup of the Cluster Resource Management Facility" in "PRIMECLUSTER Installation and Administration Guide."

3.5 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/adm/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.

3.5.1 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.

3.6 Adding a new node

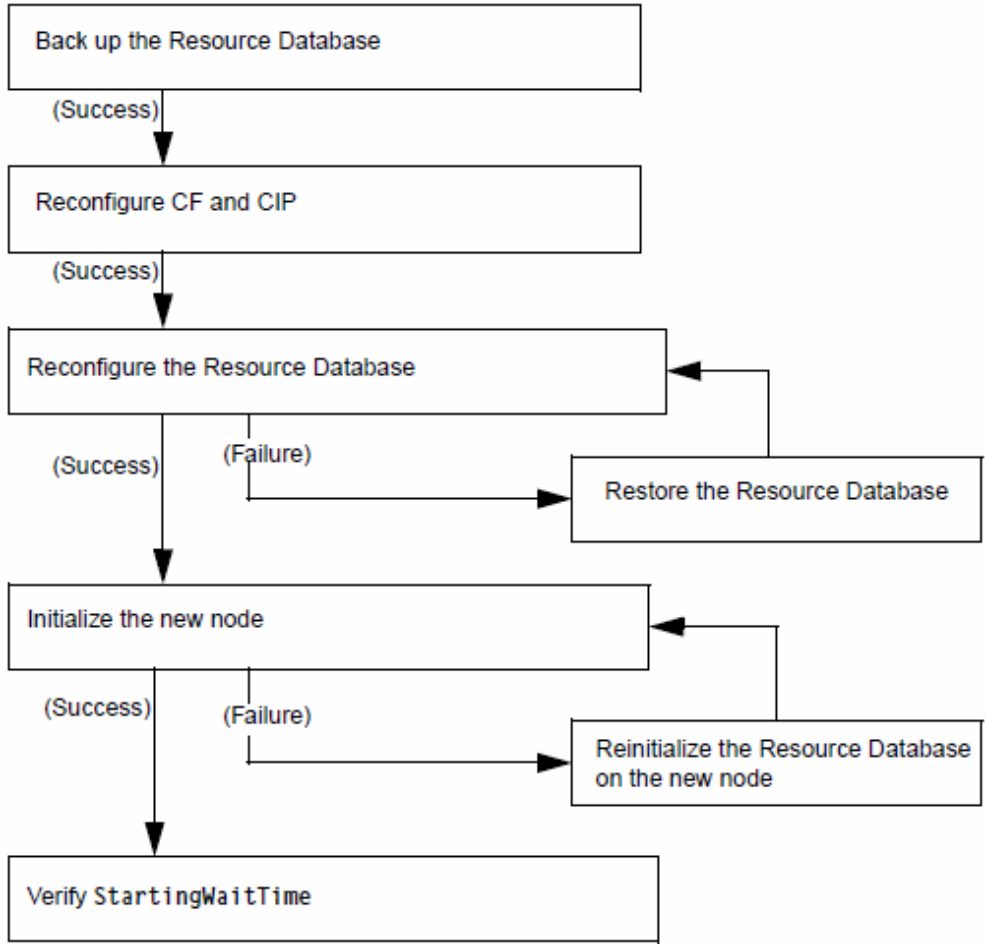
If you have a cluster where the Resource Database is already configured, and you would like to add a new node to the configuration, then you should follow the procedures in this section. You will need to make a configuration change to the currently running Resource Database and then configure the new node itself. The major steps involved are listed below:

1. Back up the currently running Resource Database. A copy of the backup is used in a later step to initialize the configuration on the new node. It also allows you to restore your configuration to its previous state if a serious error is encountered in the process.
2. Reconfigure CF and CIP to include the new nodes and initialize.

3. Reconfigure the currently running Resource Database so it will recognize the new node.
4. Initialize the Resource Database on the new node.
5. Verify that the StartingWaitTime is sufficient for the new node, and modify this parameter if necessary.

The diagram below shows these steps as a flow chart.

Figure 3.2 Adding a new node



The sections that follow describe each step in more detail.

3.6.1 Backing up the Resource Database

Before you add a new node to the Resource Database, you should first back up the current configuration. The backup will be used later to help initialize the new node. It is also a safeguard. If the configuration process is unexpectedly interrupted by a panic or some other serious error, then you may need to restore the Resource Database from the backup.

Note

The configuration process itself should not cause any panics. However, if some non-PRIMECLUSTER software panics or if the SF causes a power cycle because of a CF cluster partition, then the Resource Database configuration process could be so severely impacted that a restoration from the backup would be needed.

Note

The restoration process requires all nodes in the cluster to be in single user mode.

Since the Resource Database is synchronized across all of its nodes, the backup can be done on any node in the cluster where the Resource Database is running. The steps for performing the backup are as follows:

1. Log onto any node where the Resource Database is running with system administrator authority.
2. Run the command `clbackuprdb` to back the Resource Database up to a file.

```
/etc/opt/FJSVcluster/bin/clbackuprdb -f file
```

For example:

```
# /etc/opt/FJSVcluster/bin/clbackuprdb -f /mydir/backup_rdb
```

`clbackuprdb` stores the Resource Database as a compressed tar file. Thus, in the above example, the Resource Database would be stored in `/mydir/backup_rdb.tar.*`. * represents the extension of the type of tar compression (Z or gz).

Make sure that you do not place the backup in a directory whose contents are automatically deleted upon reboot (for example, `/tmp`).



Note

The hardware configuration must not change between the time a backup is done and the time that the restore is done. If the hardware configuration changes, you will need to take another backup. Otherwise, the restored database would not match the actual hardware configuration, and new hardware resources would be ignored by the Resource Database.

3.6.2 Reconfiguring the Resource Database

After you have backed up the currently running Resource Database, you will need to reconfigure the database to recognize the new node. Before you do the reconfiguration, however, you need to perform some initial steps.

After these initial steps, you should reconfigure the Resource Database. This is done by running the `clsetup` command on any of the nodes which is currently running the Resource Database. Since the Resource Database is synchronized across all of its nodes, the reconfiguration takes effect on all nodes. The steps are as follows:

1. Log in to any node where the Resource Database is running. Log in with system administrator authority.
2. If this node is not the same one where you made the backup, then copy the backup to this node. Then run the `clsetup` command with the `-a` and `-g` options to reconfigure the database. The configuration information of the resource database that is generated after the execution of `clsetup` command is also used in the configuration of new node of the resource database. That is why for the `--g` option, do not specify the directory (eg:/tmp) that is automatically deleted during the node restart-up.

```
# /etc/opt/FJSVcluster/bin/clsetup -a cfname -g file
```

`cfname` is the CF name of the new node to be added, and `file` is the name of the backup file without the `.tar.*` suffix. * represents the extension of the type of tar compression (Z or gz).

For example, suppose that you want to add a new node whose CF name is `fujj4` to a cluster.

If the backup file on an existing node is named `/mydir/rdb.tar.Z`, then the following command would cause the Resource Database to be configured for the new node:

```
# cd /etc/opt/FJSVcluster/bin/  
# ./clsetup -a fujj -g /mydir/rdb.tar.Z
```

If `clsetup` is successful, then you should immediately make a new backup of the Resource Database. This backup will include the new node in it. Be sure to save the backup to a place where it will not be lost upon a system reboot.

If an unexpected failure such as a panic occurs, then you may need to restore the Resource Database from an earlier backup. See "[3.6.5 Restoring the Resource Database](#)" for details.

3. To verify if the reconfiguration was successful, run the `clgettree` command. Make sure that the new node is displayed in the output for that command. If it is not present, then recheck the CIP configuration to see if it omitted the new node. If the CIP configuration is in error, then you will need to do the following to recover:
 - a. Correct the CIP configuration on all nodes. Make sure that CIP is running with the new configuration on all nodes.

- b. Restore the Resource Database from backup.
- c. Rerun the clsetup command to reconfigure the Resource Database.

3.6.3 Configuring the Resource Database on the new node

After the Resource Database has been reconfigured on the existing nodes in the cluster, you are ready to set up the Resource Database on the new node itself.

The first step is to verify the CIP configuration on the new node. The file `/etc/cip.cf` should reference the new node. The file should be the same on the new node as it is on existing nodes in the cluster.

You should also verify that the existing nodes in the cluster can ping the new node using the new node's CIP name. If the new node has multiple CIP subnetworks, then recall that the Resource Database only uses the first one that is defined in the CIP configuration file. If multiple CIPs are configured in the new node, test the connection of CIP that is initially configured.

After verifying that CIP is correctly configured and working, then you should do the following:

1. Log in to the new node with system administrator authority.
2. Copy the latest Resource Database backup to the new node. This backup was made in Step 2 of the second list in "[3.6.2 Reconfiguring the Resource Database](#)."
3. Run the command `clsetup` with the `-s` option.

```
/etc/opt/FJSVcluster/bin/clsetup -s file
```

File is the name of the backup file. If we continue our example of adding `fuji4` to the cluster and we assume that the backup file `rdb.tar.Z` was copied to `/mydir`, then the command would be as follows:

```
# /etc/opt/FJSVcluster/bin/clsetup -s /mydir/rdb.tar.Z
```

If the new node unexpectedly fails before the `clsetup` command completes, then you should execute the `clinitreset` command. After `clinitreset` completes, you must reboot the node and then retry the `clsetup` command which was interrupted by the failure.

If the `clsetup` command completes successfully, then you should run the `clgettree` command to verify that the configuration has been set-up properly. The output should include the new node. It should also be identical to output from `clgettree` run on an existing node.

If the `clgettree` output indicates an error, then recheck the CIP configuration. If you need to change the CIP configuration on the new node, then you will need to do the following on the new node after the CIP change:

- a. Run `clinitreset`.
- b. Reboot.
- c. Rerun the `clsetup` command described above.

3.6.4 Adjusting StartingWaitTime

After the Resource Database has successfully been brought up in the new node, then you need to check if the `StartingWaitTime` used in 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. Refer to "[3.5 Startup synchronization](#)" for further information.

3.6.5 Restoring the Resource Database

The procedure for restoring the Resource Database is as follows:

1. Copy the file containing the Resource Database to all nodes in the cluster.
2. Log in to each node in the cluster and shut it down with the following command:

```
#!/usr/sbin/shutdown -y -i0
```

3. Reboot each node to single user mode with the following command:

```
{0} ok boot -s
```

Note

The restore procedure requires that all nodes in the cluster must be in single user mode.

4. Reboot each node to single user mode with the following command:

```
# mountall -l
# zfs mount -a
```

5. Reboot each node to single user mode with the following command:

```
# clrestorerdb -f file
```

file is the backup file with the .tar.Z suffix omitted.

For example, suppose that a restoration was being done on a two-node cluster consisting of nodes fuji2 and fuji3, and that the backup file was copied to */mydir/backup_rdb.tar.Z* on both nodes. The command to restore the Resource Database on fuji2 and fuji3 would be as follows:

```
fuji2# cd /etc/opt/FJSVcluster/bin/
fuji2# ./clrestorerdb -f /mydir/backup_rdb.tar.Z
fuji3# cd /etc/opt/FJSVcluster/bin/
fuji3# ./clrestorerdb -f /mydir/backup_rdb.tar.Z
```

6. After Steps 1 through 5 have been completed on all nodes, then reboot all of the nodes with the following command:

```
#!/usr/sbin/shutdown -y -i6
```

Chapter 4 GUI administration

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

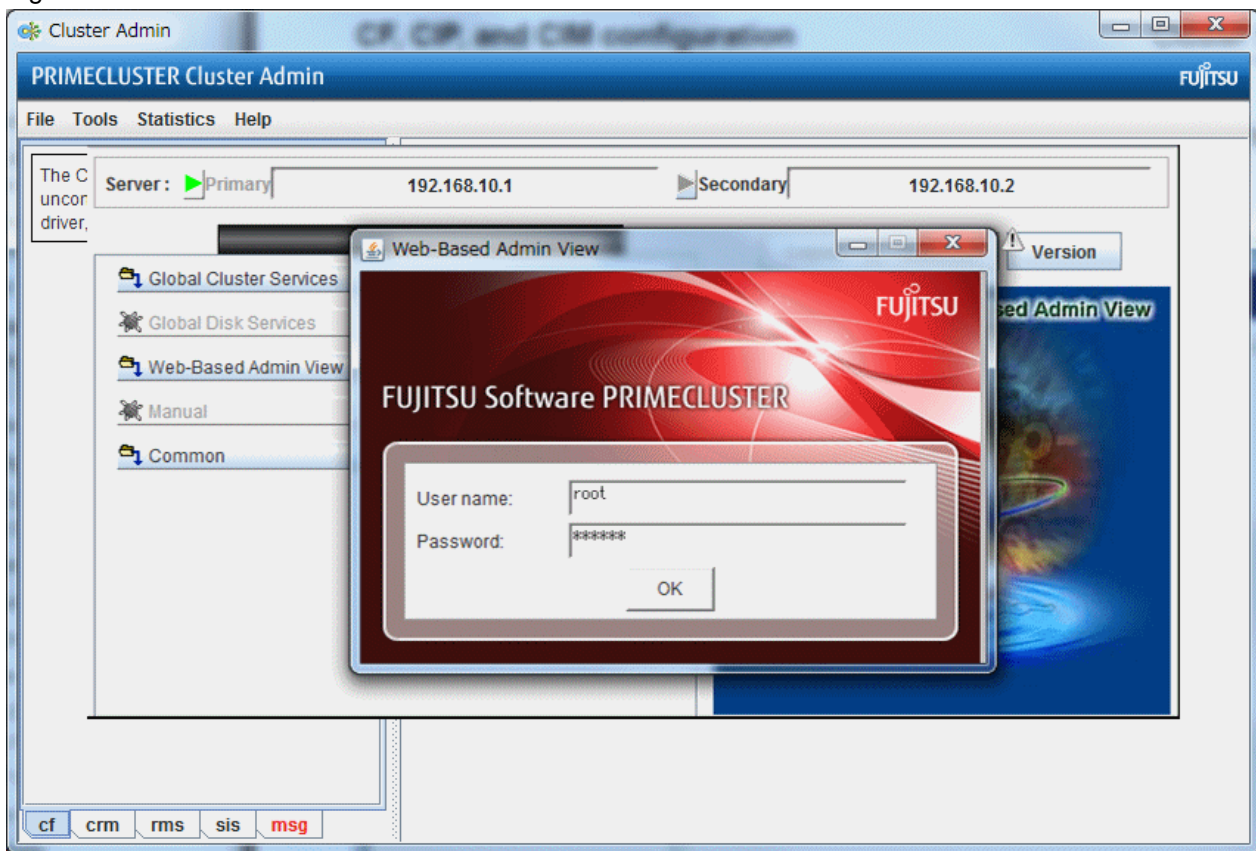
4.1 Overview

CF administration is done by means of the Cluster Admin GUI. The following sections describe the CF Cluster Admin GUI options.

4.2 Starting Cluster Admin GUI and logging in

By starting the screen from the shortcut of the Java application (PRIMECLUSTER Web-Based Admin View Startup), the Web-Based Admin View main screen below is displayed.

Figure 4.1 Main 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.

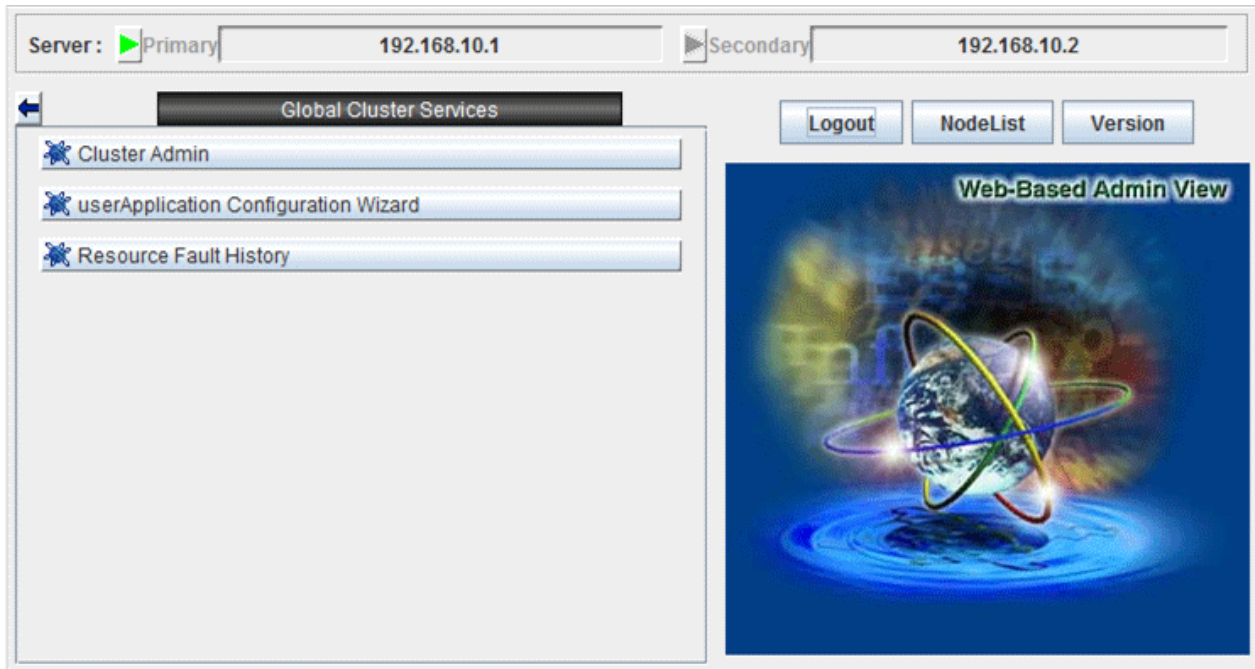
With the root privileges, you can perform all actions including configuration, administration and viewing tasks. With administrative privileges, you can view as well as execute commands but cannot make configuration changes.

Point

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

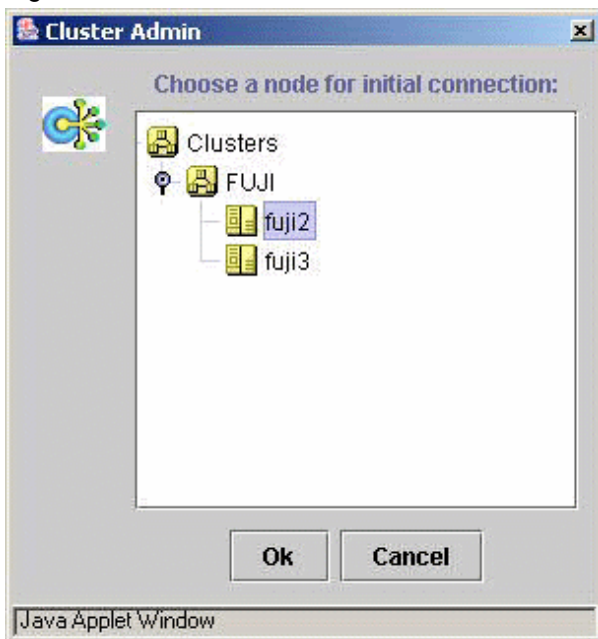
Click on the [Global Cluster Services] button and the screen below is displayed.

Figure 4.2 Cluster Admin start-up window



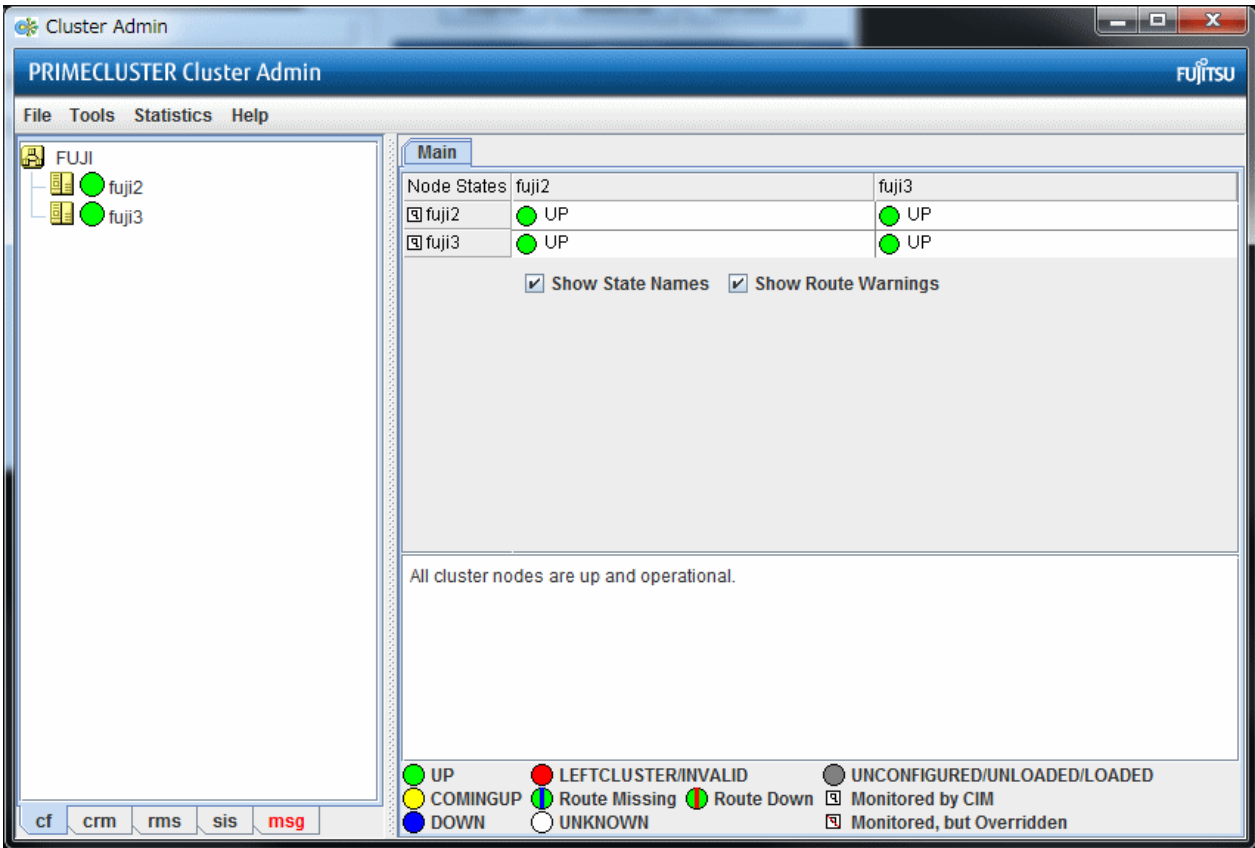
Click on the [Global Cluster Services] button and then the node selection screen below is displayed.

Figure 4.3 Initial connection choice window



Select a node and click [Ok]. The Cluster Admin main window below is displayed.

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. The tab for RMS will appear as *rms&pcs* when PCS is installed and as *rms* in configurations where PCS is not installed.

Point

Both of the terms **UP** and **Online** are represented by green circles. These terms describe the same state and are interchangeable.

4.3 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 table below lists the local states.

Table 4.1 Local states

CF state	Description
UNLOADED	The node does not have a CF driver loaded.

CF state	Description
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 node is unconfigured.

The table below lists the remote states.

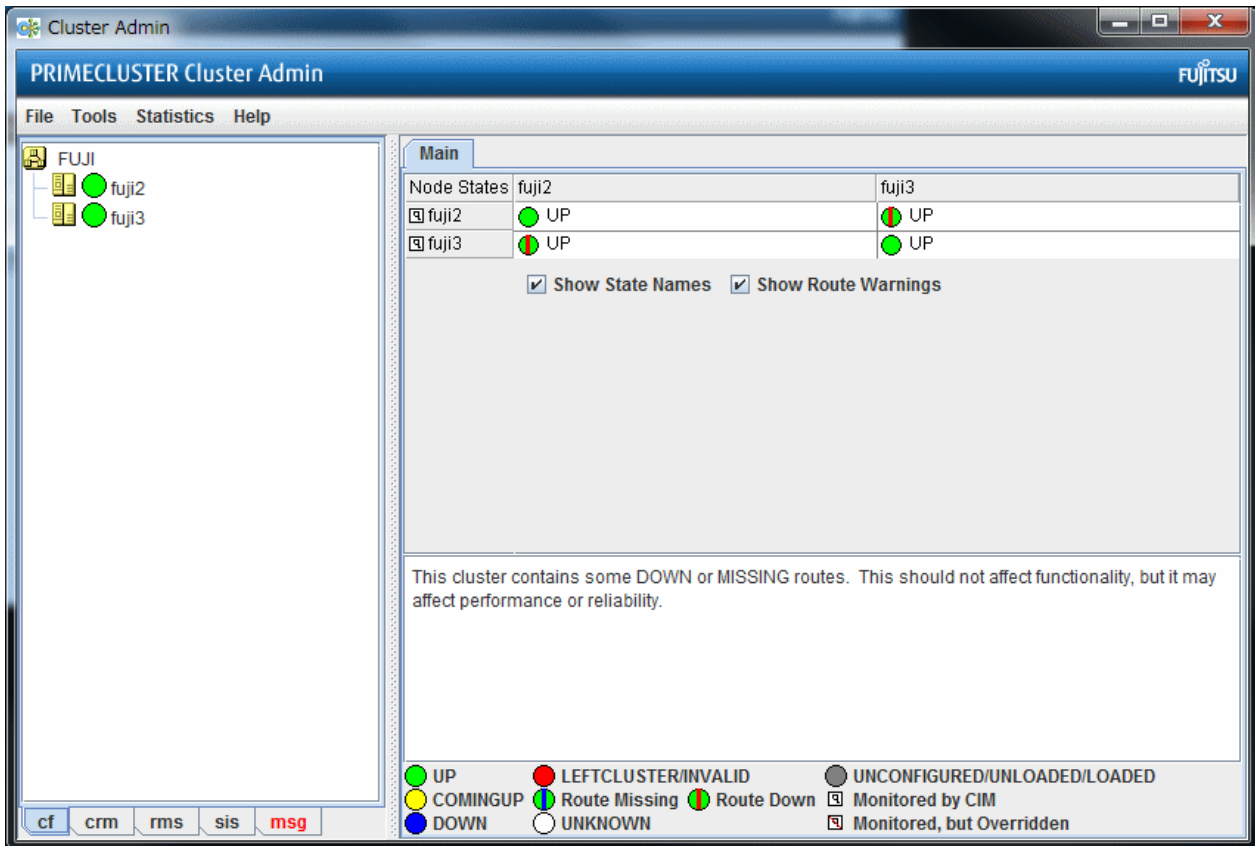
Table 4.2 Remote states

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.

4.4 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 below).

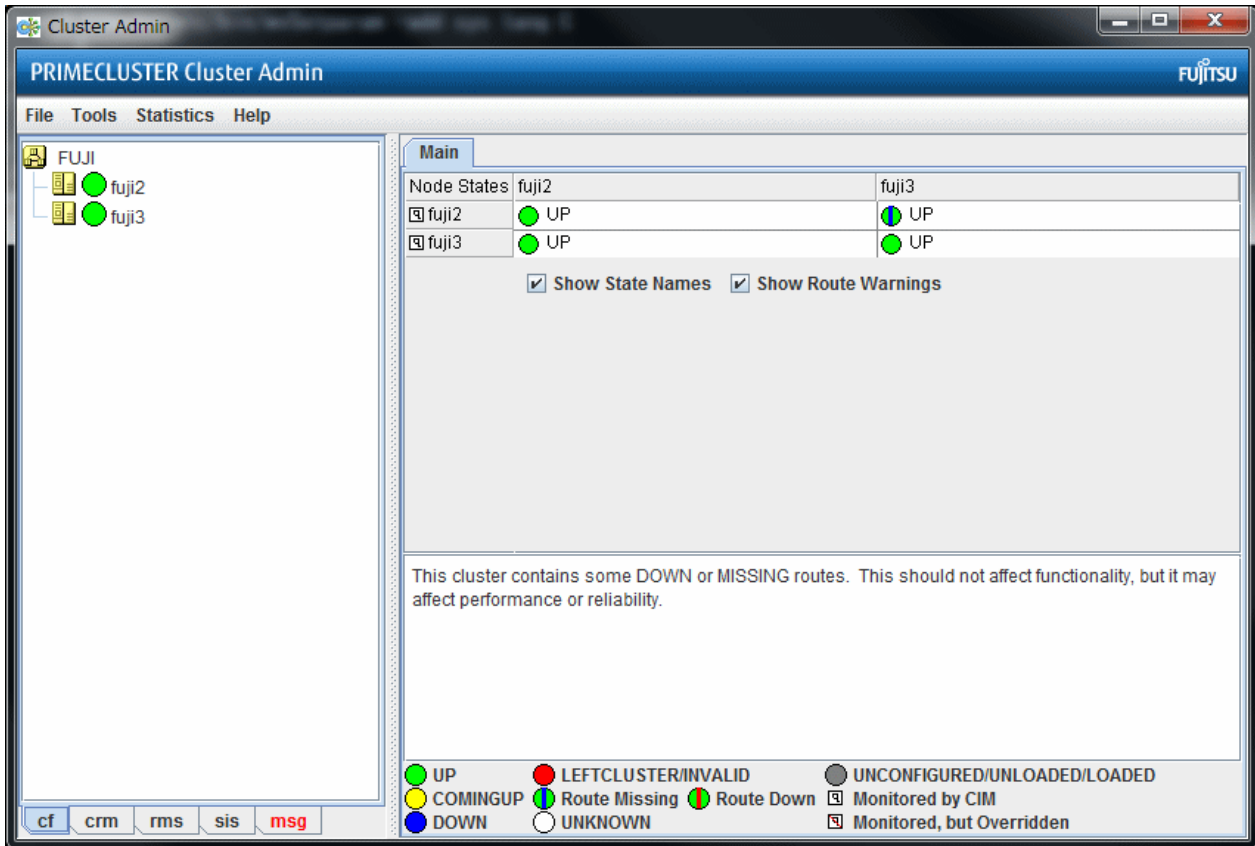
Figure 4.5 CF route DOWN



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 screen below).

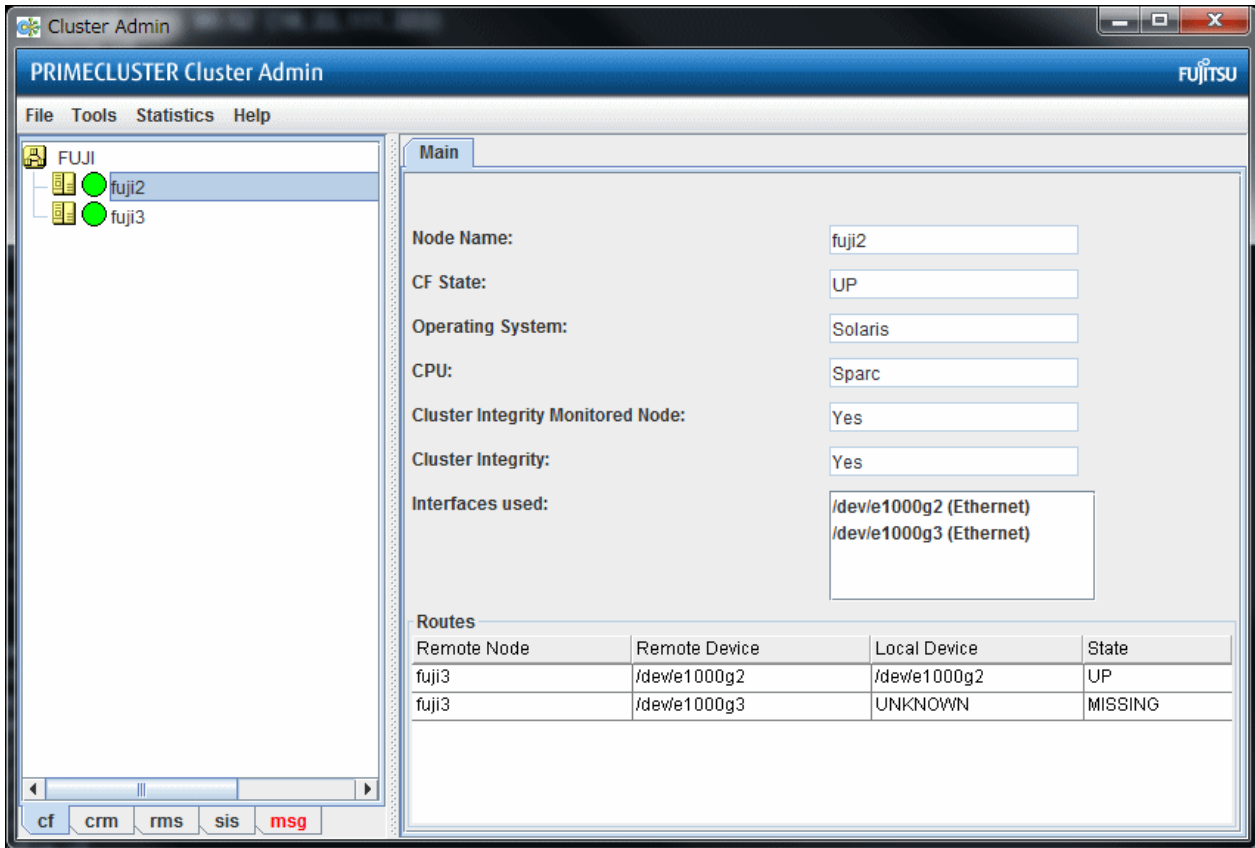
Figure 4.6 CF interface missing



In the example screen seen above, 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 e1000g3 interface on fuji3 (see below).

Figure 4.7 CF route table



4.5 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.) The panel displayed is similar below.

Figure 4.8 CF node information

Main

Node Name:

CF State:

Operating System:

CPU:

Cluster Integrity Monitored Node:

Cluster Integrity:

Interfaces used:

Routes

Remote Node	Remote Device	Local Device	State
fuji3	/dewhme1	/dewhme1	UP
fuji3	/dewhme3	/dewhme3	UP
fuji3	/dewip0	/dewip0	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.

4.6 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.

An example of the topology table is shown below.

Figure 4.9 CF topology table

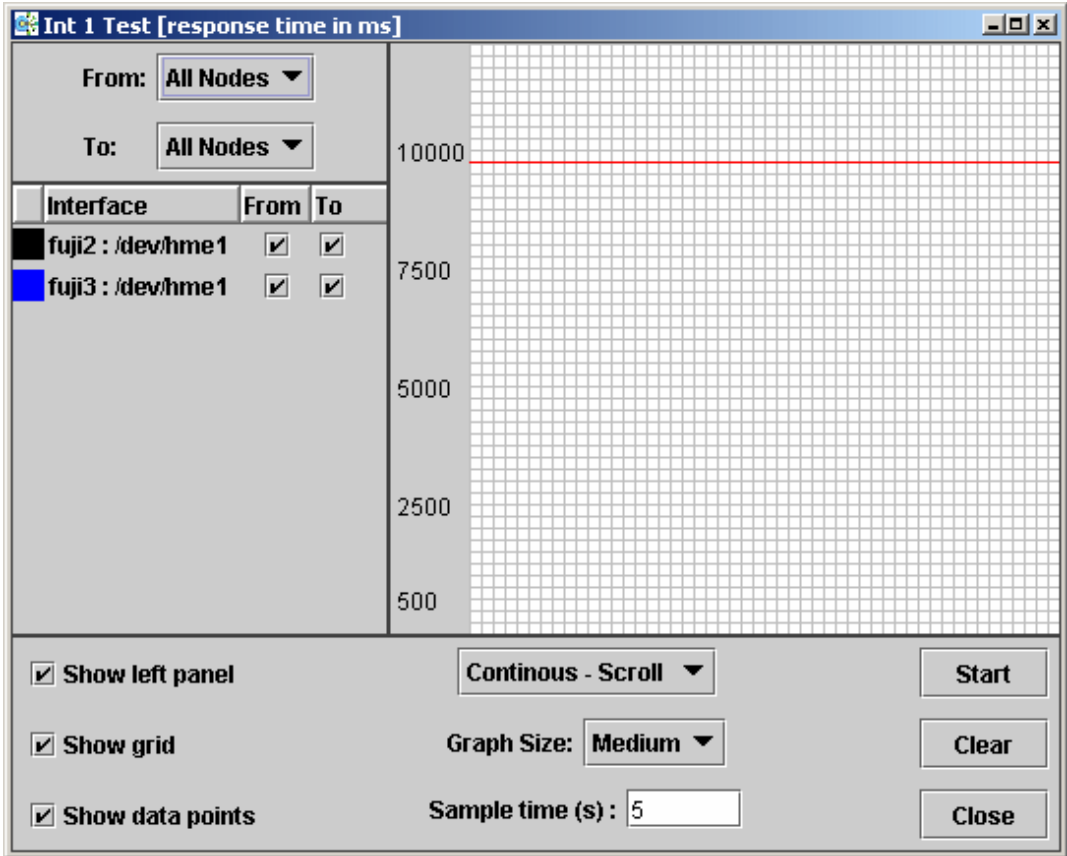
FUJI: Topology			
FUJI	Full Interconnects		
	<input checked="" type="checkbox"/> Int 1 <input type="button" value="Test"/>	<input checked="" type="checkbox"/> Int 2 <input type="button" value="Test"/>	<input checked="" type="checkbox"/> Int 3 <input type="button" value="Test"/>
fuji2 *	<input checked="" type="checkbox"/> /dev/hme1	<input checked="" type="checkbox"/> /dev/hme3	<input checked="" type="checkbox"/> /dev/ip0
fuji3 *	<input checked="" type="checkbox"/> /dev/hme1	<input checked="" type="checkbox"/> /dev/hme3	<input checked="" type="checkbox"/> /dev/ip0

This table displays the physical connectivity of the nodes in this cluster. This information is current as of (10:17:39 AM) and will not update. Nodes marked with a * will only show interfaces that are configured.

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 below).

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.

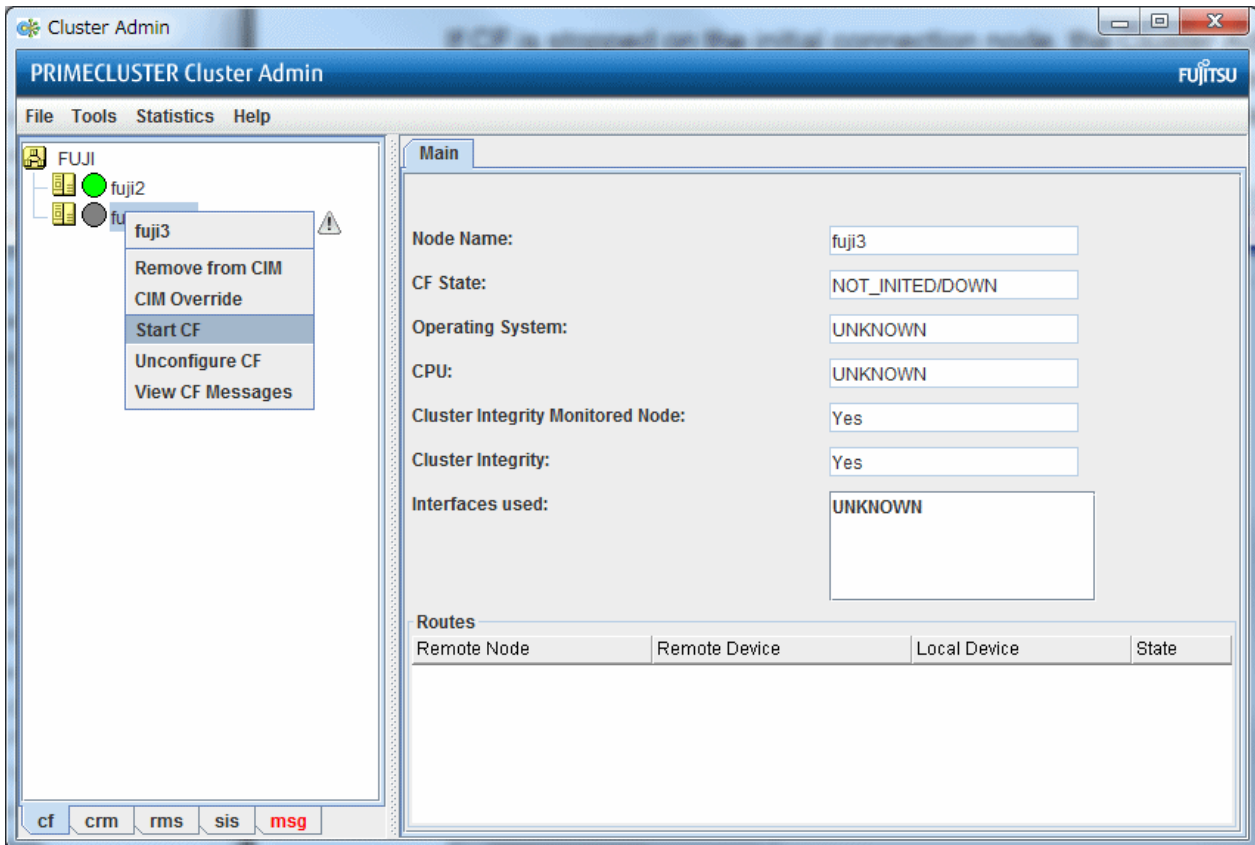
For best results, the Response Time Monitor should be run at times when CF traffic is relatively light, and the CF nodes are only lightly loaded.

4.7 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 figure below shows the content-sensitive menu pop-up when you select [Start CF].

You can also go to the [Tools] from the pull-down menu and select either [Start CF] or [Stop CF] (not shown). A pop-up listing all the nodes where CF may be started or stopped will appear. You can then select the desired node to carry out the appropriate action. The following screen shows the content-sensitive menu pop-up when you select Start CF.

Figure 4.11 Starting CF



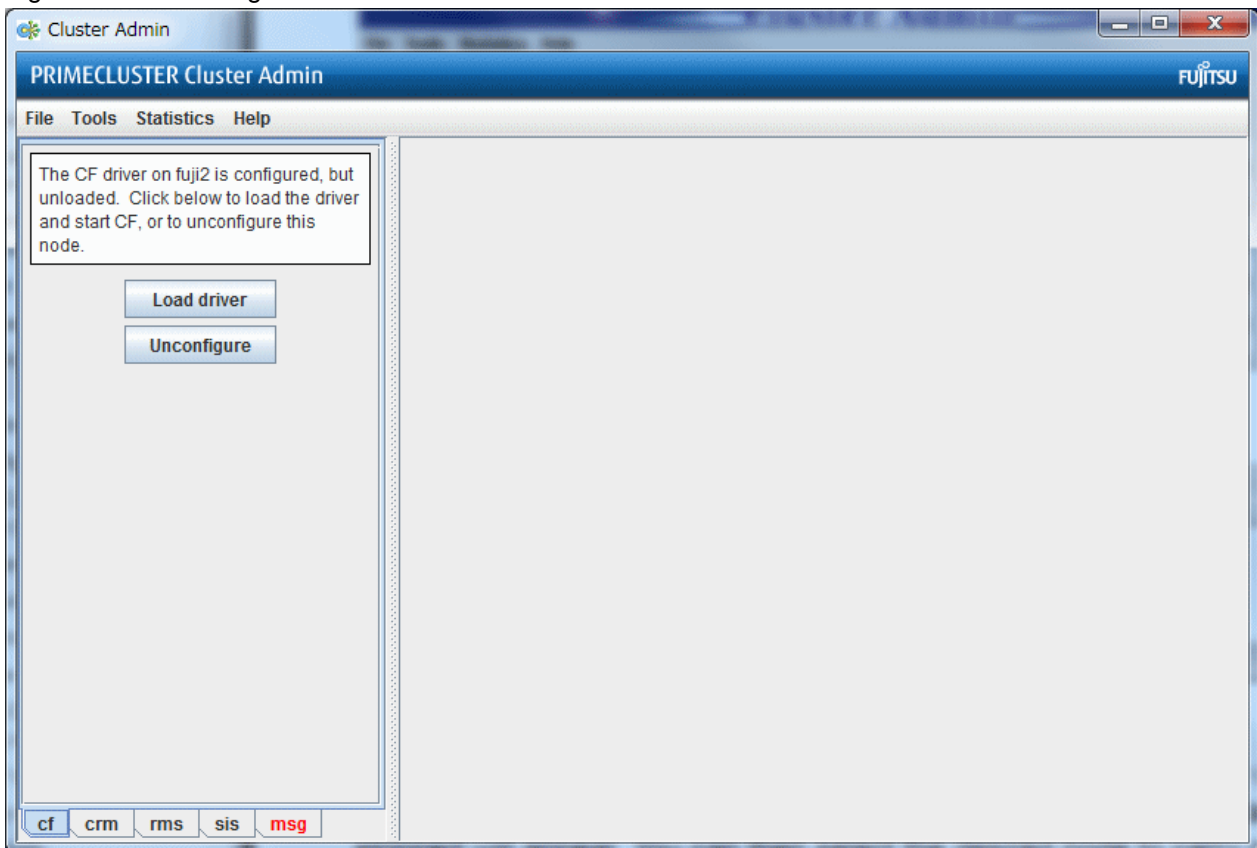
The CF GUI gets its list of CF nodes from the node you selected as the initial connection node that is shown in "[Figure 4.3 Initial connection choice window](#)". If CF is not up and running on the initial connection node, then the CF GUI will not display the list of nodes in the tree in the left panel.

Because of this, when 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.

4.7.1 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.

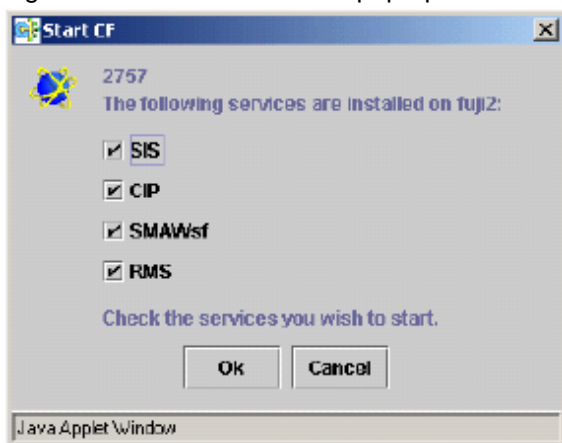
Figure 4.12 CF configured but not loaded



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

The Start CF services pop-up 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.

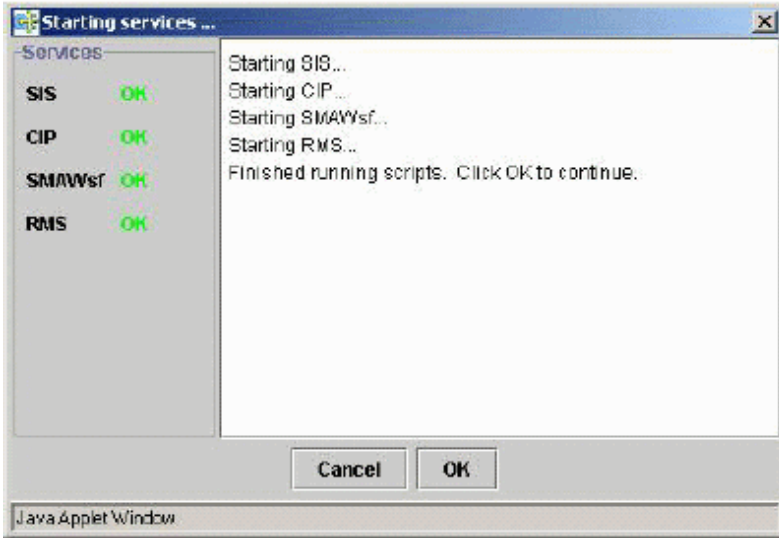
Figure 4.13 Start CF services pop-up



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 pop-up appears with the results of each service start operation (see below).

Figure 4.14 Start CF services status window



After all the services have been started, click on the [OK] button to return to the Cluster Admin main window.

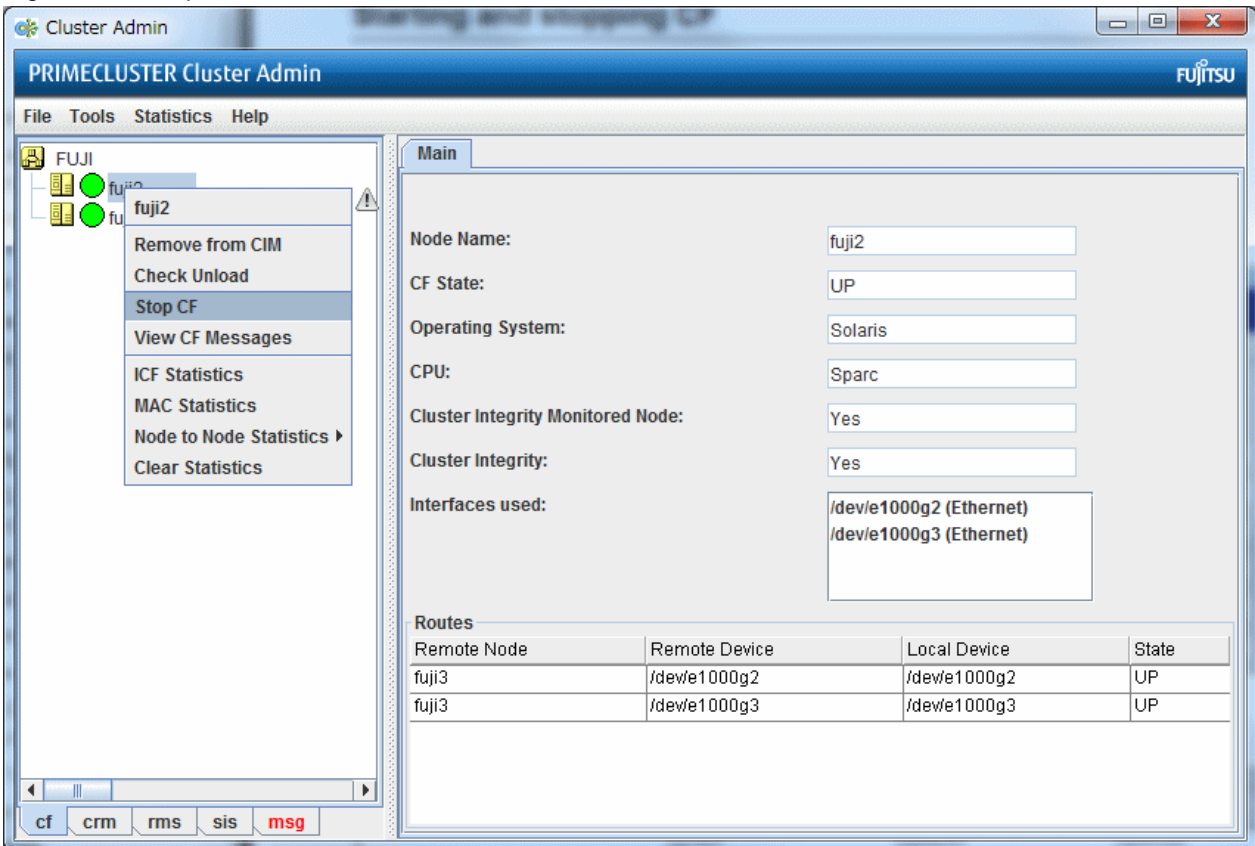
If you click the [Cancel] button, startup process of the CF services are canceled when these CF services are not yet started.

The already started CF services remain running.

4.7.2 Stopping CF

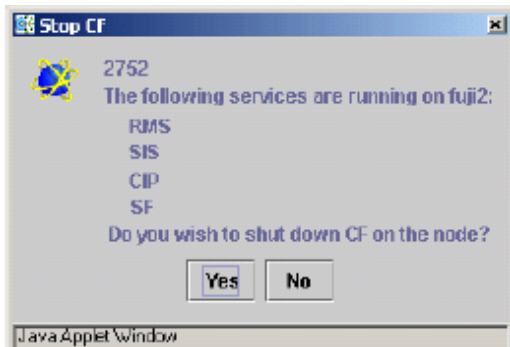
Right-click on a CF node name and select [Stop CF] (see below).

Figure 4.15 Stop CF



A confirmation pop-up appears (see below). Choose [Yes] to continue.

Figure 4.16 Stopping CF



Before stopping CF, all services that run over CF on that node should first be shut down. When you invoke [Stop CF] from the GUI, it will use the CF dependency scripts to see what services are still running. It will print out a list of these in a pop-up and ask you if you wish to continue. If you do continue, it will then run the dependency scripts to shut down these services.

If you click the [Cancel] button, stop process of the CF services are canceled when these CF services are not yet stopped.

The already stopped CF services remain suspended.



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.

4.8 Marking nodes DOWN

This section explains the procedure on how to change the state to DOWN of the node during recovery from LEFTCLUSTER state.

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

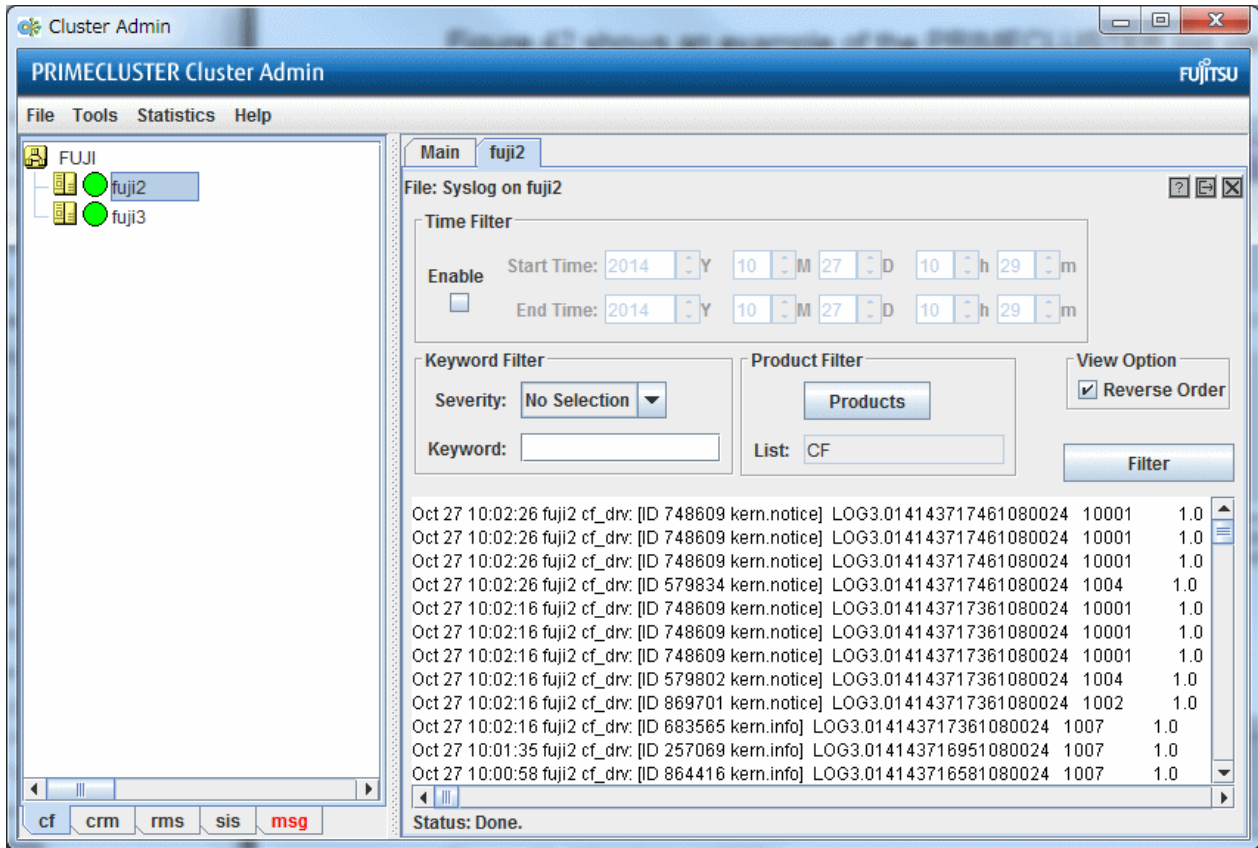
4.9 Using PRIMECLUSTER log viewer

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 in the [Product Filter] window pane on the PRIMECLUSTER log viewer screen.

Figure 4.17 PRIMECLUSTER log viewer



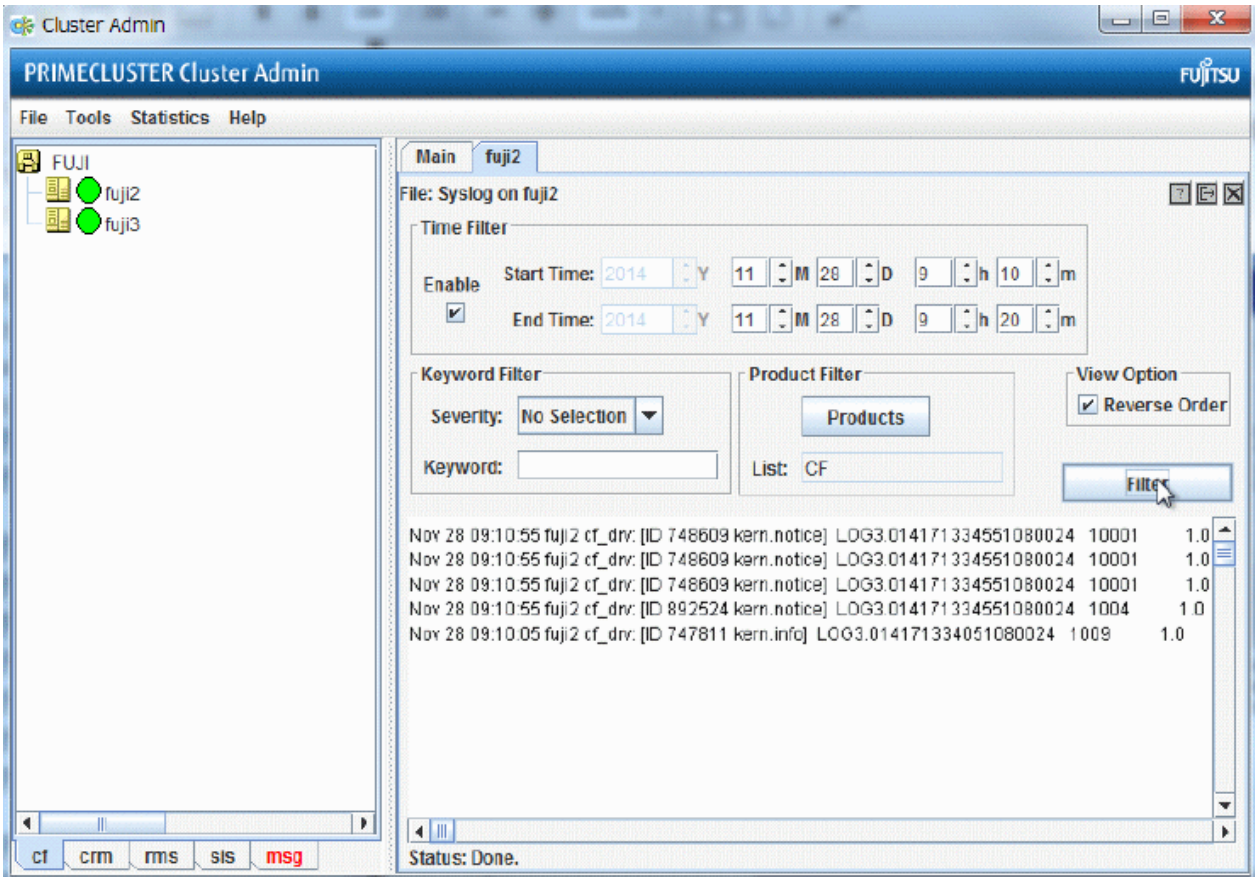
The syslog messages appears in the right-hand panel. If you click on the [Detach] button on the tab, then the syslog window appears as a separate window.

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

4.9.1 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 (see Figure below).

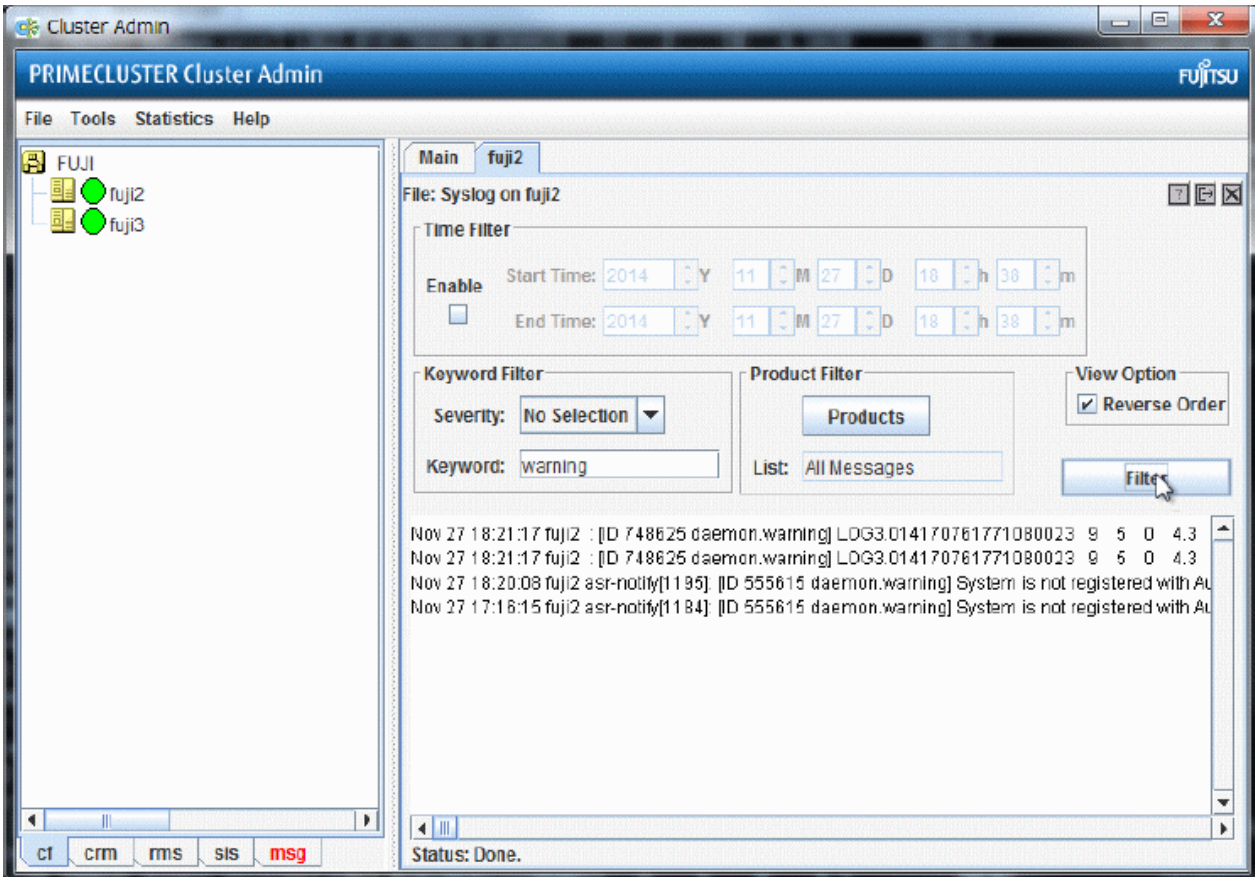
Figure 4.18 Search based on date/time



4.9.2 Search based on keyword

To perform a search based on a keyword, enter a keyword and click on the [Filter] button (see Figure below).

Figure 4.19 Search based on keyword



4.9.3 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 on the table below and click on the [Filter] button. The figure below shows the log for a search based on severity level.

Figure 4.20 Search based on severity

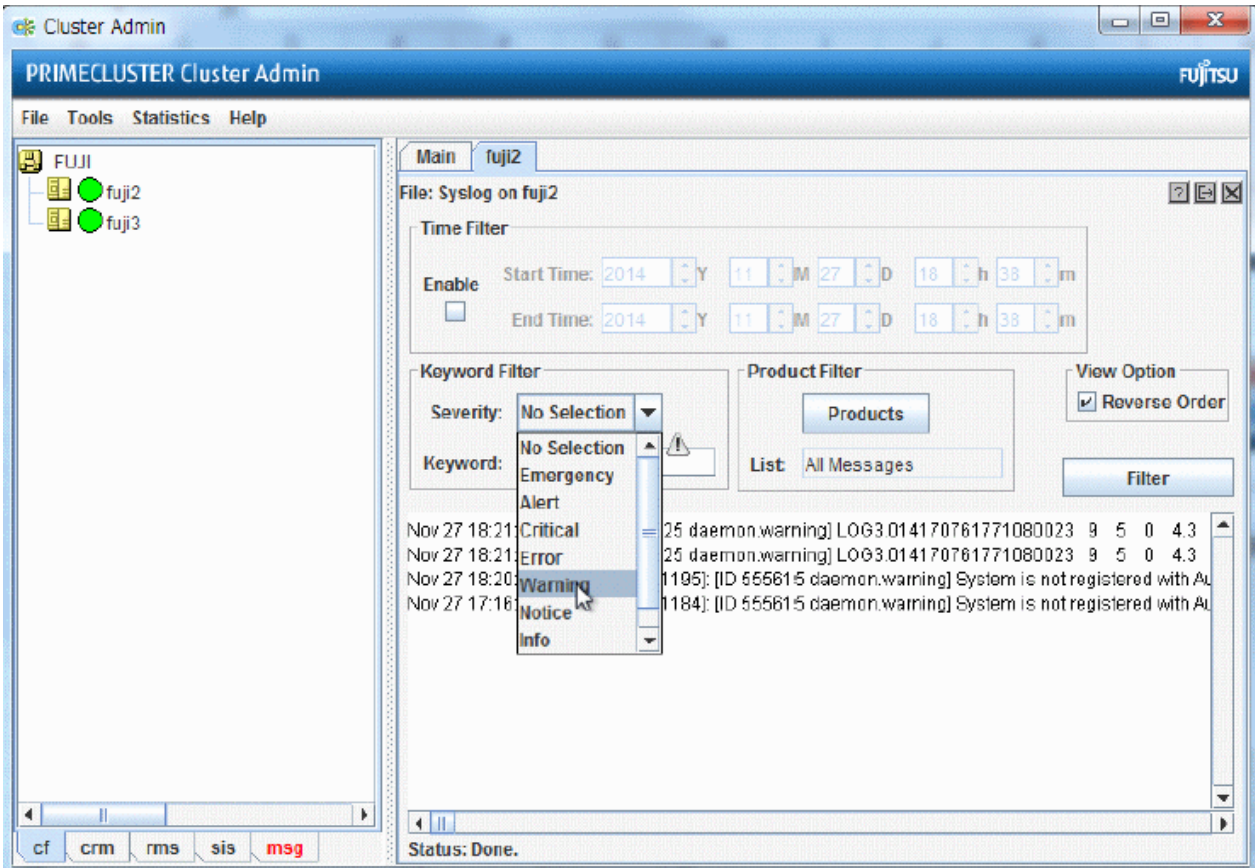


Table 4.3 PRIMECLUSTER log viewer severity levels

Severity level	Severity description
[Emergency]	Systems cannot be used
[Alert]	Immediate action is necessary
[Critical]	Critical condition
[Error]	Error condition
[Warning]	Warning condition
[Notice]	Normal but important condition
[Info]	For information
[Debug]	Debug message

4.10 Displaying statistics

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

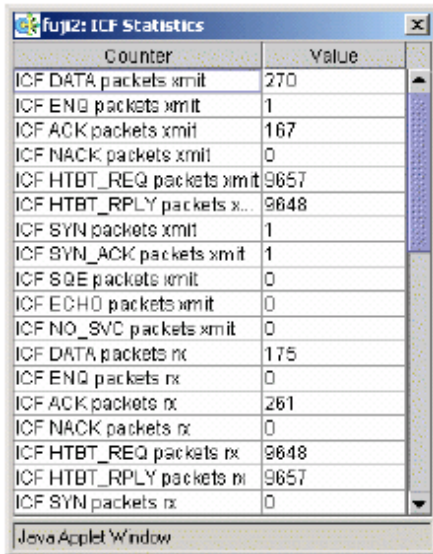
- ICF
- MAC
- Node to Node

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.

Display of ICF Statistics is shown below.

Figure 4.21 ICF statistics

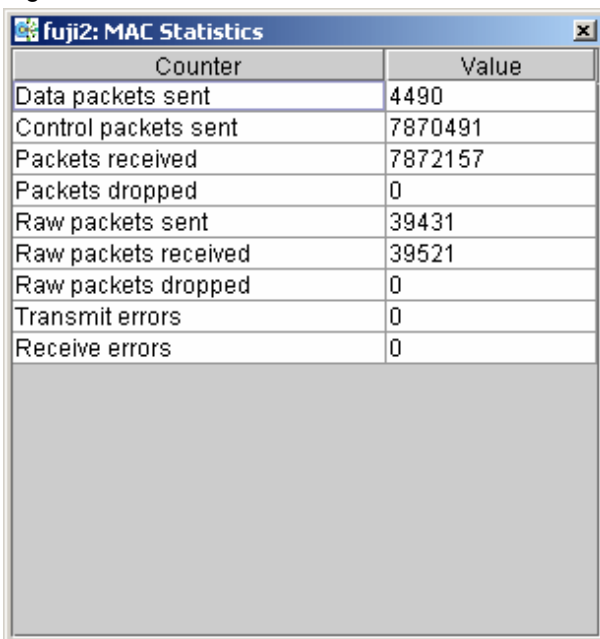


Counter	Value
ICF DATA packets xmit	270
ICF ENQ packets xmit	1
ICF ACK packets xmit	167
ICF NACK packets xmit	0
ICF HTBT_REQ packets xmit	9657
ICF HTBT_RPLY packets xmit	9648
ICF SYN packets xmit	1
ICF SYN_ACK packets xmit	1
ICF SQE packets xmit	0
ICF ECHO packets xmit	0
ICF NO_SVC packets xmit	0
ICF DATA packets rx	175
ICF ENQ packets rx	0
ICF ACK packets rx	261
ICF NACK packets rx	0
ICF HTBT_REQ packets rx	9648
ICF HTBT_RPLY packets rx	9657
ICF SYN packets rx	0

JavaApplet Window

Display of MAC Statistics is shown below.

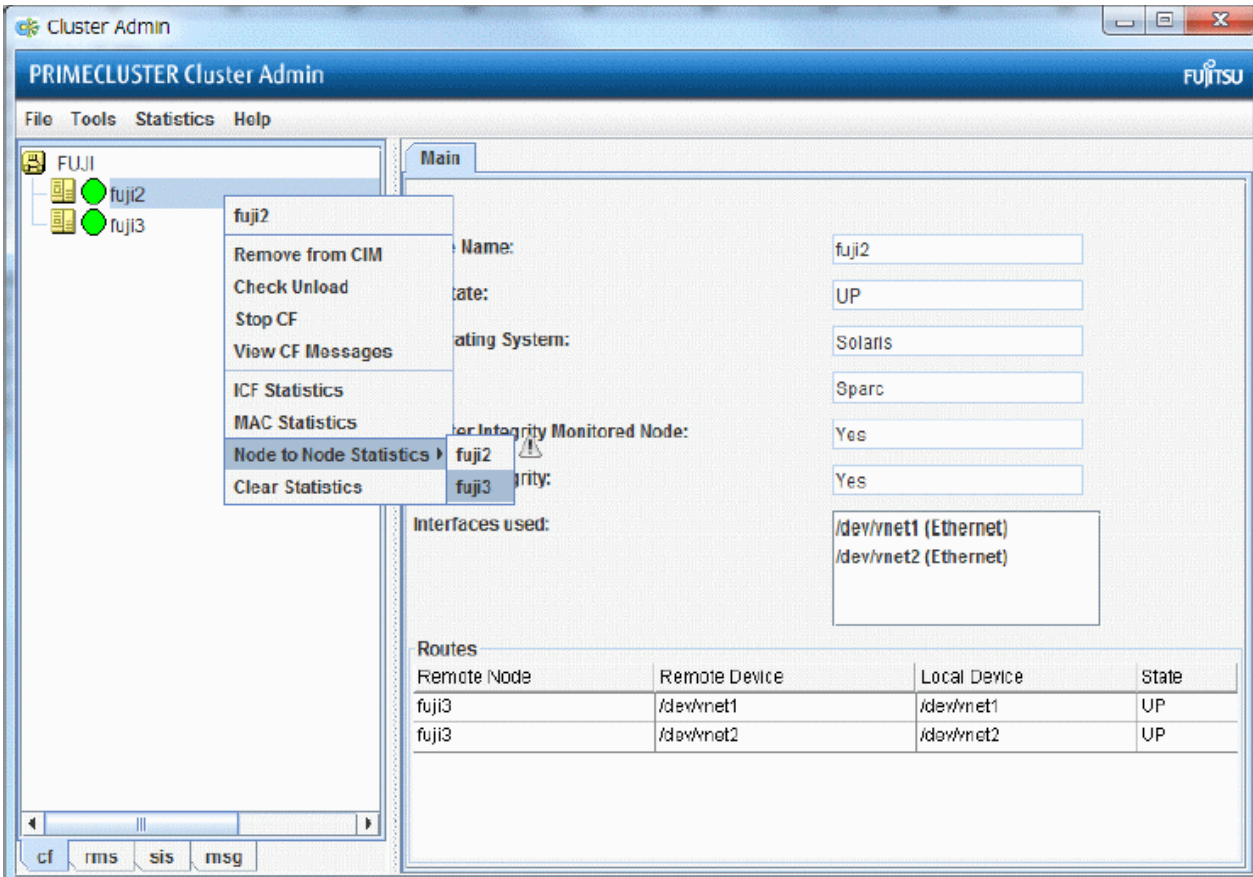
Figure 4.22 MAC statistics



Counter	Value
Data packets sent	4490
Control packets sent	7870491
Packets received	7872157
Packets dropped	0
Raw packets sent	39431
Raw packets received	39521
Raw packets dropped	0
Transmit errors	0
Receive errors	0

To display node to node statistics, choose Node to Node Statistics and click on the desired node (see below).

Figure 4.23 Selecting a node for node to node statistics



The window for Node to Node Statistics appears (see Figure below).

Figure 4.24 Node to Node statistics

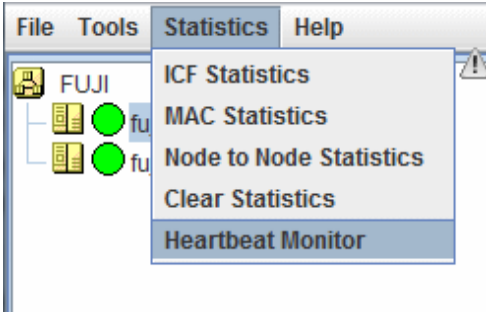
Counter	Value
ICF DATA packets xmit	310
ICF ENQ packets xmit	1
ICF ACK packets xmit	187
ICF NACK packets xmit	0
ICF HTBT_REQ packets xmit	11319
ICF HTBT_RPLY packets xmit	11310
ICF SYN packets xmit	1
ICF SYN_ACK packets xmit	1
ICF SQE packets xmit	0
ICF ECHO packets xmit	0
ICF NO_BVC packets xmit	0
ICF DATA packets rx	195
ICF ENQ packets rx	0
ICF ACK packets rx	301
ICF NACK packets rx	0
ICF HTBT_REQ packets rx	11310
ICF HTBT_RPLY packets rx	11319
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.

4.11 Heartbeat monitor

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

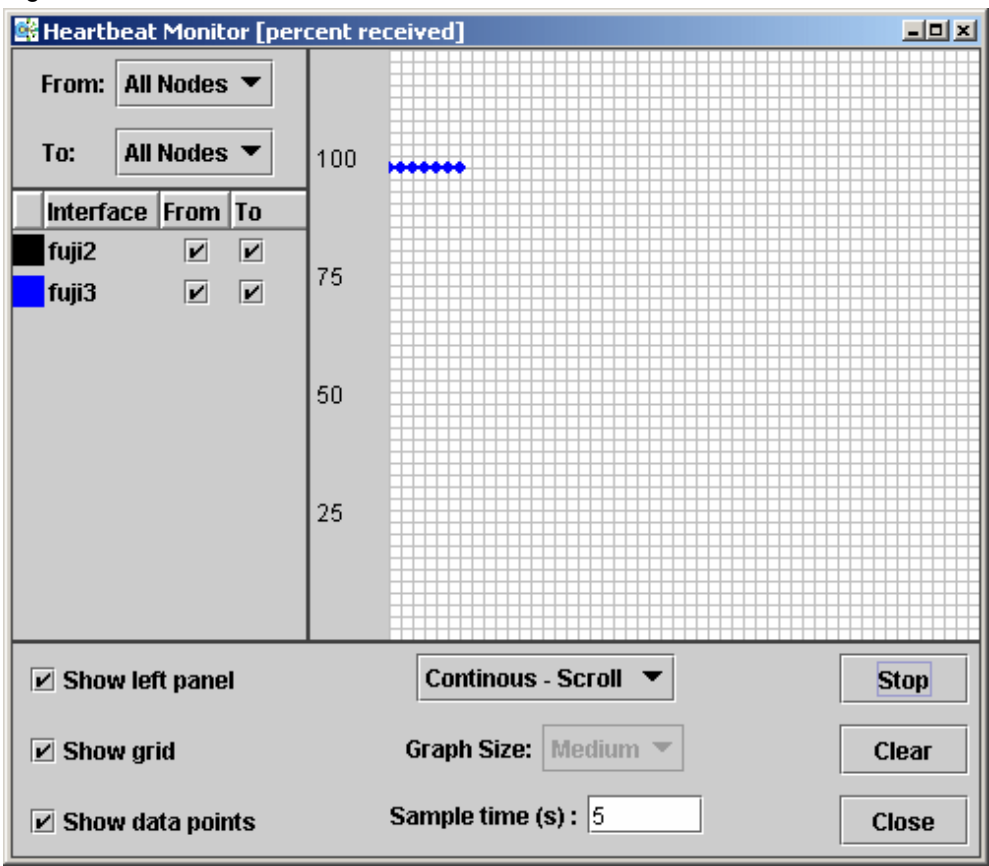
Figure 4.25 Selecting the Heartbeat monitor



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 4.26 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.

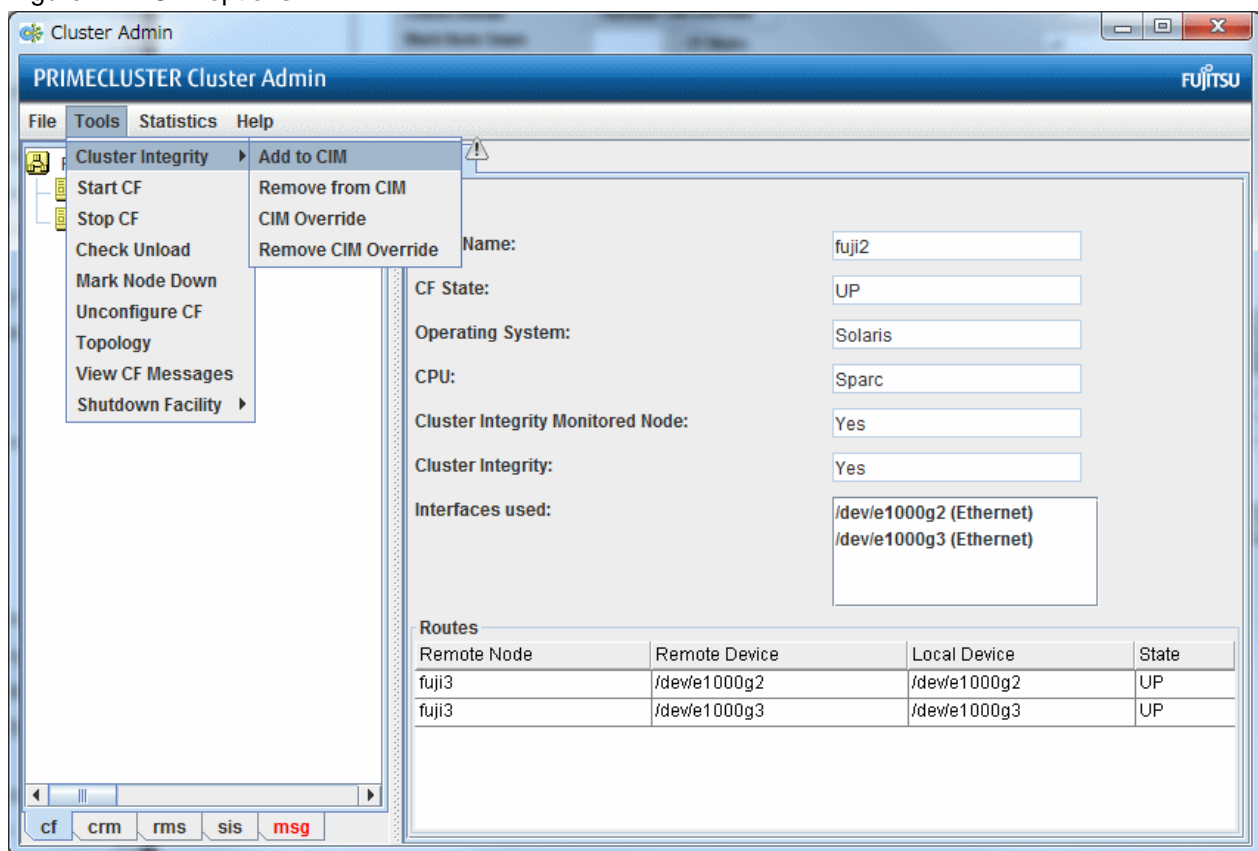
The controls in the bottom panel are as follows:

- 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.

4.12 Adding and removing a node from CIM

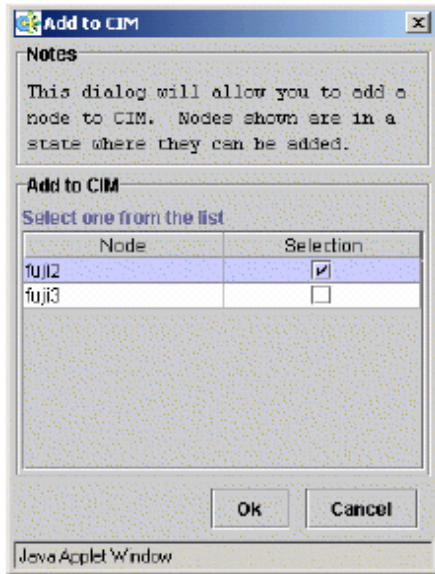
To add a node to CIM, click on the [Tools] from the pull-down menu and select [Cluster Integrity] - [Add to CIM] from the expandable pull-down menu (see below).

Figure 4.27 CIM options



The Add to CIM pop-up display appears. Choose the desired CF node and click on [Ok] (see below).

Figure 4.28 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 "[2.2 Cluster Integrity Monitor \(CIM\)](#)" for more details on CIM.

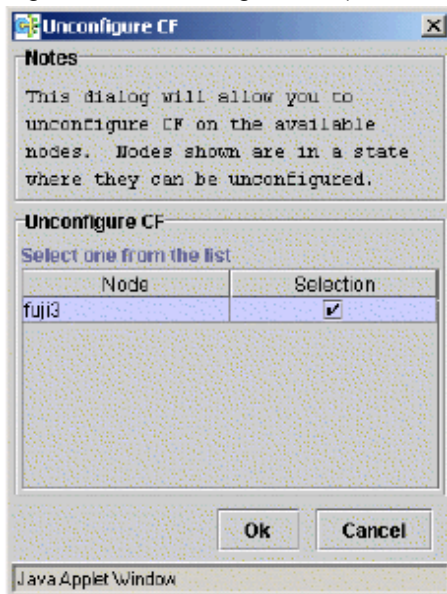
4.13 Unconfigure CF

You can delete the CF configuration of each node one at a time. To remove all of the nodes, delete the configuration of all the nodes other than the local node. Then, delete the configuration of the local node (the node in "Figure 4.3 Initial connection choice window").

1) Deleting the configuration other than the local node.

1. Refer to "[4.7.2 Stopping CF](#)" and then stop CF which is on the node to be deleted.
2. When selecting [Delete CF configuration] of [Tools] from the pull-down menu, the following pop-up [Unconfigure CF] will be displayed:

Figure 4.29 Unconfigure CF (Other than local node)



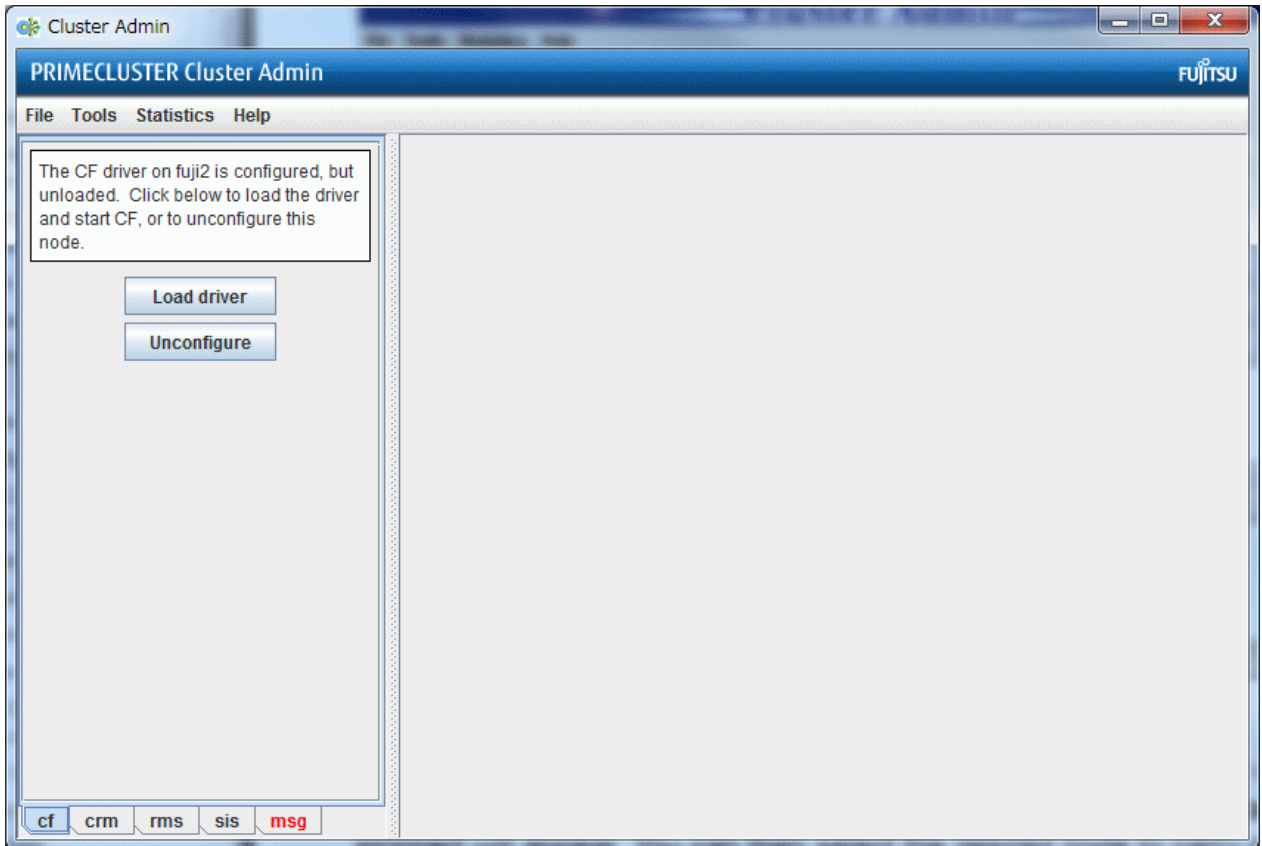
3. Check the checkbox of the CF node that is to be unconfigured and then click [Ok].

The unconfigured node will be removed from the cluster. Other cluster nodes will display the removed node as DOWN until it is rebooted.

2) Deleting the configuration of the local node.

1. Refer to "4.7.2 Stopping CF" and then stop CF which is on the node to be deleted.
2. Click [Delete configuration] when the screen below is displayed.

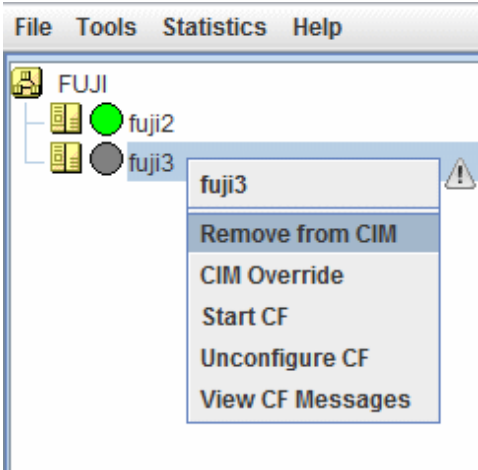
Figure 4.30 Deleting CF configuration (Local node)



4.14 CIM Override

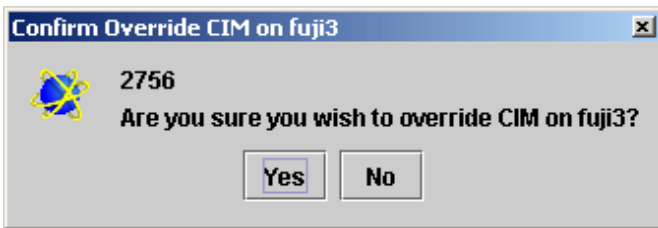
[The CIM Override option causes a node to be ignored when determining a quorum. A node cannot be overridden if its CF state is UP. To select a node for CIM Override, right-click on a node and choose [CIM Override] (see below).

Figure 4.31 CIM Override



A confirmation pop-up appears (see below).

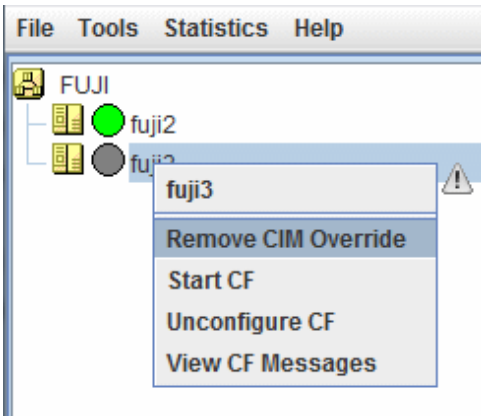
Figure 4.32 CIM Override confirmation



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 4.33 Remove CIM Override



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

Chapter 5 LEFTCLUSTER state

This chapter defines and describes the LEFTCLUSTER 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.
- Node in kernel debugger
A node is left in the kernel debugger for too long and heartbeats are missed.
- Entering the firmware monitor OBP
Will cause missed heartbeats and will result in the LEFTCLUSTER state.
- 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.

These commands will run the `rc` scripts that will allow CF to be cleanly shut down on that node. If you run the `reboot` command, the `rc` scripts are 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 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.

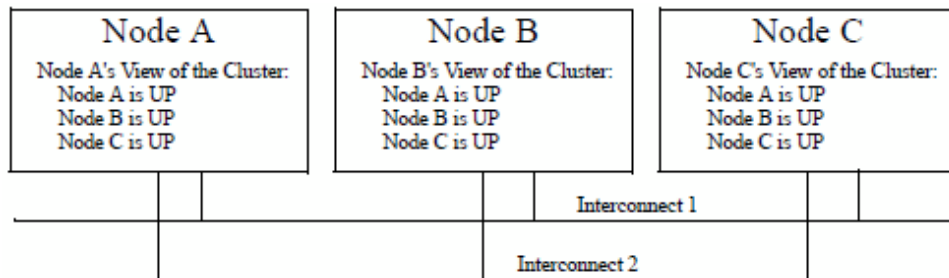
5.1 Description of the LEFTCLUSTER state

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 the figure below.

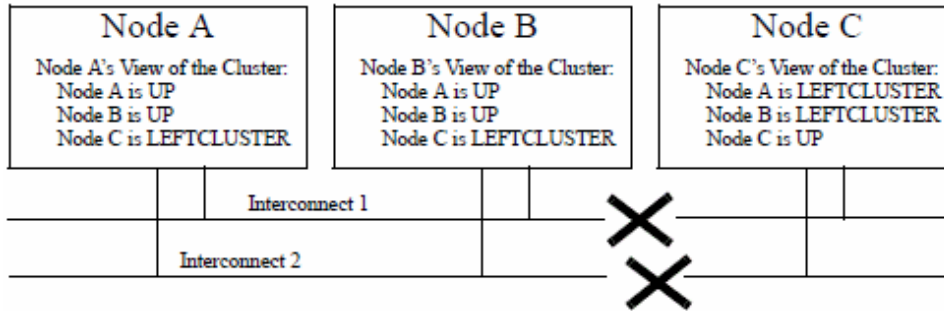
Figure 5.1 Three-node cluster with working connections



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 figure below.

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 say 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 see, 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 "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.

5.2 Recovering from LEFTCLUSTER

If SF is not running on all 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. While the offending node is down, use Cluster Admin to log on to one of the surviving nodes in the cluster. Invoke the CF GUI 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:

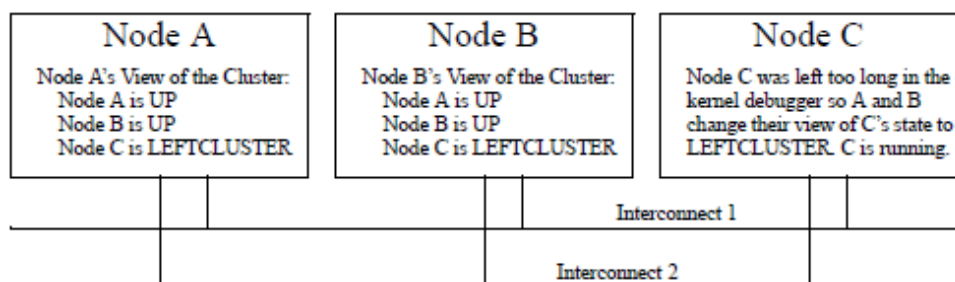
```
#cftool -k
```

3. Bring the offending node back up. It will rejoin the cluster as part of the reboot process.

5.2.2 Caused by staying in the kernel debugger too long

In the figure below, 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.
2. While Node C is down, 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. Bring Node C back up. It will rejoin the cluster as part of its reboot process.

5.2.3 Caused by a cluster partition

A cluster partition refers to 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.

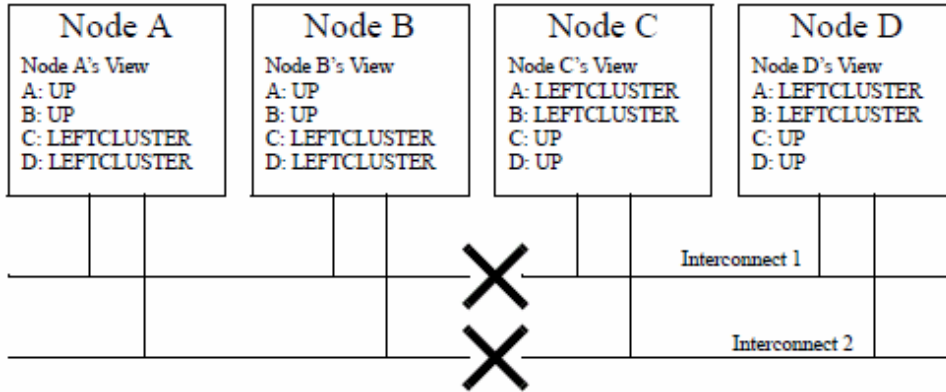
If SF runs without problem on all nodes in the cluster SF will solve any cluster partition issues. However, if SF is not running properly or the SF forced shutdown operation fails it becomes necessary to manually recover the situation.

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. While the nodes are down, 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.
4. Fix the network break so that connectivity is restored between all nodes in the cluster.
5. Bring the nodes back up. They will rejoin the cluster as part of their reboot process.

For example, consider the figure below.

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. While Nodes C and D are down, 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 Nodes C and D as DOWN.

```
#cftool -k
```

4. Fix the interconnect break on Interconnect 1 and Interconnect 2 so that both sub-clusters will be able to communicate with each other again.
5. Bring Nodes C and D back up.

5.2.4 Caused by reboot

The LEFTCLUSTER state may occur because a particular node (called the offending node) has been rebooted improperly. If a node is rebooted using the normal reboot commands like `init(1M)` or `shutdown(1M)`, the LEFTCLUSTER state should not occur.

The LEFTCLUSTER state will occur if you reboot the offending node with commands like `uadmin(1M)` or `reboot(1M)`. In this case the procedure to clear the LEFTCLUSTER state is as follows:

1. Make sure the offending node is rebooted in multi-user mode.
2. Use Cluster Admin to log on to one of the surviving nodes in the cluster. Invoke the CF GUI by selecting [Mark Node Down] from the [Tools] pull-down menu. Mark the offending node as DOWN.
3. The offending node will rejoin the cluster automatically.

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 "4.6 Displaying the topology table." in "Chapter 4 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 nodes in the cluster and then analyzing the results. `cfconfig -l` causes the driver to be loaded by pushing its modules on all possible Ethernet devices on the system, regardless of whether or not they are configured for use with CF. This allows CF pings to be done on all Ethernet devices on all nodes in the cluster. Thus, all Ethernet devices show up in the topology table.

`cfconfig -L` causes CF to push CF modules only on the Ethernet devices which are configured for use with CF.

The `-L` option offers several advantages. On systems with large disk arrays, it means that CF driver load time is reduced. On SPARC Enterprise M-series systems with dynamic hardware reconfiguration, Ethernet controllers that are not used by CF can be moved more easily between partitions. Because of these advantages, the `rc` scripts that load CF use the `-L` option.

However, the `-L` option restricts the devices which are capable of sending or receiving CF pings to only configured devices. CF has no knowledge of other Ethernet devices on the system. Thus, when the topology table displays devices for a node where CF has been loaded with the `-L` option, it only displays devices that have been configured for CF.

It is possible that a running cluster might have a mixture of nodes where some were loaded with `-l` and others were loaded with `-L`. In this case, the topology table would show all Ethernet devices for nodes loaded with `-l`, but only CF configured devices for nodes loaded with `-L`. The topology table indicates which nodes have been loaded with the `-L` option by adding an asterisk (*) after the node's name.

When a cluster is totally unconfigured, the CF Wizard will load the CF driver on each node using the `-l` option. This allows all devices on all nodes to be seen. After the configuration is complete, the CF Wizard will unload the CF driver on the newly configured nodes and reload it with `-L`. This means that if the topology table is subsequently invoked on a running cluster, only configured devices will typically be seen.

If you are using the CF Wizard to add a new CF node into an existing cluster where CF is already loaded, then the Wizard will load the CF driver on the new node with `-l` so all of its devices can be seen. However, it is likely that the already configured nodes will have had their CF drivers loaded with `-L`, so only configured devices will show up on these nodes.

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 below.

Table 6.1 Basic layout for the CF topology table

FUJI	Full interconnects		Partial interconnects		Unconnected devices
	Int 1	Int 2	Int 3	Int 4	
fuji2	hme0 hme2	hme1	hme3	hme5	hme4 hme6
fuji3	hme0	hme2	missing	hme1	
fuji4	hme1	hme2	hme3	missing	hme4

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 for the CF 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 cluster name 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 for the CF 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 for the CF topology table," Interconnect 1 is a full interconnect. It is attached to hme0 and hme2 on fuji2. On fuji3, it is attached to hme0, and on fuji4, it is attached to hme1.

Since CF runs over Ethernet devices, the hmen devices in "Table 6.1 Basic layout for the CF 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 for the CF 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 table below. 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 table.

Table 6.2 Topology table with check boxes shown

FUJI	Full interconnects		Partial interconnects		Unconnected devices
	<input checked="" type="checkbox"/> Int1	<input checked="" type="checkbox"/> Int2	<input type="checkbox"/> Int3	<input type="checkbox"/> Int 4	
fuji2	<input checked="" type="checkbox"/> hme0 <input type="checkbox"/> hme2	<input checked="" type="checkbox"/> hme1	<input type="checkbox"/> hme3	<input type="checkbox"/> hme5	<input type="checkbox"/> hme4 <input type="checkbox"/> hme6
fuji3	<input checked="" type="checkbox"/> hme0	<input checked="" type="checkbox"/> hme2	missing	<input type="checkbox"/> hme1	
fuji4	<input checked="" type="checkbox"/> hme1	<input checked="" type="checkbox"/> hme2	<input type="checkbox"/> hme3	missing	<input type="checkbox"/> hme4

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 the above table, fuji2 has both hme0 and hme2 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 hme0 or hme2 to be selected for fuji2. In the above example, if hme2 were selected for fuji2, then hme0 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.

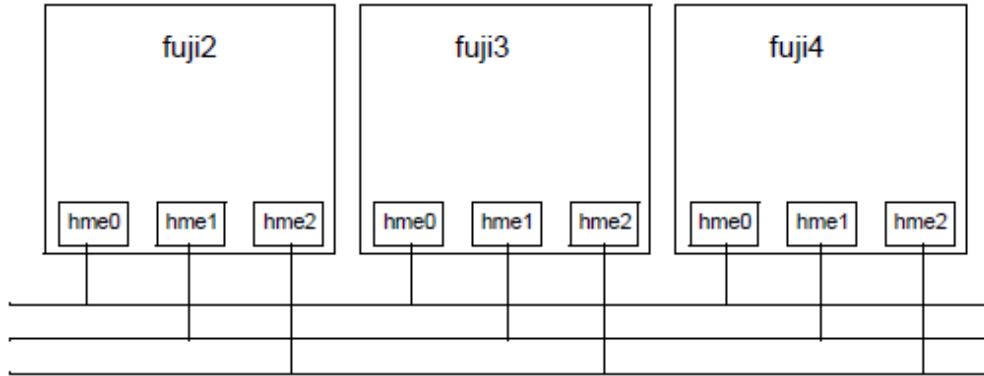
6.3 Examples

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 figure above is shown in the table below.

Table 6.3 Topology table for 3 full interconnects

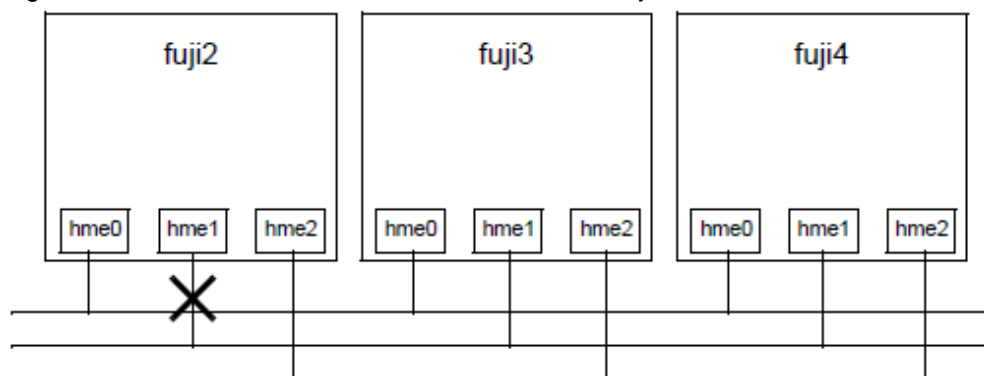
FUJI	Full interconnects		
	Int 1	Int 2	Int 3
fuji2	hme0	hme1	hme2
fuji3	hme0	hme1	hme2
fuji4	hme0	hme1	hme2

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 hme1 has been broken.

Figure 6.2 Broken Ethernet connection for hme1 on fuji2



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

Table 6.4 Topology table with broken Ethernet connection

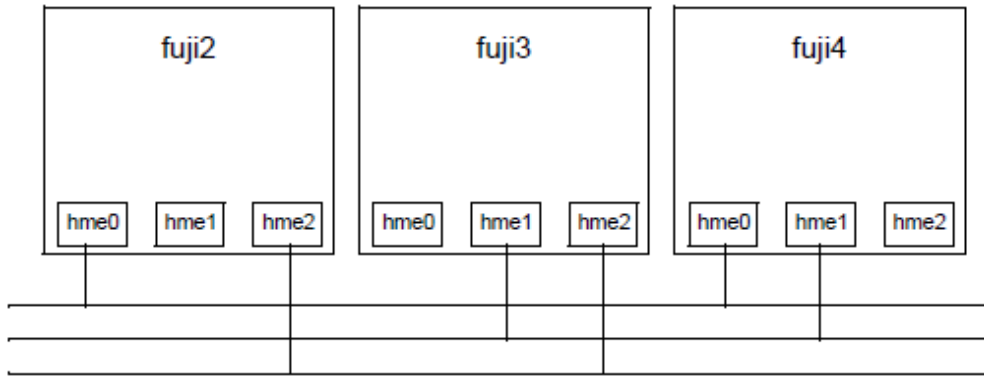
FUJI	Full interconnects		Partial interconnects	Unconnected devices
	Int 1	Int 2	Int 3	
fuji2	hme0	hme2	missing	hme1
fuji3	hme0	hme2	hme1	
fuji4	hme0	hme2	hme1	

In "Table 6.4," hme1 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 "Table 6.3" and "Table 6.4". In "Table 6.3," for example, all hme1 devices were on Int 2. In "Table 6.4," the hme1 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 table.

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 figure above is shown in the table below.

Table 6.5 Topology table with no full interconnects

FUJI	Partial interconnects			Unconnected devices
	Int 1	Int 2	Int 3	
fuji2	hme0	missing	hme2	hme1
fuji3	missing	hme1	hme2	hme0
fuji4	hme0	hme1	missing	hme2

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 your local customer-support service representative.

7.1 Overview

PRIMECLUSTER Shutdown Facility (SF) provides features to forcibly stop nodes where an error occurs in the cluster.

The SF is made up of the following four major components:

- Shutdown Daemon (SD)

Shutdown Daemon monitors the state of cluster nodes, collects the states, and provides the interface to request a shutdown manually or automatically. Also, a processing to solve a cluster partition state is performed.

- Shutdown Agents (SA)

Shutdown Agents guarantee the shutdown of other nodes. Though Shutdown Agents are attached to SF products, they vary depending on the architecture of cluster nodes for SF installation destination. SF provides the feature to shut down a node whether RMS is operating or not for each product of PRIMECLUSTER service layer.

- Monitoring Agent (MA)

Monitoring Agent monitors the state of cluster nodes by taking advantage of the characteristic of hardware and detects the node down immediately. When errors occur on other nodes, such as sudden system panic or power off, the error is reported to SF. Also, the feature as Shutdown Agents (SA) is provided to shut down nodes which errors occur.

- `sdtool(1M)` command

The `sdtool(1M)` is the command to provide I/F of Shutdown Daemon.

PRIMECLUSTER Shutdown Facility has the following features:

- It is possible to detect the shut down or cluster nodes immediately (monitoring agent).
- It is possible to shut down cluster nodes whether RMS is running or not.
- It is possible to shut down cluster nodes from any components of PRIMECLUSTER service layer.

The first section explains the initial installation of SF products. The second and subsequent sections explain the configuration setup of SF. The last section explains the changes which need to be added to other products.

7.2 Configuring SF

This section describes how to configure SF.

7.2.1 Setting procedure before configuring SF

Before creating the configuration file, take the following procedure:

1. Checking system requirements

Take the following steps to check the system requirements.

- Establishing user feature requirements
- Monitoring the cluster node and establishing the usage of SF for the shutdown

- Determining the most suitable shutdown agent
2. Planning the configuration of the shutdown agent

The following are necessary to plan the configuration of the shutdown agent.

- A node monitored by SF
- Shutdown agent

The configuration design is determined depending on the environment where SF is used, and depending on requirements specified for the node.

The monitoring by SF should also be determined in detail.
(Shutdown agent or flow of usage, for example)

3. Defining Shutdown Agent (SA) to be configured in SF

When all the cluster interconnects are disabled due to a hung or a failure of the node that configures the cluster system, the node should be forcibly stopped. For this reason, SA should be defined. When defining SA, check the hardware model and configure the suitable shutdown agent for the model.

7.2.2 Configuration file of SF

Information

See the format of the configuration file described below as a reference. How to configure the shutdown agent is described in "[7.5 Configuring the Shutdown Facility](#)."

Create a configuration file at `/etc/opt/SMAW/SMAWsf` directory and change the configuration file name to `rcsd.cfg`.

Here is the format of the configuration file.

```
CFName[ ,weight=weight ]
[ ,admIP=myadmIP ]:agent=SA_name ,timeout=SA_timeout { :agent=SA_name2 ,timeout=SA_timeout2 : }
```

`Weight` is an option keyword. If this option is not specified, assign the weight 1 to `rcsd`. This keyword is an option so that the existing configuration works without any change.

`admIP` is an option keyword. `myadmIP` is the IP address of the administrative LAN on the `CFName` machine. This keyword is also an option because this is the case of backward compatibility. However, the setting is required to avoid the inappropriate cluster partition. Set the address `myadmIP`, which does not exist on the CIP Interface.

`CFName` is the CF node name of the machine in the cluster.

`agent` and `timeout` are reserved words.

`SA_name` is the command name of the shutdown agent.

`SA_timeout` is the maximum time (second) during which the shutdown agent can work until a fault occurrence is detected.

The shutdown agent described first in the configuration file is the first priority SA. When the first priority SA sends a shutdown request and the response shows that the shutdown fails, the second priority SA sends a shutdown request. Requests and responses are continuously sent until a response shows the successful shutdown, or until a shutdown request is sent by all SAs. If SA fails to shut down the cluster node, operation by an operator is necessary, and the node remains in LEFTCLUSTER state.

The log file is stored in the `/var/opt/SMAWsf/log/rcsd.log`. Make sure to use the same `rcsd.cfg` file on all cluster nodes. This should be secured for administrative reasons.

The `rcsd.cfg.template` file exists in the `/etc/opt/SMAW/SMAWsf` directory. This file is the sample configuration file of the shutdown daemon configured by a dummy machine and an agent.

7.3 Available SAs

This section describes the following set of supported SAs:

- RCI-Remote Cabinet Interface
- XSCF-eXtended System Control Facility
- XSCF SNMP- XSCF Simple Network Management Protocol
- ALOM-Advanced Lights Out Management
- ILOM-Integrated Lights Out Manager
- KZONE-Oracle Solaris Kernel Zones
- ICMP-Internet Control Message Protocol

Table 7.1 Available SAs

SA	Name	Hardware
RCI	SA_pprcip, SA_pprcir	SPARC Enterprise M-series
XSCF	SA_xscfp, SA_xscfr, SA_rccu, SA_rccux	SPARC Enterprise M-series
XSCF SNMP	SA_xscfsnmpg0p, SA_xscfsnmpg1p, SA_xscfsnmpg0r, SA_xscfsnmpg1r, SA_xscfsnmp0r, SA_xscfsnmp1r	SPARC M10, M12
ALOM	SA_sunF	SPARC Enterprise T1000, T2000
ILOM	SA_ilomp, SA_ilomr	SPARC Enterprise T5120, T5220, T5140, T5240, T5440, SPARC T3, T4, T5, T7, S7 series
KZONE	SA_kzonep, SA_kzoner, SA_kzchkhost	SPARC M10, M12, SPARC T4, T5, T7, S7 series
ICMP	SA_icmp	SPARC M10, M12

7.3.1 RCI

RCI SA is the Shutdown Agent for SPARC Enterprise M-series.

Setup and configuration

Hardware setup of the RCI is performed only by qualified support personnel. Contact field engineers for more information. In addition, you can refer to the manual shipped with the unit and to any relevant PRIMECLUSTER Release Notices for more details on configuration.

Shutdown Agent

There are two kinds of RCI SAs:

- SA_pprcip-panics the node through RCI.
- SA_pprcir-resets the node through RCI.

The RCI log files are as follows:

```
/var/opt/SMAwsf/log/SA_pprcip.log
/var/opt/SMAwsf/log/SA_pprcir.log
```

How to check the RCI Monitoring Agent when an RCI error is detected

The RCI Monitoring Agent only discontinues monitoring the node when an RCI error is detected, so the monitoring function is not disrupted on the other nodes.

For how to restore the RCI Monitoring Agent, see "4.5 Error Messages" in "PRIMECLUSTER Messages." See below for how to check the RCI monitoring status.

How to check the RCI monitoring status

Check the Shutdown Facility on all the nodes as follows:

```
# /opt/SMAW/bin/sdtool -s
```

An RCI error is detected before the Shutdown Facility is started.

If InitFailed is displayed for Init State of the Agent SA_pprcip.so and SA_pprcir.so on any one of cluster nodes, an RCI transmission failure occurred between the node and the other nodes. This node is excluded from monitoring and elimination.

For example, an RCI transmission failure occurred between nodes, where the sdtool command was executed, and the other nodes in the following:

```
# /opt/SMAW/bin/sdtool -s
Cluster Host      Agent              SA State   Shut State   Test State   Init State
-----
node01           SA_pprcip.so      Idle       Unknown     Unknown     InitFailed
node01           SA_pprcir.so      Idle       Unknown     Unknown     InitFailed
node02           SA_pprcip.so      Idle       Unknown     Unknown     InitFailed
node02           SA_pprcir.so      Idle       Unknown     Unknown     InitFailed
node03           SA_pprcip.so      Idle       Unknown     Unknown     InitFailed
node03           SA_pprcir.so      Idle       Unknown     Unknown     InitFailed
```

Refer to /var/adm/messages and take corrective action according to the error message instructions.

An RCI error is detected after the Shutdown Facility is started.

If Unknown or TestFailed is displayed for Test State of the Agent SA_pprcip.so and SA_pprcir.so on any one of the nodes, an RCI transmission failure occurred between the node and the other nodes. This node is excluded from monitoring and elimination.

For example, an RCI transmission failure occurred between node02, where the sdtool command was executed, and the other nodes in the following:

```
# /opt/SMAW/bin/sdtool -s
Cluster Host      Agent              SA State   Shut State   Test State   Init State
-----
node01           SA_pprcip.so      Idle       Unknown     TestWorked  InitWorked
node01           SA_pprcir.so      Idle       Unknown     TestWorked  InitWorked
node02           SA_pprcip.so      Idle       Unknown     TestFailed  InitWorked
node02           SA_pprcir.so      Idle       Unknown     TestFailed  InitWorked
node03           SA_pprcip.so      Idle       Unknown     TestWorked  InitWorked
node03           SA_pprcir.so      Idle       Unknown     TestWorked  InitWorked
```

Refer to /var/adm/messages and take corrective action according to the error message instructions.

Note

- When RCI transmission failures are detected, the node which uses the failed transmission route is excluded from monitoring and elimination until the Shutdown Facility is restarted.
- If nodes use the same RCI address, the No.7004 error message is output, and the RCI Monitoring Agent daemon is abnormally terminated.
- If you turn off a node for maintenance, the No.7003 error message appears on the other nodes. Take corrective action after the node is started after maintenance.

7.3.2 XSCF

XSCF SA is the Shutdown Agent for SPARC Enterprise M-series.

Note

XSCF is the system monitoring agent provided on SPARC Enterprise M series.

Following functions are added to XSCF to enhance the existing system monitoring agent:

- Remote resetting of the main unit and power-on/off by using http, telnet, and SNMP protocols
- Notifying to the specified email address when an error occurs
- SSL supported
- Monitoring the configuration of RCI device
- Providing XSCF shell
- Supporting the hot swap of main components including power or FAN

Refer to the "XSCF (eXtended System Control Facility) User's Guide" for details on how to configure XSCF.

Setup and configuration

The XSCF must be configured according to the "XSCF (eXtended System Control Facility) User's Guide." Moreover, you must set the user name and password to allow the operation in XSCF.

When using XSCF, check the configuration of XSCF by referring to "5.1.2.2.1 Checking Console Configuration" of the "PRIMECLUSTER Installation and Administration Guide."

Shutdown Agent

The different types of XSCF SAs provide shutdown mechanisms as follows:

- SA_xscfp-panics the node through XSCF.
 - SA_xscfr-resets the node through XSCF.
 - SA_rccu-sends a control break signal to the node through XSCF.
 - SA_rccux-sends a control break signal to the node through XSCF. (*)
- *) XSCF has duplex configuration and does not use the takeover IP address of XSCF.

It is recommended that using XSCF with RCI. In this case, the priority of each agent is as follows:

- (1) RCI Panic (SA_pprcip)
- (2) XSCF Panic (SA_xscfp)
- (3) XSCF Break signal (SA_rccu, SA_rccux)
- (4) RCI Reset (SA_pprcir)
- (5) XSCF Reset (SA_xscfr)

The XSCF log files are as follows:

```
/var/opt/SMAWsf/log/SA_xscfp.log  
/var/opt/SMAWsf/log/SA_xscfr.log  
/var/opt/SMAWsf/log/SA_rccu.log  
/var/opt/SMAWsf/log/SA_rccux.log
```

Note

- The IP address of XSCF belongs to the same segment as the Administrative LAN.

However, if the network routing is configured, the IP address of XSCF does not need to belong to the same segment as the Administrative LAN of the cluster node.

- If XSCF is used for the console, the No.7040 error message might appear on the other nodes in the following cases:
 - Turning off a node for maintenance
 - Changing the network configuration of XSCF
 - Updating the XSCF firmware

If the error message is displayed, take corrective action of the No.7040 error message after each operation is completed.

- After Shutdown Facility (SF) startup, it can take up to 30 seconds for the console Monitoring Agent to detect hardware failures such as RCCU or XSCF errors or a disconnected cable, and setting errors like incorrect IP addresses.

7.3.3 XSCF SNMP

XSCF SNMP shutdown agent is the shutdown agent for SPARC M10 and M12.

For details on XSCF, see "Fujitsu SPARC M12 and Fujitsu M10/SPARC M10 System Operation and Administration Guide."

Setup and configuration

The XSCF must be configured according to "Fujitsu SPARC M12 and Fujitsu M10/SPARC M10 System Operation and Administration Guide."

Moreover, you must set the user name and password to allow the operation in XSCF.

For details, see "5.1.2.1.1 Checking XSCF Information" and "5.1.2.1.2 Setting SNMP" of the "PRIMECLUSTER Installation and Administration Guide."

Shutdown Agent

There are six kinds of XSCF SNMP SAs:

- SA_xscfsnmpg0p-panics the domain by using XSCF-LAN#0
- SA_xscfsnmpg1p-panics the domain by using XSCF-LAN#1
- SA_xscfsnmpg0r-resets the domain by using XSCF-LAN#0
- SA_xscfsnmpg1r-resets the domain by using XSCF-LAN#1
- SA_xscfsnmp0r-resets PPAR by using XSCF-LAN#0 (Used only on the control domain)
- SA_xscfsnmp1r-resets PPAR by using XSCF-LAN#1 (Used only on the control domain)

The priority of each agent is as follows:

1. SA_xscfsnmpg0p
2. SA_xscfsnmpg1p
3. SA_xscfsnmpg0r
4. SA_xscfsnmpg1r
5. SA_xscfsnmp0r
6. SA_xscfsnmp1r

The XSCF log files are as follows:

```

/var/opt/SMAWsf/log/SA_xscfsnmpg0p.log
/var/opt/SMAWsf/log/SA_xscfsnmpg1p.log
/var/opt/SMAWsf/log/SA_xscfsnmpg0r.log
/var/opt/SMAWsf/log/SA_xscfsnmpg1r.log
/var/opt/SMAWsf/log/SA_xscfsnmp0r.log
/var/opt/SMAWsf/log/SA_xscfsnmp1r.log
```



Note

- To make XSCF-LAN redundant, XSCF-LAN#0 and XSCF-LAN#1 should use different subnets.
- After Shutdown Facility (SF) startup, it can take up to 30 seconds to detect hardware failures such as XSCF errors or disconnected cable, and setting errors like incorrect IP addresses.

7.3.4 ALOM

ALOM SA is the Shutdown Agent for SPARC Enterprise T1000, T2000.

Setup and configuration

The ALOM must be configured according to the directions in the "Advanced Lights out Management (ALOM) CMT guide."

Moreover, you must set the user name and password to allow the operation in ALOM.

When using ALOM, check the configuration of ALOM by referring to "5.1.2.4.1 Checking Console Configuration" of the "PRIMECLUSTER Installation and Administration Guide."

Shutdown Agent

SA_sunF-sends a control break signal to the node through ALOM.

The ALOM log file is as follows:

```
/var/opt/SMAWsf/log/SA_sunF.log
```



Note

- The IP address of ALOM belongs to the same segment as the Administrative LAN. However, if the network routing is configured, the IP address of ALOM does not need to belong to the same segment as the Administrative LAN of the cluster node.
- After Shutdown Facility (SF) startup, it can take up to 50 seconds to detect hardware failures such as ALOM errors or a disconnected cable, and setting errors like incorrect IP addresses.

7.3.5 ILOM

ILOM shutdown agent is the shutdown agent for SPARC Enterprise T5120, T5220, T5140, T5240, T5440 series, and SPARC T3, T4, T5, T7, S7 series.

Setup and configuration

The ILOM must be configured according to the directions below:

- For ILOM 2.x
 - Integrated Lights Out Manager User's Guide
- For ILOM 3.0
 - Integrated Lights Out Manager (ILOM) 3.0 Concepts Guide
 - Integrated Lights Out Manager (ILOM) 3.0 Web Interface Procedures Guide
 - Integrated Lights Out Manager (ILOM) 3.0 CLI Procedures Guide
 - Integrated Lights Out Manager (ILOM) 3.0 Getting Started Guide

Moreover, you must set the user name and password to allow the operation in ILOM.

When using ILOM, check the configuration of ILOM by referring to "5.1.2.3.1 Checking Console Configuration" of the "PRIMECLUSTER Installation and Administration Guide."

Shutdown Agent

There are two kinds of ILOM SAs:

- SA_ilomp-panics the node through ILOM.
- SA_ilomr-resets the node through ILOM.

The ILOM log file is as follows:

```
/var/opt/SMAwsf/log/SA_ilomp.log  
/var/opt/SMAwsf/log/SA_ilomr.log
```



- The IP address of ILOM belongs to the same segment as the Administrative LAN.

However, if the network routing is configured, the IP address of ILOM does not need to belong to the same segment as the Administrative LAN of the cluster node.

- If ILOM is used for the console, the No.7040 error message might appear on the other nodes in the following cases:
 - Turning off a node for maintenance
 - Changing the network configuration of ILOM
 - Updating the ILOM firmware

If the error message is displayed, take corrective action of the No.7040 error message after each operation is completed.

- After Shutdown Facility (SF) startup, it can take up to 30 seconds to detect hardware failures such as ILOM errors or a disconnected cable, and setting errors like incorrect IP addresses.

7.3.6 KZONE

KZONE shutdown agent is the shutdown agent for SPARC M10 and SPARC T4, T5, T7, S7 series. It is used in the Oracle Solaris Kernel Zones environment when using PRIMECLUSTER.

Setup and configuration

See "Oracle VM Server for SPARC Guide" and "Creating and Using Oracle Solaris Kernel Zones" for how to configure Oracle Solaris Kernel Zones.

As KZONE shutdown agent, use XSCF when building PRIMECLUSTER on SPARC M10, and use ILOM when building PRIMECLUSTER on SPARC T4, T5, T7, S7 series.

- For XSCF

The XSCF must be configured according to "Fujitsu SPARC M12 and Fujitsu M10/SPARC M10 System Operation and Administration Guide."

Moreover, you must set the user name and password to allow the operation in XSCF.

For details, see "5.1.2.5.1 Checking XSCF Information" of the "PRIMECLUSTER Installation and Administration Guide."

- For ILOM

The ILOM must be configured according to the directions below:

- Integrated Lights Out Manager (ILOM) 3.0 Concepts Guide
- Integrated Lights Out Manager (ILOM) 3.0 Web Interface Procedures Guide
- Integrated Lights Out Manager (ILOM) 3.0 CLI Procedures Guide

- Integrated Lights Out Manager (ILOM) 3.0 Getting Started Guide

Moreover, you must set the user name and password to allow the operation in ILOM.

For details, see "5.1.2.5.2 Checking ILOM Information" of the "PRIMECLUSTER Installation and Administration Guide."

Shutdown Agent

There are three kinds of KZONE shutdown agents:

- SA_kzonep-panics the node (kernel zones)
- SA_kzoner-resets the node (kernel zones)
- SA_kzchkhost-detects an error in the node on which the kernel zones are working

KZONE log file

```
/var/opt/SMAWsf/log/SA_kzonep.log  
/var/opt/SMAWsf/log/SA_kzoner.log  
/var/opt/SMAWsf/log/SA_kzchkhost.log
```



- To make XSCF-LAN redundant, XSCF-LAN#0 and XSCF-LAN#1 should use different subnets.
- The IP address of ILOM should belong to the same segment as the Administrative LAN.

However, if the network routing is configured, the IP address of ILOM does not need to belong to the same segment as the Administrative LAN of the cluster node.

7.3.7 ICMP

ICMP shutdown agent is the shutdown agent of SPARC M10, M12.

It is used when I/O fencing function is used in SPARC M10, M12.

Configuration settings

For the settings of ICMP shutdown agent, refer to "5.1.2.6 Using ICMP Shutdown Agent in SPARC M10 and M12" of "PRIMECLUSTER Installation and Administration Guide."

Shutdown agent

SA_icmp: Checks the startup/suspended status of the node by using the network route.

ICMP log file

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



- ICMP shutdown agent checks the startup/suspended status of the node by using the network route.

For all the designated network routes, if there is no response from the node, consider the node is in a suspended status and end normally. If there are more than one network routes that responded, consider the node is in startup status.

- ICMP shutdown agent does not forcibly stop the other nodes.

7.4 SF split-brain handling

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

7.4.1 Administrative LAN

In PRIMECLUSTER, the administrative LAN is used to handle the split-brain.

For faster and more accurate split-brain handling, make sure to configure the administrative LAN when configuring the Shutdown Facility.

When configuring the administrative LAN in the Shutdown Facility, the public LAN can be used as well. However, due to network load, using the public LAN may require a longer time to handle the split-brain. For this reason, using the administrative LAN is highly recommended.

7.4.2 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 subcluster by calculating the total node weight and application weight of each subcluster. The subcluster with the greatest total weight survives.

Node weights are defined in the SF configuration file `rcsd.cfg`. Typically, you use Cluster Admin's SF Wizard to set the node weights.

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.

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` object's `ShutdownPriority` is transferred to the new node.

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

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.



- 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 "[7.4.4 Configuration notes](#)" for help you with the configuration.
-

7.4.3 Runtime processing

Split-brain handling may be performed by 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.

7.4.4 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.

As can be seen from the default values of both the SF weight and the RMS ShutdownPriority, if no specific action is taken by the administrator to define a split-brain resolution outcome, Largest Sub-cluster Survival is selected by default.

Specific Hardware Survival

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 PCS or 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 PCS or 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.

As an example, in a four-node cluster there are three applications. Set the SF weight of all nodes to 1, and set the ShutdownPriority of the three applications to 50, 10, 10. This would define that the application with a

ShutdownPriority of 50 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 50, 10 and 10 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 51 and A,B,C with 23. The heaviest sub-cluster (D) would win.

7.5 Configuring the Shutdown Facility

For information on configuring the Shutdown Facility, refer to "5.1.2 Configuring the Shutdown Facility" of the "PRIMECLUSTER Installation and Administration Guide."

7.6 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.

7.6.1 Starting and stopping SF

This section describes the procedures for 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 [-bcCsSre] [-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 sdttool manual pages for details.

7.7 Logging

Whenever there is a recurring problem, this will display the debugging information into the `/var/opt/SMAWsf/log/rscd.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 `rscd` 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 CF over IP

This chapter describes CF over IP and how it is configured.

This function is available for Solaris 10. It is not supported on Solaris 11 or later version.

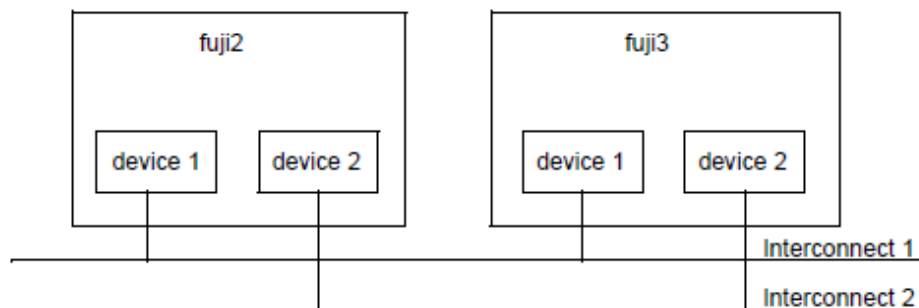
8.1 Overview

Note

- All IP configuration must be done prior to using CF over IP. The devices must be initialized with a unique IP address and a broadcast mask. IP must be configured to use these devices. If the configuration is not done, cfconfig(1M) will fail to load CF, and CF will not start.
- The devices used for CF over IP must not be controlled by an RMS userApplication that could unconfigure a device due to Offline processing.

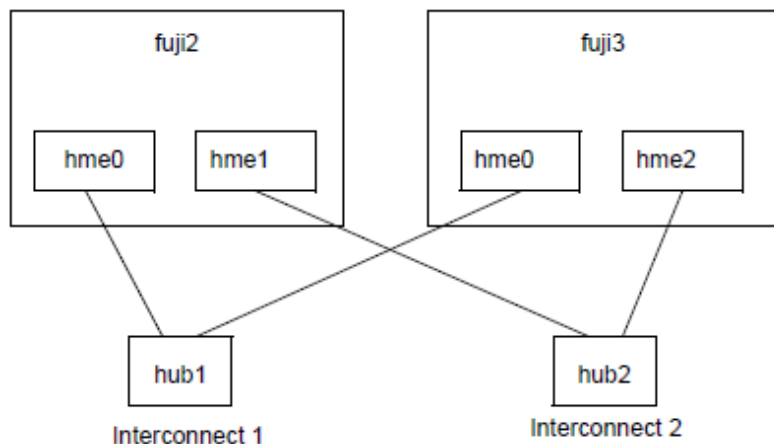
CF communications are based on the use of interconnects. An interconnect is a communications medium which can carry CF's link-level traffic between the CF nodes. A properly configured interconnect will have connections to all of the nodes in the cluster through some type of device. The figure below illustrates this configuration.

Figure 8.1 Conceptual view of CF interconnects



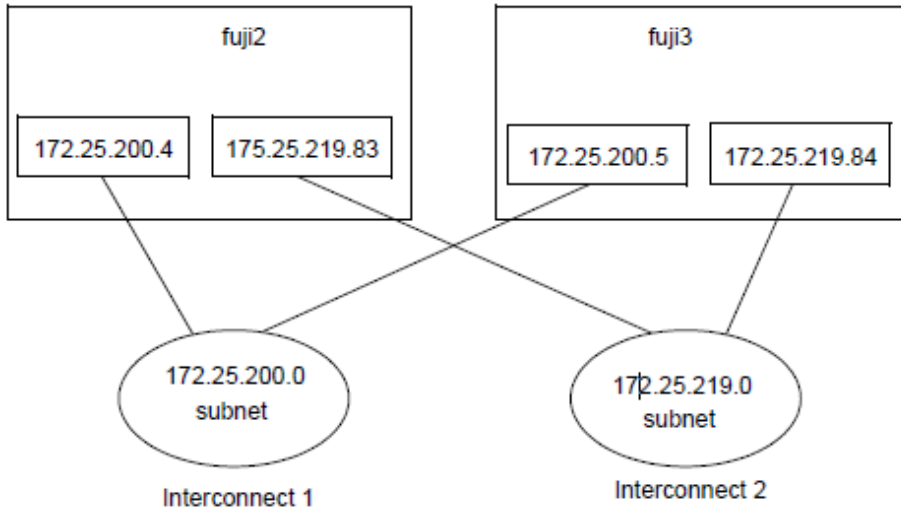
When CF is used over Ethernet, Ethernet devices are used as the interfaces to the interconnects. The interconnects themselves are typically Ethernet hubs or switches. An example of this is shown in the following figure.

Figure 8.2 CF with Ethernet interconnects



When using CF over IP, the IP interface is used as a device to connect to the interconnect and ensures redundancy using multiple IP subnetworks. The figure below illustrates the configuration of CF over IP.

Figure 8.3 CF with IP interconnects



It is also possible to use mixed configurations in which CF is run over both Ethernet devices and IP subnetworks.

When using CF over IP, you should make sure that each node in the cluster has an IP interface on each subnetwork used as an interconnect. You should also make sure that all the interfaces for a particular subnetwork use the same IP broadcast address and the same netmask on all cluster nodes. This is particularly important since CF depends on an IP broadcast on each subnet to do its initial cluster join processing.

Note

- IPv4 address is used for CF over IP.
- CF is not allowed to reach nodes that are on different subnets.

Note

When selecting a subnetwork to use for CF, you should use a private subnetwork that only cluster nodes can access. CF security is based on access to its interconnects. Any node that can access an interconnect can join the cluster and acquire root privileges on any cluster node. When CF over IP is used, this means that any node on the subnetworks used by CF must be trusted. You should not use the public interface to a cluster node for CF over IP traffic unless you trust every node on your public network.

8.2 Configuring CF over IP

To configure CF over IP, you should do the following:

- Designate which subnetworks you want to use for CF over IP. Up to four subnetworks can be used.
- Make sure that each node that is to be in the cluster has IP interfaces properly configured for each subnetwork. Make sure the IP broadcast and netmasks are correct and consistent on all nodes for the subnetworks.
- Make sure that all of these IP interfaces are up and running.
- Run the CF Wizard in Cluster Admin.

The CF Wizard has a window which allows CF over IP to be configured. The Wizard will probe all the nodes that will be in the cluster, find out what IP interfaces are available on each, and then offer them as choices in the CF over IP window. It will also try to group the choices for each node by subnetworks. See "[1.1 CF, CIP, and CIM configuration](#)" for details.

CF uses special IP devices to keep track of CF over IP configuration. There are four of these devices named as follows:

```
/dev/ip0  
/dev/ip1
```

```
/dev/ip2  
/dev/ip3
```

These devices do not actually correspond to any device files under /dev in the Solaris. Instead, they are just place holders for CF over IP configuration information within the CF product. Any of these devices can have an IP address and broadcast address assigned by the cfconfig(1M) command (or by Cluster Admin which invokes the cfconfig(1M) command in the Wizard).

If you run cfconfig(1M) by hand, you may specify any of these devices to indicate you want to run CF over IP. The IP device should be followed by an IP address and broadcast address of an interface on the local node. The addresses must be in internet dotted-decimal notation. For example, to configure CF on fuji2 in "Figure 8.3 CF with IP interconnects," the cfconfig(1M) command would be as follows:

```
fuji2 # cfconfig -s A clustername /dev/ip0  
\172.25.200.4 172.25.200.255 /dev/ip1 172.25.219.83
```

It really does not matter which IP device you use. The above command could equally have used /dev/ip2 and /dev/ip3.

 **Note**

.....
The cfconfig(1M) command does not do any checks to make sure that the IP addresses are valid.
.....

The IP devices chosen in the configuration will appear in other commands such as cftool -d and cftool -r. IP interfaces will not show up in CF pings using cftool -p unless they are configured for use with CF and the CF driver is loaded.

 **Note**

.....
cftool -d shows a relative speed number for each device, which is used to establish priority for the message send. If the configured device is IP, the relative speed 100 is used. This is the desired priority for the logical IP device. If a Gigabit Ethernet hardware device is also configured, it will have priority.
.....

Chapter 9 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 as described in the PRIMECLUSTER Software Release Guide and Installation Guide.

9.1 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/adm/messages` file. You can use the `cftool(1M)` command for checking states, configuration information. To use the CF log viewer click on the Tools pulldown menu and select View Syslog messages. 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/adm/messages`. You might have to look in multiple files (`/var/adm/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/adm/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. "4.5 Error Messages" in "PRIMECLUSTER Messages" contains messages that can be found in the `/var/adm/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/adm/messages` file:

```
Nov 9 08:51:45 fuji2 unix: LOG3.0973788705 1080024 1008 4 0 1.0 cf:ens CF: Icf Error:
(service_err_type route_src route_dst). (0 0 0 0 0 0 0 2 0 0 0 5 0 0 0 5)
```

The first 80 bytes are the log3 prefix as in the following:

```
Nov 9 08:51:45 fuji2 unix: LOG3.0973788705 1080024 1008 4 0 1.0 cf:ens
```

This part 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 in the following:

```
CF: Icf Error: (service_err_type route_src route_dst). (0 0 0 0 0 0 0 2 0 0 0 5 0 0 0 5)
```

This message is from the `cf:ens` service (that is, the Cluster Foundation, Event Notification Service) and the error is CF: Icf Error. This error is described in "5.1.4 Error Messages" in "PRIMECLUSTER Messages" as signifying a missing heartbeat and/or a route down. This gives us direction to look into the cluster interconnect further. A larger piece of the `/var/adm/messages` file shows as follows:

```
fuji2# tail /var/adm/messages
Nov 9 08:51:45 fuji2 unix: SUNW,pci-gem1: Link Down - cable problem?
Nov 9 08:51:45 fuji2 unix: SUNW,pci-gem0: Link Down - cable problem?
Nov 9 08:51:45 fuji2 unix: LOG3.0973788705 1080024 1008 4 0 1.0 cf:ens CF:
Icf Error: (service_err_type route_src route_dst). (0 0 0 0 0 0 0 2 0 0 0 5 0 0 0 5)
```

```

Nov  9 08:51:46 fuji2 unix: SUNW,pci-gem0: Link Down - cable problem?
Nov  9 08:51:48 fuji2 last message repeated 1 time
Nov  9 08:51:48 fuji2 unix: LOG3.0973788708 1080024 1008 4 0 1.0 cf:ens CF:
Icf Error: (service err_type route_src route_dst). (0 0 0 0 0 0 0 2 0 0 0 4 0 0 0 4)
Nov  9 08:51:50 fuji2 unix: SUNW,pci-gem0: Link Down - cable problem?
Nov  9 08:51:52 fuji2 last message repeated 1 time
Nov  9 08:51:53 fuji2 unix: LOG3.0973788713 1080024 1008 4 0 1.0 cf:ens CF:
Icf Error: (service err_type route_src route_dst). (0 0 0 0 0 0 0 2 0 0 0 4 0 0 0 4)
Nov  9 08:51:53 fuji2 unix: LOG3.0973788713 1080024 1015 5 0 1.0 cf:ens CF:
Node fuji2 Left Cluster POKE. (0 0 2)
Nov  9 08:51:53 fuji2 unix: Current Nodee Status = 0

```

Here we see that there are error messages from the Ethernet controller indicating that the link is down, possibly because of a cable problem. This is the clue we need to solve this problem; the Ethernet used for the interconnect has failed for some reason. The investigation in this case should shift to the cables and hubs to insure that they are all powered up and securely connected.

Several options for the command cftool are listed above as sources for information. Some examples are as follows:

```

fuji2# cftool -l
Node    Number State      Os      Cpu
fuji2   2      UP          Solaris Sparc

```

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:

```

fuji2# cftool -l
Node    Number State      Os
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/adm/messages file. In this case, we see the following:

```

fuji2# tail /var/adm/messages
May 30 17:36:39 fuji2 unix: pseudo-device: fcp0
May 30 17:36:39 fuji2 unix: fcp0 is /pseudo/fcp@0
May 30 17:36:53 fuji2 unix: LOG3.0991269413 1080024 1007 5 0 1.0 cf:eventlog CF:
(TRACE): JoinServer: Startup.
May 30 17:36:53 fuji2 unix: LOG3.0991269413 1080024 1009 5 0 1.0 cf:eventlog CF:
Giving UP Mastering (Cluster already Running).
May 30 17:36:53 fuji2 unix: LOG3.0991269413 1080024 1006 4 0 1.0 cf:eventlog CF:
fuji4: 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 LEFTCLUSTER state](#)" for the description and the instructions for resolving the state.

The next option to cftool shows the device states as follows:

```

fuji2# cftool -d
Number Device    Type Speed    Mtu    State Configured Address
1      /dev/hme0 4    100     1432   UP     YES     00.80.17.28.21.a6
2      /dev/hme3 4    100     1432   UP     YES     08.00.20.ae.33.ef
3      /dev/hme4 4    100     1432   UP     YES     08.00.20.b7.75.8f
4      /dev/ge0  4    1000    1432   UP     YES     08.00.20.b2.1b.a2
5      /dev/ge1  4    1000    1432   UP     YES     08.00.20.b2.1b.b5

```

Here we can see the 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 as shown in the following:

```

fuji2# cftool -n
Node    Number State      Os      Cpu

```



```
fuji2 1 UP Solaris Sparc
fuji3 2 UP Solaris Sparc
```

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 shown in the following example:

```
fuji2# cftool -r
Node    Number Srcdev Dstdev Type State Destaddr
fuji2   1      4      4      4    UP   08.00.20.b2.1b.cc
fuji2   1      5      5      4    UP   08.00.20.b2.1b.94
fuji3   2      4      4      4    UP   08.00.20.b2.1b.a2
fuji3   2      5      5      4    UP   08.00.20.b2.1b.b5
```

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 as shown in the following:

```
fuji2# cftool -r
Node    Number Srcdev Dstdev Type State Destaddr
fuji2   2      3      2      4    UP   08.00.20.bd.5e.a1
fuji3   1      3      3      4    UP   08.00.20.bd.60.e4
```

```
fuji2# cftool -n
Node    Number State      Os      Cpu
fuji2   2      UP          Solaris Sparc
fuji3   1      LEFTCLUSTER Solaris Sparc
```

9.2 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 your customer service representative. Samples from the error log (`/var/adm/messages`) have the log3 header stripped from them in this section.

9.2.1 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 PRIMECLUSTER installation manual 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.

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 follows:

```
fuji2# cftool -l
Node      Number  State   Os
fuji2    --      COMINGUP --
```

This indicates the driver is loaded and 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/adm/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 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:

```
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.

Information

Note that the log3 prefix is stripped from all of the error message text displayed below. Messages in the error log will appear as follows:

```
Mar 10 09:47:55 fuji2 unix: LOG3.0952710475 1080024 1014 4 0 1.0 cf:ens
CF: Local node is missing a route from node: fuji3
```

However they are shown here as follows:

```
CF: Local node is missing a route from node: fuji3
```

Join problems

Problem:

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 fuji4 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 to send an echo request to all the nodes of the cluster:

```
fuji2# cftool -e
Localdev Srcdev Address Cluster Node Number Joinstate
3 2 08.00.20.bd.5e.a1 FUJI fuji2 2 6
3 3 08.00.20.bd.60.ff 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 then look under the Interconnect Problems heading for the problem "The node only sees itself on the configured

interconnects." 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:

```
fuji2# cfconfig -u
fuji2# cfconfig -l
```

Note

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

If this attempt to join the cluster succeeds, then look under the Problem: "The node intermittently fails to join the cluster." If the node did not join the cluster then proceed with the problem below "The node does not join the cluster and some or all nodes respond to cftool -e."

Problem:

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

Diagnosis:

At this point, we know that the CF device is loading properly and that this node can communicate to 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 as in the following example:

```
fuji2# cftool -e
Localdev  Srcdev  Address                Cluster  Node    Number  Joinstate
3         2       08.00.20.ae.33.ef     FUJI    fuji1   3       6
3         2       08.00.20.bd.5e.a1    FUJI    fuji2   2       6
3         3       08.00.20.bd.60.ff    FUJI    fuji3   1       6
```

```
fuji2# cftool -e
Localdev  Srcdev  Address                Cluster  Node    Number  Joinstate
3         2       08.00.20.ae.33.ef     FUJI    fuji1   3       6
3         2       08.00.20.bd.5e.a1    FUJI    fuji2   2       6
3         3       08.00.20.bd.60.ff    FUJI    fuji3   1       6
3         3       08.00.20.bd.60.e4    FUJI    fuji4   1       6
```

```
fuji2# cftool -e
Localdev  Srcdev  Address                Cluster  Node    Number  Joinstate
3         2       08.00.20.ae.33.ef     FUJI    fuji1   3       6
3         2       08.00.20.bd.5e.a1    FUJI    fuji2   2       6
3         3       08.00.20.bd.60.ff    FUJI    fuji3   1       6
```

```
fuji2# cftool -e
Localdev  Srcdev  Address                Cluster  Node    Number  Joinstate
3         2       08.00.20.ae.33.ef     FUJI    fuji1   3       6
3         2       08.00.20.bd.5e.a1    FUJI    fuji2   2       6
3         3       08.00.20.bd.60.ff    FUJI    fuji3   1       6
3         3       08.00.20.bd.60.e4    FUJI    fuji4   1       6
```

```
fuji2# cftool -e
Localdev  Srcdev  Address                Cluster  Node    Number  Joinstate
3         2       08.00.20.ae.33.ef     FUJI    fuji1   3       6
3         2       08.00.20.bd.5e.a1    FUJI    fuji2   2       6
3         3       08.00.20.bd.60.ff    FUJI    fuji3   1       6
3         3       08.00.20.bd.60.e4    FUJI    fuji4   1       6
```

```
fuji2# cftool -e
Localdev  Srcdev  Address                Cluster  Node    Number  Joinstate
3         2       08.00.20.ae.33.ef     FUJI    fuji1   3       6
3         2       08.00.20.bd.5e.a1    FUJI    fuji2   2       6
```

3	3	08.00.20.bd.60.ff	FUJI	fuji3	1	6
3	3	08.00.20.bd.60.e4	FUJI	fuji4	1	6

Notice that the node fuji4 does not show up in each of the echo requests. This indicates that the connection to the node fuji4 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 utilities on that node show any errors. If we log on to fuji4 and look at the network devices, we see the following:

Number	Device	Type	Speed	Mtu	State	Configured	Address
1	/dev/hme0	4	100	1432	UP	NO	00.80.17.28.2c.fb
2	/dev/hme1	4	100	1432	UP	NO	00.80.17.28.2d.b8
3	/dev/hme2	4	100	1432	UP	YES	08.00.20.bd.60.e4

The netstat(1M) utility in Solaris reports information about the network interfaces. The first attempt will show the following:

```
fuji4# netstat -i
```

Name	Mtu	Net/Dest	Address	Ipkts	Ierrs	Opkts	Oerrs	Collis	Queue
lo0	8232	loopback	localhost	65	0	65	0	0	0
hme0	1500	fuji4	fuji4	764055	8	9175	0	0	0
hme1	1500	fuji4-priv	fuji4-priv	2279991	0	2156309	0	7318	0

Notice that the hme2 interface is not shown in this report. This is because Solaris does not report on interconnects that are not configured for TCP/IP. To temporarily make Solaris report on the hme2 interface, enter the ifconfig plumb command as follows:

```
fuji4# ifconfig hme2 plumb
```

Repeat the command as follows:

```
fuji4# netstat -i
```

Name	Mtu	Net/Dest	Address	Ipkts	Ierrs	Opkts	Oerrs	Collis	Queue
lo0	8232	loopback	localhost	65	0	65	0	0	0
hme0	1500	fuji4	fuji4	765105	8	9380	0	0	0
hme1	1500	fuji4-priv	fuji4-priv	2282613	0	2158931	0	7319	0
hme2	1500	default	0.0.0.0	752	100	417	0	0	0

Here we can see that the hme2 interface has 100 input errors (Ierrs) from 752 input packet (Ipkts). This means that one in seven packets had an error; this rate is too high for PRIMECLUSTER to use successfully. This also explains why fuji4 sometimes responded to the echo request from fuji2 and sometimes did not.

Point

.....

It is always safe to plumb the interconnect. This will not interfere with the operation of PRIMECLUSTER.

.....

To resolve these errors further, we can look at the undocumented -k option to the Solaris netstat command as follows:

```
fuji4# netstat -k hme2
hme2:
ipackets 245295 ierrors 2183 opackets 250486 oerrors 0 collisions 0
defer 0 framing 830 crc 1353 sqe 0 code_violations 38 len_errors 0
ifspeed 100 buff 0 oflo 0 uflo 0 missed 0 tx_late_collisions 0
retry_error 0 first_collisions 0 nocarrier 0 inits 15 nocanput 0
allocbfail 0 runt 0 jabber 0 babble 0 tmd_error 0 tx_late_error 0
rx_late_error 0 slv_parity_error 0 tx_parity_error 0 rx_parity_error 0
slv_error_ack 0 tx_error_ack 0 rx_error_ack 0 tx_tag_error 0
rx_tag_error 0 eop_error 0 no_tmids 0 no_tbufs 0 no_rbufs 0
rx_late_collisions 0 rbytes 22563388 obytes 22729418 multircv 0 multixmt 0
brdcstrcv 472 brdcstxmt 36 norcvbuf 0 noxmtbuf 0 phy_failures 0
```

Most of this information is only useful to specialists for problem resolution. The two statistics that are of interest here are the framing and crc errors. These two error types add up to exactly the number reported in ierrors. Further resolution of this problem consists of trying each of the following steps:

- Ensure the Ethernet cable is securely inserted at each end.

- Try repeated cftool -e and look at the netstat -i. If the results of the cftool 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 your support personnel will have to further diagnose the problem.

Problem:

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

```
Mar 10 09:47:55 fuji2 unix: LOG3.0952710475 1080024 1014 4 0 1.0 cf:ens CF: Local node
is missing a route from node: fuji3
Mar 10 09:47:55 fuji2 unix: LOG3.0952710475 1080024 1014 4 0 1.0 cf:ens CF: missing
route on local device: /dev/hme3
Mar 10 09:47:55 fuji2 unix: LOG3.0952710475 1080024 1014 4 0 1.0 cf:ens CF: Node fuji3
Joined Cluster FUJI. (0 1 0)
```

Diagnosis:

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

Same message as on console.

No console messages on node fuji3.

Look in /var/adm/messages on node fuji3:

```
fuji2# cftool -d
Number Device      Type Speed    Mtu      State Configured Address
1      /dev/hme0 4      100     1432    UP      NO      08.00.06.0d.9f.c5
2      /dev/hme1 4      100     1432    UP      YES     00.a0.c9.f0.15.c3
3      /dev/hme2 4      100     1432    UP      YES     00.a0.c9.f0.14.fe
4      /dev/hme3 4      100     1432    UP      NO      00.a0.c9.f0.14.fd
```

```
fuji3# cftool -d
Number Device      Type Speed    Mtu      State Configured Address
1      /dev/hme0 4      100     1432    UP      NO      08.00.06.0d.9f.c5
2      /dev/hme1 4      100     1432    UP      YES     00.a0.c9.f0.15.c3
3      /dev/hme2 4      100     1432    UP      YES     00.a0.c9.f0.14.fe
4      /dev/hme3 4      100     1432    UP      YES     00.a0.c9.f0.14.fd
```

```
/dev/hme3 is not configured on node fuji2
Mar 10 11:00:28 fuji2 unix:WARNING:hme3:no MII link detected
Mar 10 11:00:31 fuji2 unix:LOG3.0952714831 1080024 1008 4 0 1.0cf:ens
CF:Icf Error:(service err_type route_src route_dst).(0 0 0 0 2 0 0 0 3 0 0 0 3 0 0 0)
Mar 10 11:00:53 fuji2 unix:NOTICE:hme3:100 Mbps full-duplex link up
Mar 10 11:01:11 fuji2 unix:LOG3.0952714871 1080024 1007 5 0 1.0cf:ens
CF (TRACE):Icf:Route UP:node src dest.(0 2 0 0 0 3 0 0 0 3 0 0 0)
The hme3 device or interconnect temporarily failed.
```

```
fuji2# cftool -n
Node  Number State      Os      Cpu
fuji2 1      LEFTCLUSTER Solaris Sparc
fuji3 2      UP          Solaris Sparc
```

Problem:

/dev/hme3 is not configured on node fuji2.

```
Mar 10 11:00:28 fuji2 unix: WARNING: hme3: no MII link detected
Mar 10 11:00:53 fuji2 unix: NOTICE: hme3: 100 Mbps full-duplex link up
```

Diagnosis:

Look in /var/adm/messages on node fuji2:

```
Mar 10 11:00:28 fuji2 unix: WARNING: hme3: no MII link detected
Mar 10 11:00:31 fuji2 unix: LOG3.0952714831 1080024 1008 4 0 1.0 cf:ens CF: Icf
Error: (service err_type route_src route_dst). (0 0 0 0 0 2 0 0 0 3 0 0 0 0)
Mar 10 11:00:53 fuji2 unix: NOTICE: hme3: 100 Mbps full-duplex link up
Mar 10 11:01:11 fuji2 unix: LOG3.0952714871 1080024 1007 5 0 1.0 cf:ens CF (TRACE):
Icf: Route UP: node src dest. (0 2 0 0 0 3 0 0 0 3 0 0 0)
```

Problem:

The hme3 device or interconnect temporarily failed. It could be the NIC on either of the cluster nodes or a cable or hub problem.

Node in LEFTCLUSTER state

IF SF is not configured, and node fuji2 panicked and has rebooted. The following console message appears on node fuji2:

```
Mar 10 11:23:41 fuji2 unix: LOG3.0952716221 1080024 1012 4 0 1.0
cf:ens CF: fuji2: busy: local node not down: retrying.
```

Diagnosis:

Look in /var/adm/messages on node fuji2:

```
Mar 10 11:23:41 fuji2 unix: LOG3.0952716221 1080024 1007 5 0 1.0 cf:ens CF (TRACE):
JoinServer: Startup.
Mar 10 11:23:41 fuji2 unix: LOG3.0952716221 1080024 1009 5 0 1.0 cf:ens CF: Giving
UP Mastering (Cluster already Running).
Mar 10 11:23:41 fuji2 unix: LOG3.0952716221 1080024 1012 4 0 1.0 cf:ens CF: Join
postponed, server fuji3 is busy.
```

...last message repeats.

No new messages on console or in /var/adm/messages on fuji2:

```
fuji2: cftool -n
Node Number State Os Cpu
fuji2 1 LEFTCLUSTER Solaris Sparc
fuji3 2 UP Solaris Sparc
```

Identified problem:

Node fuji2 has left the cluster and has not been declared DOWN.

Fix:

To fix this problem, enter the following command:

```
# 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 fuji2:

```
Mar 10 11:34:21 fuji2 unix: LOG3.0952716861 1080024 1005 5 0 1.0
cf:ens          CF: MYCLUSTER: fuji2 is Down. (0 1 0)
Mar 10 11:34:29 fuji2 unix: LOG3.0952716869 1080024 1004 5 0 1.0
cf:ens          CF: Node fuji2 Joined Cluster MYCLUSTER. (0 1 0)
```

The following console message appears on node fuji2:

```
Mar 10 11:32:37 fuji2 unix: LOG3.0952716757 1080024 1004 5 0 1.0
cf:ens          CF: Node fuji2 Joined Cluster MYCLUSTER. (0 1 0)
```

9.3 Collecting troubleshooting information

If the failure occurs in the PRIMECLUSTER system, collect the necessary information from all of the cluster nodes. For details about procedures on collecting troubleshooting information, see "Appendix C Troubleshooting" of "PRIMECLUSTER Installation and Administration Guide." Then, contact field engineers.

Glossary

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

In PRIMECLUSTER configurations, an administrative LAN is a private local area network (LAN) on which machines such as the system console and cluster console reside. Because normal users do not have access to the administrative LAN, it provides an extra level of security. The use of an administrative LAN is required.

See also public LAN.

API

See Application Program Interface.

application (RMS)

A resource categorized as a userApplication used to group resources into a logical collection.

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 RMS Application Wizards to create object definitions for a specific type of application.

Application Wizards

See RMS Application Wizards.

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 power control

This function is provided by the Enhanced Support Facility (ESF), and it automatically switches the server power on and off.

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 PRIMECLUSTER module resides on top of the basic OS and provides internal interfaces for the CF (Cluster Foundation) functions that the PRIMECLUSTER services use in the layer above.

See also Cluster Foundation.

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.

CCBR

See Cluster Configuration Backup and Restore.

CF node name

The CF cluster node name, which is configured when a CF cluster is created.

Cluster Configuration Backup and Restore

CCBR provides a simple method to save the current PRIMECLUSTER configuration information of a cluster node. It also provides a method to restore the configuration information.

Cluster Interconnect Protocol

CIP is an interface such as hme0eth0 except the physical layer is built on top of the cluster interconnect.

CF

See Cluster Foundation.

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.

Cluster Foundation

The set of PRIMECLUSTER modules that provides basic clustering communication services.

See also base cluster foundation (CF).

cluster interconnect (CF)

The set of private network connections used exclusively for PRIMECLUSTER communications.

Cluster Join Services (CF)

This PRIMECLUSTER module handles the forming of a new cluster and the addition of nodes.

concatenated virtual disk

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. (Applies to transitioning users of existing Fujitsu Technology Solutions products only.)

See also mirror virtual disk, simple virtual disk, striped virtual disk, virtual disk.

[configuration file \(RMS\)](#)

The RMS configuration file that defines the monitored resources and establishes the interdependencies between them. The default name of this file is config.us.

[custom detector \(RMS\)](#)

See [detector \(RMS\)](#).

[custom type \(RMS\)](#)

See [generic type \(RMS\)](#).

[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.

[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 \(CF\)](#).

[ENS \(CF\)](#)

See [Event Notification Services \(CF\)](#).

[environment variables \(RMS\)](#)

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.

[Event Notification Services \(CF\)](#)

This PRIMECLUSTER 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 \(RMS\)](#).

[Fast switching mode](#)

One of the LAN duplexing modes presented by Global Link Service.

This mode uses a multiplexed LAN simultaneously to provide enhanced communication scalability between servers and high-speed switchover if a LAN failure occurs.

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).

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 PRIMECLUSTER optional module provides network high availability solutions by multiplying a network route.

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).

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.

high availability

This concept applies to the use of redundant resources to avoid single points of failure.

interconnect (CF)

See cluster interconnect (CF).

Internet Protocol address

A numeric address that can be assigned to computers or applications.

See also IP aliasing.

Internode Communications facility

This module is the network transport layer for all PRIMECLUSTER 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.

JOIN (CF)

See Cluster Join Services (CF).

keyword

A word that has special meaning in a programming language. For example, in the 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 public LAN.

local node

The node from which a command or process is initiated.

See also remote node, node.

log file

The file that contains a record of significant system events or messages. The base monitor, wizards, and detectors can have their own log files.

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.

Meta Data Server

GFS daemon that centrally manages the control information of a file system (meta-data).

mirror virtual disk

Mirror virtual disks consist of two or more physical devices, and all output operations are performed simultaneously on all of the devices. (Applies to transitioning users of existing Fujitsu Technology Solutions products only.)

See also concatenated virtual disk, simple virtual disk, striped virtual disk, virtual disk.

mixed model cluster

A cluster system that is built from different SPARC Enterprise models. For example, one node is SPARC Enterprise M3000, and another node is SPARC Enterprise M4000. The models are divided into the following groups: SPARC M12-2/M12-2S, SPARC M10-1/M10-4/M10-4S, SPARC S7-2/S7-2L, SPARC T7-1/T7-2/T7-4, SPARCT5-2/T5-4/T5-8, SPARC T4-1/T4-2/T4-4, SPARC T3-1/T3-2/T3-4, SPARC Enterprise T1000/T2000, SPARC Enterprise T5120/T5220/T5140/T5240/T5440, and SPARC Enterprise M3000/M4000/M5000/M8000/M9000.

mount point

The point in the directory tree where a file system is attached.

multihosting

Multiple controllers simultaneously accessing a set of disk drives. (Applies to transitioning users of existing Fujitsu Technology Solutions products only.)

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.

NIC switching mode

One of the LAN duplexing modes presented by Global Link Service. The duplexed NIC is used exclusively, and LAN monitoring between the server and the switching HUB, and switchover if an error is detected are implemented.

node

A host which is a member of a cluster. A computer node is the same as a computer.

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)

In the configuration file or a system graph, this is a representation of a physical or virtual resource.

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. The keyword associated with an object definition is object.

See also attribute (RMS), object type (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 generic type (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 PRIMECLUSTER modules depend upon.

OPS

See Oracle Parallel Server.

Oracle Parallel Server

Oracle Parallel Server allows access to all data in a database to users and applications in a clustered or MPP (massively parallel processing) platform.

OSD (CF)

See operating system dependent (CF).

parent (RMS)

An object in the configuration file or system graph that has at least one child.

See also child (RMS), configuration file (RMS), system graph (RMS).

primary node (RMS)

The default node on which a user application comes online when RMS is started. This is always the nodename of the first child listed in the userApplication object definition.

private network addresses

Private network addresses are a reserved range of IP addresses specified by the Internet Assigned Numbers Authority. They may be used internally by any organization but, because different organizations can use the same addresses, they should never be made visible to the public internet.

private resource (RMS)

A resource accessible only by a single node and not accessible to other RMS nodes.

See also resource (RMS), shared resource.

queue

See message queue.

PRIMECLUSTER services (CF)

Service modules that provide services and internal interfaces for clustered applications.

redundancy

This is the capability of one object to assume the resource load of any other object in a cluster, and the capability of RAID hardware and/or RAID software to replicate data stored on secondary storage devices.

public LAN

The local area network (LAN) by which normal users access a machine.

See also Administrative LAN.

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 telecommunications line or LAN.

See also local node.

remote node

See remote 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 Reliant Monitor Services (RMS).

RMS Application Wizards

RMS Application Wizards add new menu items to the RMS Wizard Tools for a specific application.

See also RMS Wizard Tools, Reliant Monitor Services (RMS).

RMS commands

Commands that enable RMS resources to be administered from the command line.

RMS configuration

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 Tools

A software package composed of various configuration and administration tools used to create and manage applications in an RMS configuration.

See also RMS Application Wizards, Reliant Monitor Services (RMS).

SAN

See Storage Area Network.

Scalable Internet Services (SIS)

Scalable Internet Services is a TCP connection load balancer, and dynamically balances network access loads across cluster nodes while maintaining normal client/server sessions for each connection.

scalability

The ability of a computing system to dynamically handle any increase in work load. Scalability is especially important for Internet-based applications where growth caused by Internet usage presents a scalable challenge.

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

simple virtual disk

Simple virtual disks define either an area within a physical disk partition or an entire partition. (Applies to transitioning users of existing Fujitsu Technology Solutions products only.)

See also concatenated virtual disk, striped virtual disk, virtual disk.

SIS

See Scalable Internet Services (SIS).

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

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). (Applies to transitioning users of existing Fujitsu Technology Solutions products only.)

See also concatenated virtual disk, mirror virtual disk, simple virtual disk, virtual disk.

switching mode

LAN duplexing mode presented by Global Link Service.

There are mode types: fast switching mode, NIC switching mode, GS/SURE linkage mode (Solaris), GS linkage mode (Linux), virtual NIC mode, and multipath mode (Solaris).

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), symmetrical switchover (RMS), failover (RMS, SIS), switchover (RMS).

synchronized power control

When the power of one node is turned in the cluster system, this function turns on all other powered-off nodes and disk array unit that are connected to nodes through RCI cables.

system graph (RMS)

A visual representation (a map) of monitored resources used to develop or interpret the 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

With virtual disks, a pseudo device driver is inserted between the highest level of the Solaris logical Input/Output (I/O) system and the physical device driver. This pseudo device driver then maps all logical I/O requests on physical disks. (Applies to transitioning users of existing Fujitsu Technology Solutions products only.)

See also concatenated virtual disk, mirror virtual disk, simple virtual disk, striped virtual disk.

Web-Based Admin View

This is a common base to utilize the Graphic User Interface of PRIMECLUSTER. This interface is in Java.

wizard (RMS)

An interactive software tool that creates a specific type of application using pretested object definitions. An enabler is a type of wizard.

Index

[Special characters]	
/etc/cip.cf.....	26
/var/opt/SMAWsf/log/rscd.log.....	91
[A]	
Adding and removing a node from CIM.....	67
Adding a new node to CF.....	24
Add to CIM.....	68
ALOM.....	85
application weight.....	89
Automatic resource registration.....	40
[B]	
Basic layout.....	75
Beginning the process.....	95
Broken interconnects.....	71
[C]	
Caused by a cluster partition.....	73
Caused by reboot.....	74
Caused by a panic/hung node.....	73
Caused by staying in the kernel debugger too long.....	73
CCBR.....	26
CF, CIP, and CIM configuration.....	1
CF/CIP Wizard.....	2
CF name.....	2
CF node information.....	53
CF node name.....	1
CF over IP.....	92
CF Registry (CFREG).....	31
CF Registry and Integrity Monitor.....	31
CF route table.....	52
CF route tracking.....	50
CF security.....	5
cfset.....	4
CF topology.....	14
CF topology table.....	54,75
CIM options.....	67
CIM Override.....	69
CIP.....	2
CIP configuration file.....	26
CIP interface.....	18,19
CIP traffic.....	2
CIP wizard (IPv4) window.....	17
CIP wizard (IPv6) window.....	18
clgettree.....	37
Cluster Foundation.....	1
Cluster Foundation (CF).....	1
Cluster Integrity Monitor (CIM).....	31
Cluster name.....	1
Cluster resource management.....	34
Collecting troubleshooting information.....	103
configure CIP.....	18
Configuring CF over IP.....	93
Configuring CIM.....	31
Configuring the Shutdown Facility.....	90
Connections table.....	15
[D]	
default value for StartingWaitTime.....	42
Description of the LEFTCLUSTER state.....	71
Diagnostics and troubleshooting.....	95
Differences between CIP and CF over IP.....	3
Displaying statistics.....	63
Displaying the topology table.....	53
[E]	
Example of creating a cluster.....	5
Examples.....	77
Exclusive device list for EMC Symmetrix.....	38
[F]	
Full interconnect.....	15
[G]	
GUI administration.....	47
[H]	
Heartbeat monitor.....	65
[I]	
ICF statistics.....	64
ICMP.....	87
ILOM.....	85
init command.....	71
Interconnect.....	1
IP name.....	18
IP over CF.....	3
[J]	
Join-related problems.....	97
[K]	
Kernel parameters for Resource Database.....	34
KZONE.....	86
[L]	
LEFTCLUSTER state.....	71
[M]	
MAC statistics.....	64
Main CF table.....	49
Marking nodes DOWN.....	59
Monitoring Agent (MA).....	79
[N]	
Node details.....	52
Node to Node statistics.....	65
[O]	
Overview.....	79
[P]	
Panicked nodes.....	71
Partial interconnect.....	15

