

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 (CF)

Configuration and Administration

Guide 4.3

Oracle Solaris

J2S2-1588-04ENZ0(01)
September 2017

Copyright and Trademarks

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

Dell EMC, EMC, PowerPath, Symmetrix, SAN Manager and SRDF are trademarks or registered trademarks of EMC Corporation.

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

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.

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.

All Rights Reserved, Copyright (C) FUJITSU LIMITED 2012-2017.

Contents

1	Preface	1
1.1	Contents of this manual	1
1.2	Related documentation	2
1.3	Conventions	3
1.3.1	Notation	4
1.3.1.1	Prompts	4
1.3.1.2	The keyboard	4
1.3.1.3	Typefaces	4
1.3.1.4	Example 1	4
1.3.1.5	Example 2	5
1.3.2	Command syntax	5
1.4	Important notes and cautions	5
1.5	Abbreviations	6
1.6	Revision history	6
2	Cluster Foundation	7
2.1	CF, CIP, and CIM configuration	7
2.1.1	Differences between CIP and CF over IP	11
2.1.2	cfset	13
2.1.3	CF security	15
2.1.4	Example of creating a cluster	16
2.1.5	Adding a new node to CF	38
2.2	CIP configuration file	39
2.3	Cluster Configuration Backup and Restore (CCBR)	41
3	CF Registry and Integrity Monitor	47
3.1	CF Registry	47
3.2	Cluster Integrity Monitor	48
3.2.1	Configuring CIM	48
3.2.2	Query of the quorum state	49
3.2.3	Reconfiguring quorum	50
4	Cluster resource management	53
4.1	Overview	53
4.2	Kernel parameters for Resource Database	53
4.3	Resource Database configuration	57
4.4	Registering hardware information	59
4.4.1	Setup exclusive device list	59
4.4.2	Exclusive device list for Dell EMC Symmetrix	60
4.4.2.1	emcpower Devices and native Devices	61
4.4.2.2	BCV, R2, GateKeeper, CKD	61

Contents

4.4.2.3	VCMDB	62
4.4.2.4	Simplified setup for exclusive device list - clmkdiskinfo, clmkdiskinfo 62	
4.4.3	Automatic resource registration	63
4.5	Start up synchronization	65
4.5.1	Start up synchronization and the new node	67
4.6	Adding a new node	67
4.6.1	Backing up the Resource Database	69
4.6.2	Reconfiguring the Resource Database	70
4.6.3	Configuring the Resource Database on the new node	71
4.6.4	Adjusting StartingWaitTime	72
4.6.5	Restoring the Resource Database	72
5	GUI administration	75
5.1	Overview	76
5.2	Starting Cluster Admin GUI and logging in	76
5.3	Main CF table	79
5.4	CF route tracking	81
5.5	Node details	84
5.6	Displaying the topology table	85
5.7	Starting and stopping CF	87
5.7.1	Starting CF	89
5.7.2	Stopping CF	92
5.8	Marking nodes DOWN	93
5.9	Using PRIMECLUSTER log viewer	94
5.9.1	Search based on time filter	96
5.9.2	Search based on keyword	97
5.9.3	Search based on severity levels	98
5.10	Displaying statistics	99
5.11	Heartbeat monitor	104
5.12	Adding and removing a node from CIM	105
5.13	Unconfigure CF	108
5.14	CIM Override	109
6	LEFTCLUSTER state	111
6.1	Description of the LEFTCLUSTER state	112
6.2	Recovering from LEFTCLUSTER	114
6.2.1	Caused by a panic/hung node	114
6.2.2	Caused by staying in the kernel debugger too long	114
6.2.3	Caused by a cluster partition	115
6.2.4	Caused by reboot	117
7	CF topology table	119

7.1	Basic layout	121
7.2	Selecting devices	122
7.3	Examples	123
8	Shutdown Facility	127
8.1	Overview	127
8.2	Configuring SF	128
8.2.1	Setting procedure before configuring SF	128
8.2.2	Configuration file of SF	129
8.3	Available SAs	130
8.3.1	RCI	131
8.3.2	XSCF	133
8.3.3	XSCF SNMP	135
8.3.4	ALOM	137
8.3.5	ILOM	137
8.3.6	KZONE	139
8.3.7	RPDU	140
8.3.8	NPS	141
8.4	SF split-brain handling	142
8.4.1	Administrative LAN	142
8.4.2	SF split-brain handling	143
8.4.2.1	RMS ShutdownPriority attribute	143
8.4.2.2	Shutdown Facility weight assignment	144
8.4.2.3	Disabling split-brain handling	144
8.4.3	Runtime processing	144
8.4.4	Configuration notes	145
8.5	Configuring the Shutdown Facility	147
8.6	SF administration	147
8.6.1	Starting and stopping SF	148
8.6.1.1	Starting and stopping SF manually	148
8.6.1.2	Starting and stopping SF automatically	148
8.7	Logging	148
9	CF over IP	149
9.1	Overview	149
9.2	Configuring CF over IP	151
10	Diagnostics and troubleshooting	153
10.1	Beginning the process	153
10.2	Symptoms and solutions	157
10.2.1	Join-related problems	158
10.3	Collecting troubleshooting information	167
10.3.1	Executing the fjsnap command	168
10.3.2	System dump	169

Contents

10.3.3	XSCF log	170
11	Manual pages	171
11.1	CCBR	171
11.2	CF	171
11.3	CFS	172
11.4	CIP	172
11.5	CPAT	173
11.6	Monitoring Agent	173
11.7	PAS	173
11.8	PCS	174
11.9	Resource Database	174
11.10	RMS	175
11.11	RMS Wizards	176
11.12	SF	176
11.13	Web-Based Admin View	177
12	Release information	179
	Glossary	183
	Abbreviations	201
	Figures	205
	Tables	209
	Index	211

1 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. These services include the following:

- Internode communications
- Node state management
- Cluster-wide configuration information
- Management and administration
- Distributed lock management

This document assumes that the reader is familiar with the contents of the PRIMECLUSTER *Concepts Guide* and that the PRIMECLUSTER software has been installed as described in the PRIMECLUSTER *Software Release Guide and Installation Guide*.

1.1 Contents of this manual

This manual contains the configuration and administrative information for the PRIMECLUSTER components. This manual is organized as follows:

- The Chapter “Cluster Foundation” describes the administration and configuration of the Cluster Foundation (CF).
- The Chapter “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).
- The Chapter “Cluster resource management” discusses the database which is a synchronized clusterwide database holding information to multiple PRIMECLUSTER products.
- The Chapter “GUI administration” describes the administrative features in the CF portion of the Cluster Admin graphical user interface (GUI).
- The Chapter “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.

- The Chapter “CF topology table” discusses the CF topology table as it relates to the CF portion of the Cluster Admin GUI.
- The Chapter “Shutdown Facility” describes the components and advantages of PRIMECLUSTER SF and provides administrative information.
- The Chapter “CF over IP” discusses CF communications based on the use of interconnects.
- The Chapter “Diagnostics and troubleshooting” provides help for troubleshooting and problem resolution for PRIMECLUSTER Cluster Foundation.
- The Chapter “Manual pages” lists the manual pages for PRIMECLUSTER.
- The Chapter “Release information” describes primary changes in this manual.

1.2 Related documentation

The documentation listed in this section contains information relevant to PRIMECLUSTER.

In addition to this manual, the following manuals are also available for PRIMECLUSTER:

- Release notices for all products—These documentation files are included as HTML files on the PRIMECLUSTER Framework DVD. Release notices provide late-breaking information about installation, configuration, and operations for PRIMECLUSTER. Read this information first.
- *PRIMECLUSTER Concepts Guide*—Provides concepts on the PRIMECLUSTER family of products.
- *PRIMECLUSTER Installation and Administration Guide (Oracle Solaris)*—Provides instructions for installing and upgrading PRIMECLUSTER products.
- *PRIMECLUSTER Software Release Guide and Installation Guide*—Provides instructions for installing and upgrading PRIMECLUSTER products.
- *PRIMECLUSTER Reliant Monitor Services (RMS) with Wizard Tools Configuration and Administration Guide*—Provides instructions for configuring and administering PRIMECLUSTER RMS using the interface of PRIMECLUSTER Wizard Tools. The diagnostic procedures to solve RMS configuration problems, and how to view and interpret RMS log files are also provided.

- *PRIMECLUSTER Global Disk Services Configuration and Administration Guide*— Provides information on configuring and administering Global Disk Services (hereinafter GDS).
- *PRIMECLUSTER Global File Services Configuration and Administration Guide*— Provides information on configuring and administering Global File Services (hereinafter GFS).
- *PRIMECLUSTER Global Link Services Configuration and Administration Guide: Redundant Line Control Function for Virtual NIC Mode*— Provides information on configuring and administering the Virtual NIC Mode for Global Link Services (hereinafter GLS).
- *PRIMECLUSTER Global Link Services Configuration and Administration Guide: Redundant Line Control Function*—Provides information on configuring and administering the redundant line control function for GLS.
- *PRIMECLUSTER Global Link Services Configuration and Administration Guide: Multipath Function*—Provides information on configuring and administering the multipath function for GLS.
- *PRIMECLUSTER Web-Based Admin View Operation Guide*—Provides information on using the Web-Based Admin View management GUI.
- *PRIMECLUSTER DR/PCI Hot Plug User's Guide*—Provides the operation of Dynamic Reconfiguration function and PCI Hot Plug function by using PRIMECLUSTER.
- *PRIMECLUSTER Messages*— Provides how to set the environment for PRIMECLUSTER. Messages output during system operation are also provided.
- *RMS Wizards documentation package*—Available on the PRIMECLUSTER DVD. These documents deal with topics such as the configuration of file systems and IP addresses. They also describe the different kinds of wizards.

1.3 Conventions

In order to standardize the presentation of material, this manual uses a number of notational, typographical, and syntactical conventions.

1.3.1 Notation

This manual uses the following notational conventions.

1.3.1.1 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*.

1.3.1.2 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.

1.3.1.3 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>	Variables that you must replace with an actual value. Items or buttons in a GUI window.
Bold	Items in a command line that you must type exactly as shown.

Typeface conventions are shown in the following examples.

1.3.1.4 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):/:
```

1.3.1.5 Example 2

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

```
$ cat file
```

1.3.2 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.

1.4 Important notes and cautions



Important

Indicates important information.



Caution

Indicates a situation that can cause harm to data.

**Note**

Indicates information that needs special attention.

1.5 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 "Solaris 10 operating system (Solaris 10)" or the "Solaris 11 operating system (Solaris 11)".

1.6 Revision history

Revision	Location	Manual code
SPARC T5/T7/S7 series were added to the supported hardware.	8.3	J2S2-1588-04ENZ0(01)

2 Cluster Foundation

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

This chapter discusses the following:

- The Section “CF, CIP, and CIM configuration” describes CF, Cluster Interconnect Protocol (CIP) and Cluster Integrity Monitor (CIM) configuration that must be done prior to other cluster services.
- The Section “CIP configuration file” describes the format of the CIP configuration file.
- The Section “Cluster Configuration Backup and Restore (CCBR)” details a method to save and restore PRIMECLUSTER configuration information.

2.1 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—This can be any name that you choose as long as it is 31 characters or less per name and each character comes from the set of printable ASCII characters, excluding white space, newline, and tab characters. Cluster names are always mapped to upper case.

- 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 names are the same as the Web-Based Admin View names; however, you can use the CF Wizard to change them.

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`. `cfset(1M)` can be used to change the 10 second default. See the Section “`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 much as it does an Ethernet device. However, all the IP addresses for all the cluster nodes on that interconnect must have the same IP subnetwork, and their IP broadcast addresses must be the same (refer to the Chapter “CF over IP” for more information).

The IP interfaces used by CF must be completely configured by the System Administrator before they are used by CF. You can run CF over both Ethernet devices and IP devices.

Higher level services, such as RMS, SF, 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.



Interconnects should be redundant to avoid a single point of failure in the cluster.

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

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

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

To ensure that you can communicate between nodes using CIP, the IP address on each node for a specific CIP interface should use the same subnet.

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

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

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

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

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

*cfname*RMS

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



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. Refer to the *PRIME-CLUSTER Software Release Guide and Installation Guide* for more details.



Caution

Do not add CIP nodenames manually to the `/etc/hosts` or `/etc/inet/ipnodes` file because the CIP Wizard automatically updates the `/etc/hosts` and `/etc/inet/ipnodes` files on each node in the cluster.

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 "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.

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. You can start the GUI by entering the following URL with a browser running a proper version of the Java plug-in:

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

management_server is the primary or secondary management server you configured for this cluster. See "3.1.3 Web environment" in "PRIMECLUSTER Web-Based Admin View Operation Guide" for details on configuring the primary and secondary management service and on which browsers and Java plug-ins are required for the Cluster Admin GUI.

2.1.1 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 1).

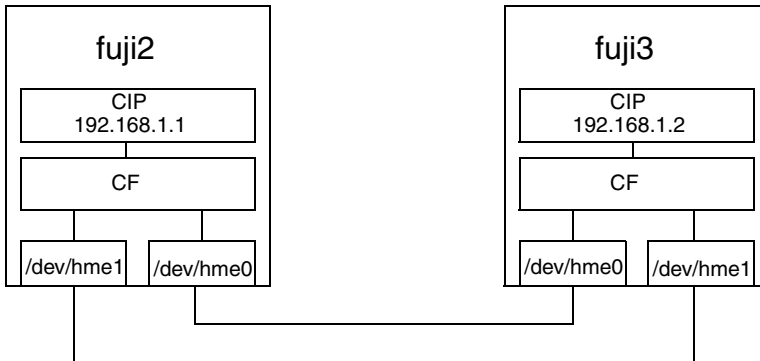


Figure 1: CIP diagram

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.

- i** 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 2).

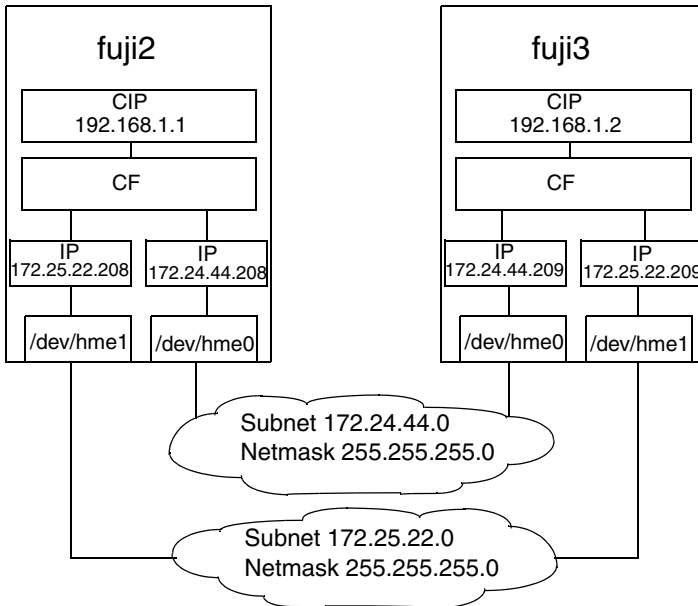


Figure 2: CF over IP diagram



Note

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

2.1.2 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 `cfcp` and `cfsh` during initial cluster configuration) or by using a text editor.

- The file consists of the following tuple entries, *Name* and *Value*:

Name:

- This is the name of a CF configuration parameter. It must be the first token in a line.
- Maximum length for *Name* is 31 bytes. The name must be unique.
- 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”)
- `CFCP` (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 while 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.

2.1.3 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.

2.1.4 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 `fujii2` and `fujii3`, and the cluster name is `FUJII`.

This example assumes that Web-Based Admin View configuration has already been done. `fujii2` is assumed to be configured as the primary management server for Web-Based Admin View, and `fujii3` is the secondary management server.

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

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

If the host name of the management server is `fujii2`, enter the following:

```
http://fujii2:8081/Plugin.cgi
```

After a few moments, a login pop-up appears asking for a user name and password (see Figure 3).

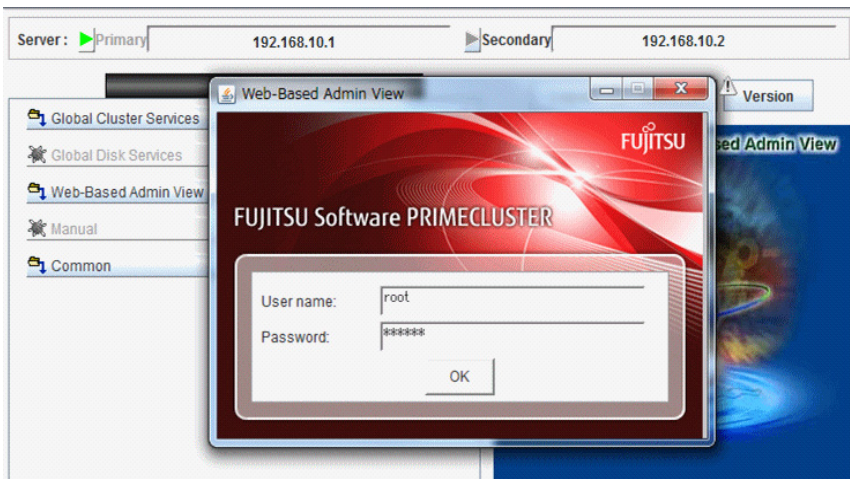


Figure 3: Login pop-up

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 `fujii2`. 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 `fujii2` 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 `fujii2` 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 appears (see Figure 4). Click on the button labeled *Global Cluster Services*.

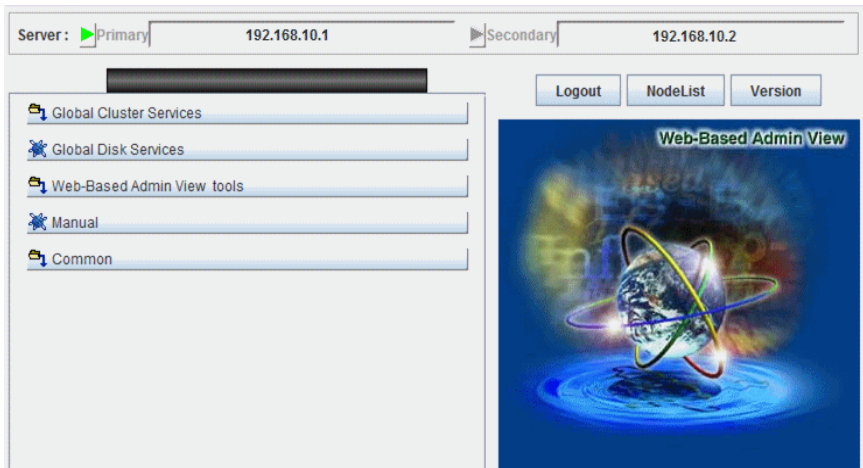


Figure 4: Main Web-Based Admin View window after login

The Cluster Admin selection window appears (see Figure 5).

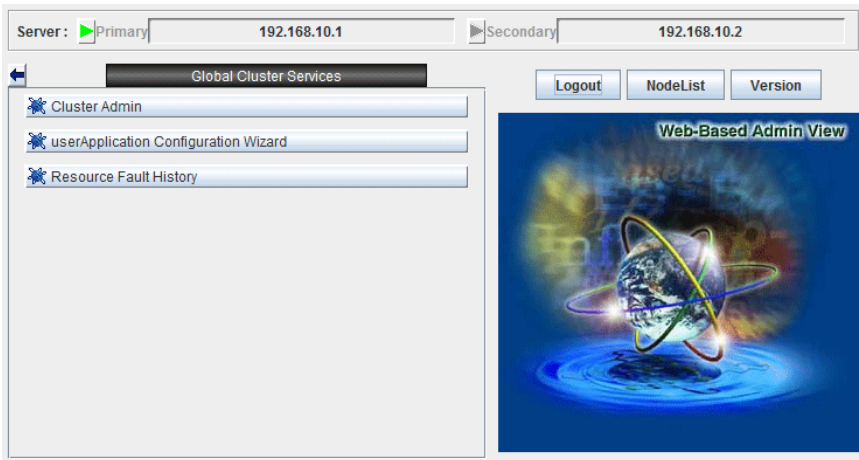


Figure 5: Global Cluster Services window in Web-Based Admin View

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

The *Choose a node for initial connection* window appears (see Figure 6).

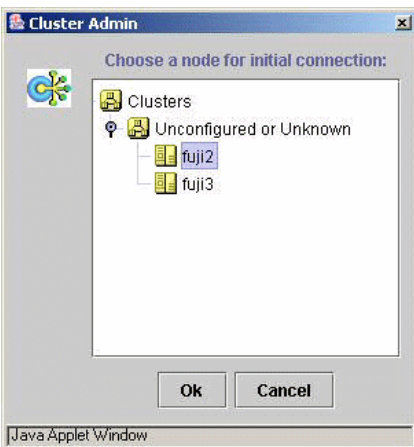


Figure 6: Initial connection pop-up

The *Choose a node for initial connection* window (see Figure 6) 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 `fujii2` nor `fujii3` have had CF configured, so either would be acceptable as a choice. In Figure 6, `fujii2` is selected. Clicking on the *OK* button causes the main Cluster Admin GUI to appear. Since CF is not configured on `fujii2`, a window similar to Figure 7 appears.

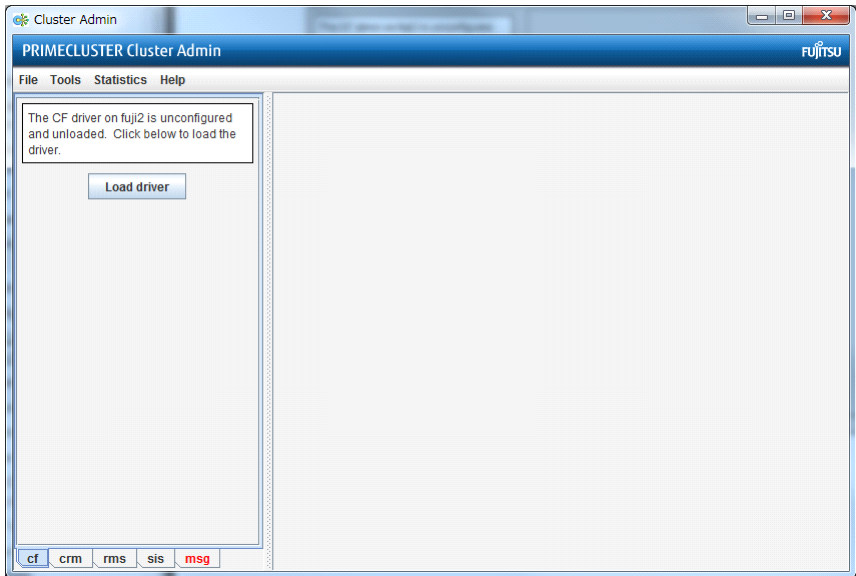


Figure 7: CF is unconfigured and unloaded

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

A window indicating that CF is loaded but not configured appears (see Figure 8).

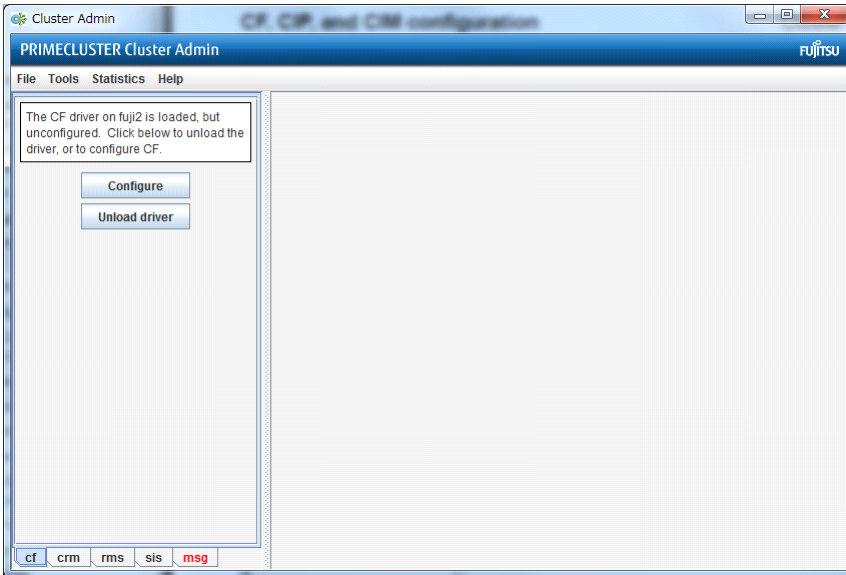


Figure 8: CF loaded but not configured

Click on the *Configure* button to bring up the CF Wizard.

The CF Wizard begins by looking for existing clusters (see Figure 9).

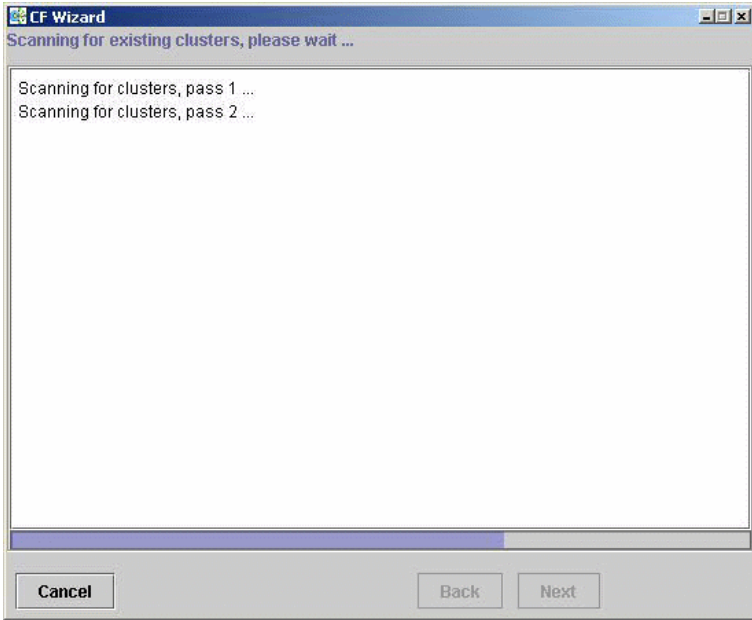


Figure 9: Scanning for clusters

After the CF Wizard finishes looking for clusters, a window similar to Figure 10 appears.

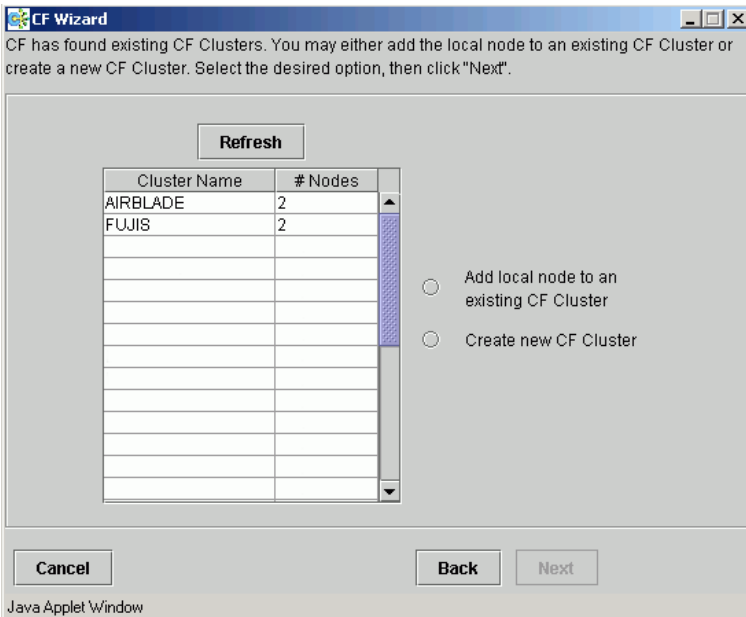


Figure 10: Creating or joining a cluster

This window lets you decide if you want to join an existing cluster or create a new one. To create a new cluster, ensure that the *Create new CF Cluster* button is selected. Then, click on the *Next* button.

The window for creating a new cluster appears (see Figure 11).

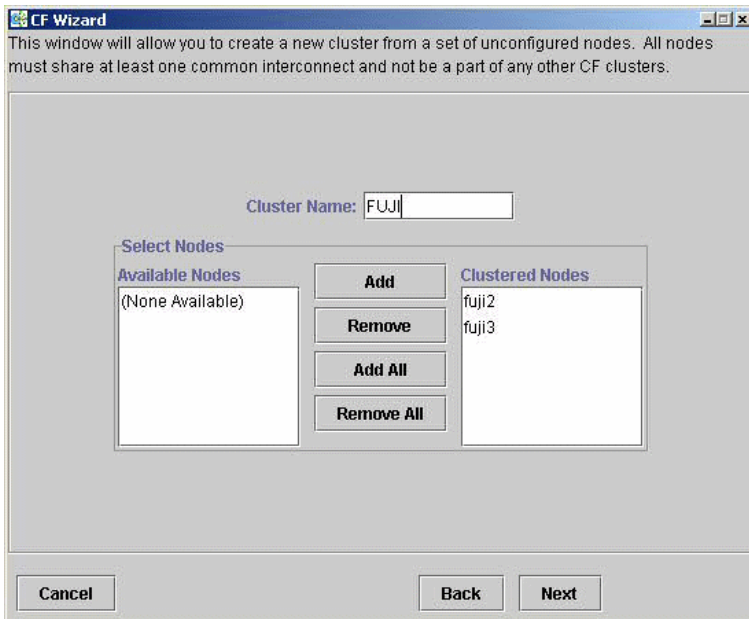


Figure 11: Selecting cluster nodes and the cluster name

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

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

If you get to this window and you do not see all of the nodes that you want to be part of this cluster, then there is a very good chance that you have not configured Web-Based Admin View properly. When Web-Based Admin View is initially installed on the nodes in a potential cluster, it configures each node as if it were a primary management server independent of every other node. If no additional Web-Based Admin View configuration were done, and you started up

Cluster Admin on such a node, then Figure 11 would show only a single node in the right-hand box and no additional nodes on the left-hand side. If you see this, then it is a clear indication that proper Web-Based Admin View configuration has not been done.

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 Figure 12 appears.

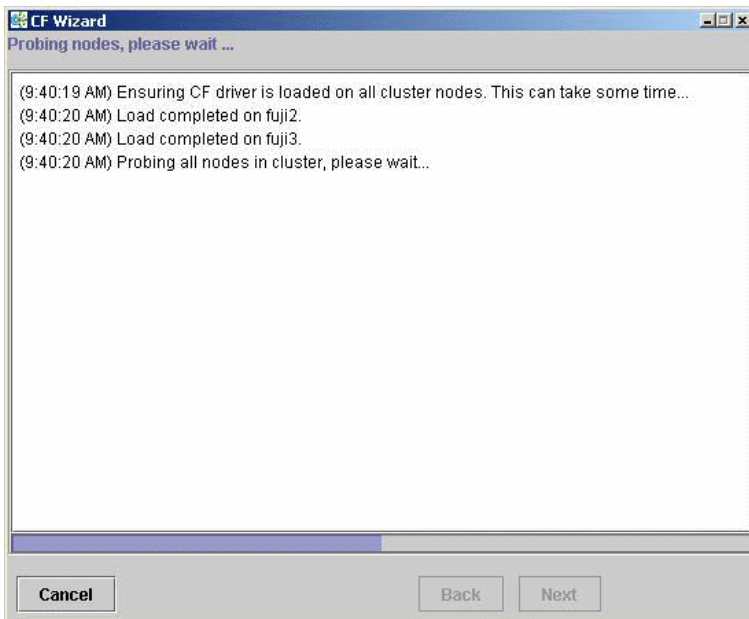


Figure 12: CF loads and pings

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.

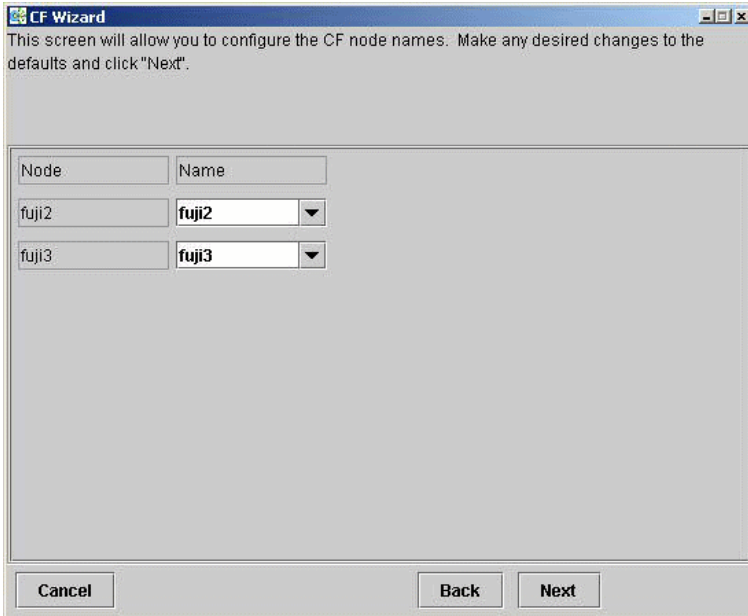


Figure 13: Edit CF node names

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 (see Figure 14).

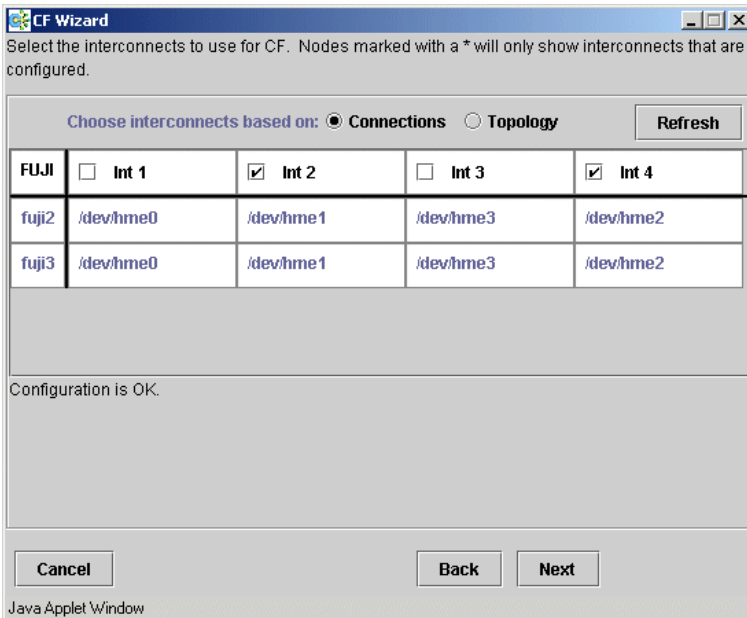


Figure 14: CF topology and connection table

Before using the CF topology and connection table in Figure 14, 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 14.

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 14, for example, Interconnect 1 (`Int 1`) has `/dev/hme0` on `fuji2` and `fuji3` attached to it. The cluster name is `FUJI`.

i 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 14, 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 will list problems or warnings concerning the configuration.

If the CF interconnect and the device are configured successfully, click *Next*. The CF over IP window is displayed (see Figure 15).

Since CF over IP is not supported in a Solaris 11 environment, the CIP Wizard window as shown in the Figure 15 is not displayed. The CIP Wizard window as shown in the Figure 16 is displayed.

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

The CIP Wizard window as shown in the Figure 16 is displayed.

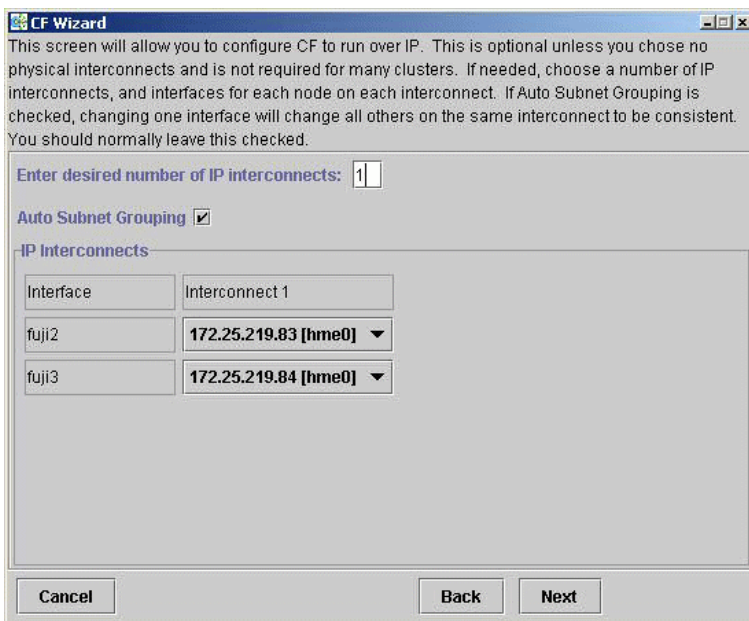



Figure 15: CF over IP window

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.



Note

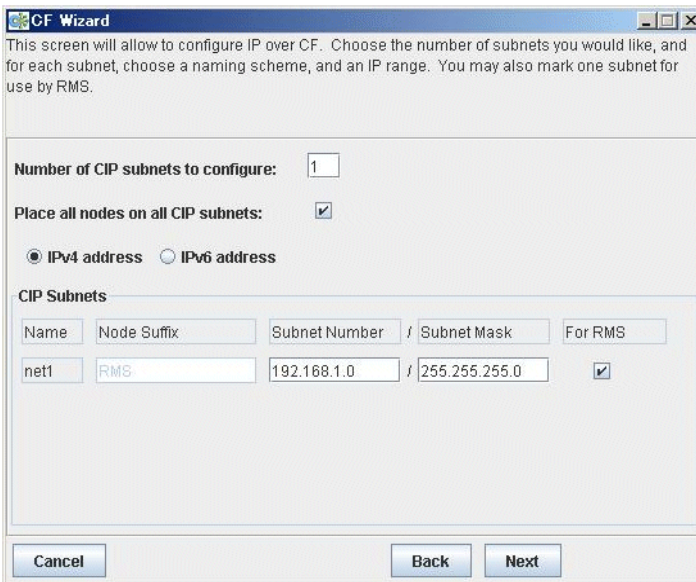
Since CF over IP is not supported in a Solaris 11 environment, the CIP Wizard window as shown in the Figure 15 is not displayed.

 Only interfaces that are configured at system boot can be used for CF over IP.

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 on this window, and click on *Next*. The CIP Wizard window appears (see Figure 16 or Figure 17).



CF Wizard

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

Number of CIP subnets to configure:

Place all nodes on all CIP subnets:

IPv4 address IPv6 address

CIP Subnets

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

Cancel Back Next

Figure 16: CIP Wizard window

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

Number of CIP subnets to configure:

Place all nodes on all CIP subnets:

IPv4 address IPv6 address

CIP Subnets

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

Cancel Back Next

Figure 17: CIP Wizard window (IPv6)

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 the window shown in Figure 16: CIP wizard (IPv4) window and Figure 17: CIP wizard (IPv6) window.

When using IPv4 for CIP interface

The following values are assigned for CIP interface:

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

The CIP Wizard will automatically fill in a default value for the subnet number 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 16, 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 *fujj2* and *fujj3*. This will result in the following configuration:

- On *fujj2*, a CIP interface will be configured with the following:
IP nodename: *fujj2RMS*
IP address: 192.168.1.1
Subnet Mask: 255.255.255.0
- On *fujj3*, a CIP interface will be configured with the following:
IP nodename: *fujj3RMS*
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.

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 17, 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, CIM configuration window appears (see Figure 18).

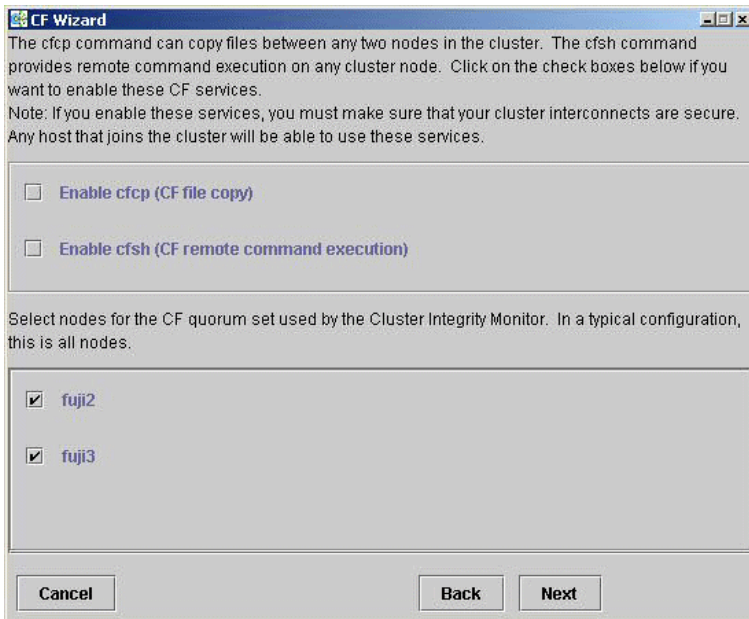


Figure 18: CIM configuration window

The CIM configuration window in Figure 18 has the following parts:

- The upper portion allows you to enable `cfcp` and `cfsh`.
`cfcp` is a CF-based file copy program. It allows files to be copied among the cluster hosts. `cfsh` is a remote command execution program that 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.

**Caution**

Do not change the default selection of the nodes that are members of the CIM set unless you fully understand the ramifications of this change.

Do not add CIP nodenames manually to the `/etc/hosts` or `/etc/inet/ipnodes` file because the CIP Wizard automatically updates the `/etc/hosts` and `/etc/inet/ipnodes` files on each node in the cluster.

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

**Caution**

- Enabling either of these means that you must trust all 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.
- Make sure to enable `cfcp` and `cfsh` when using RMS.

Click on the *Next* button to go to the summary window (see Figure 19).

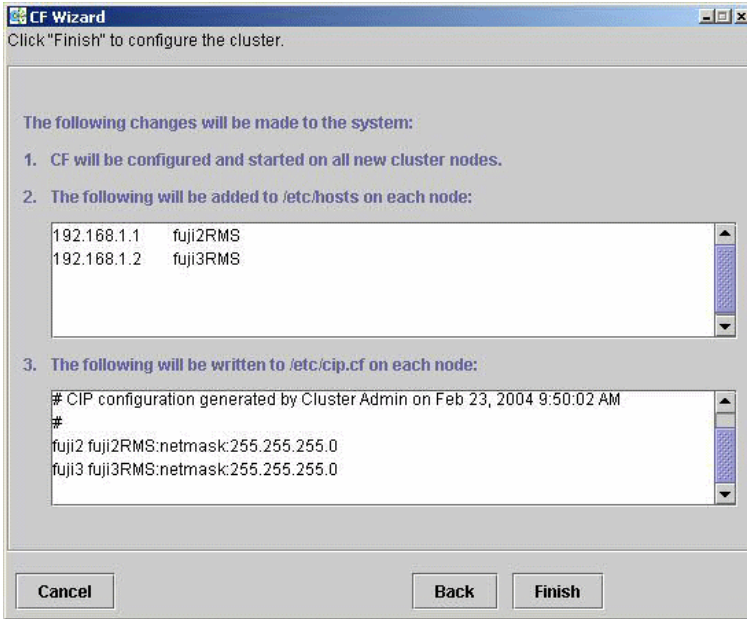


Figure 19: Summary window

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

A window similar to Figure 20 is displayed while the configuration is being done.

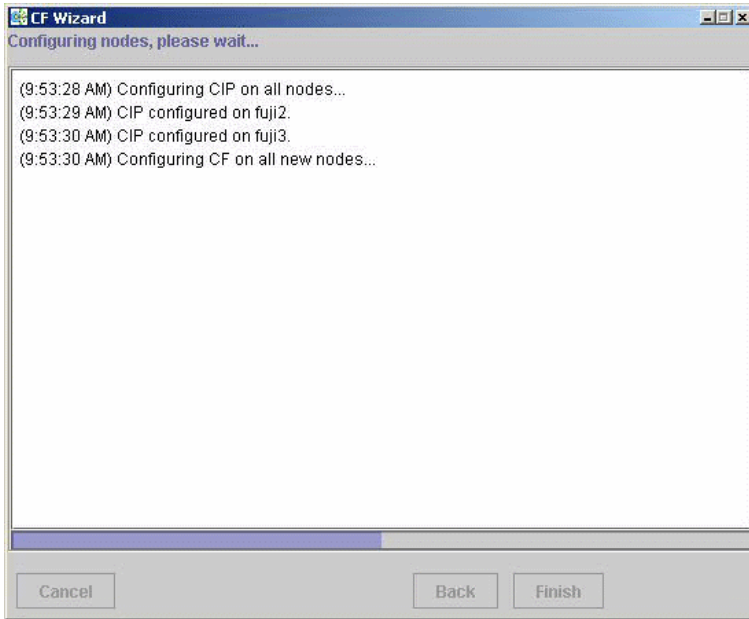


Figure 20: Configuration processing window

This window is updated after each configuration step. When configuration is complete, a pop-up appears announcing this fact (see Figure 21).

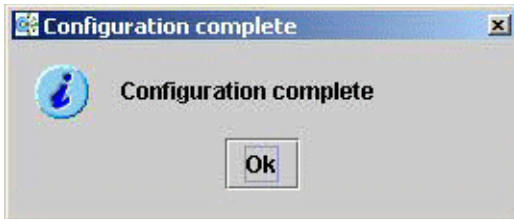


Figure 21: Configuration completion pop-up

Click on the *OK* button, and the pop-up is dismissed. The configuration processing window now has a *Finish* button (see Figure 22).

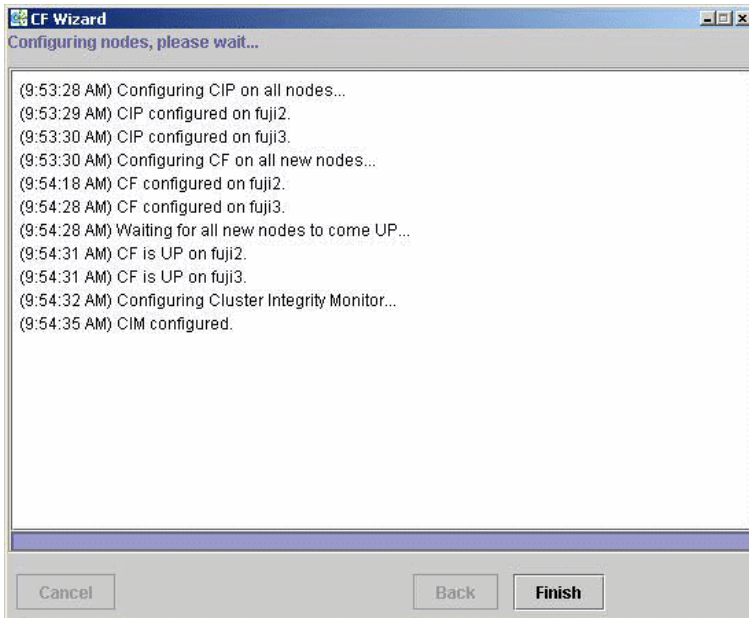


Figure 22: Configuration window after completion

You might see the following error message in the window shown in Figure 22.

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

You can safely ignore this message.

When the CF Wizard is run on an unconfigured node, it will ask the CF driver to push its modules on every Ethernet device on the system. This allows CF to do CF pings on each interface so that the CF Wizard can discover the network topology.

Occasionally, this unload will fail. To correct this problem, you need to unload and reload the CF driver on the node in question. This can be done easily through the GUI (refer to the Section “Starting and stopping CF”).

Click on the *Finish* button to dismiss the window in Figure 22. A small pop-up appears asking if you would like to run the SF Wizard. Click on *yes*, and run the SF Wizard (described in the section “Configuring the Shutdown Facility”).

After the CF (and optionally the SF) Wizards are done, you see the main CF window. After several moments, the window will be updated with new configuration and status information (see Figure 23).

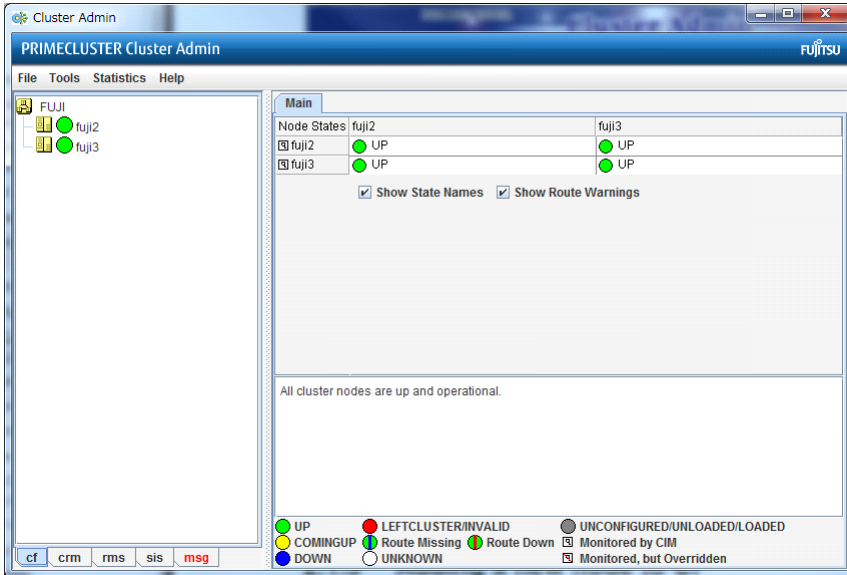


Figure 23: Main CF window

2.1.5 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 the *PRIMECLUSTER Software Release Guide and Installation Guide* for additional details on Web-Based Admin View configuration options.

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

The first window that Cluster Admin displays is the small initial connection pop-up window (see Figure 6). 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 8). 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 also run the SF Wizard to configure the Shutdown Facility on the new node.

You will also need to do additional configuration work for other PRIME-CLUSTER products you might be using such as CRM, RMS, GDS, GFS, and so forth.

2.2 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.

The `qip.cf` configuration file typically contains configuration information for all CIP interfaces on all nodes in the cluster.

- 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`.

The IP address can also have additional options following it. These options are passed to the configuration command `ifconfig`. Each option is separated from the IP address and other option by colons (:).

- 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 expressed in the Internet standard colon-separated hexadecimal format and prefix length.

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

For example, the CIP configuration done in Section “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.

If you make changes to the `cip.cf` file by hand, you should be sure that the file exists on all nodes, and all nodes are specified in the file. Be sure to update all nodes in the cluster with the new file. Changes to the CIP configuration file will not take effect until CIP is stopped and restarted. If you stop CIP, be sure to stop all applications that use it. In particular, RMS needs to be shut down before CIP is stopped.

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

For instructions on starting and stopping CF, see Section “Starting and stopping CF”.

2.3 Cluster Configuration Backup and Restore (CCBR)



Caution

CCBR only saves PRIMECLUSTER configuration information. It does not replace an external, full backup facility.

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



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 add 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 command must be executed as `root`. `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.



You cannot run `cfbackup(1M)` and `cfrestore(1M)` at the same time on the same node.



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/MAW/ccbr/plugins` directory contains executable CCBR plugins. The installed PRIMECLUSTER products supply them.
- The `/opt/MAW/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/MAW/MAWccbr` is supplied as part of the `MAWccbr` package.

The system administrator can change the `CCBRHOME` pathname at anytime. It is recommended that the system administrator verify that there is enough disk space available for the archive file before setting `CCBRHOME`. The system administrator might need to change the `CCBRHOME` pathname to a file system with sufficient disk space.



It is important to remember that re-installing the `MAWccbr` package will reset the contents of the `/opt/MAW/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/MAW/MAWccbr
export CCBRHOME
```

- The `/opt/MAW/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:

`CCBRHOME/nodename_ccbrN`

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

`CCBR00T` 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.
- The `CCBR00T/err.log` log file contains specific error information when a plug-in fails. All the messages are time stamped.

- The `CCBR00T/plugin.blog` or `CCBR00T/plugin.rlog` log files contain startup and completion messages from each backup/restore attempt for each plug-in. These messages are time stamped.

cfbackup example

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:

```
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 the Chapter “Diagnostics and troubleshooting” for more details on troubleshooting CCBR.

cfrestore example

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 `CCBR00T/root` directory. They are copied from the `CCBR00T/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.

3 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).

This chapter discusses the following:

- The Section “CF Registry” discusses the purpose and physical characteristics of the CF synchronized registry.
- The Section “Cluster Integrity Monitor” discusses the purpose and implementation of CIM.

3.1 CF Registry

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.

3.2 Cluster Integrity Monitor

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 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 (`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 state.

`False`—One or more CIM nodes in the cluster are in an unknown state.

3.2.1 Configuring CIM

You can perform CIM procedures through the following methods:

- Cluster Admin GUI—This is the preferred method of operation. Refer to the Section “Adding and removing a node from CIM” for the GUI procedures.
- CLI—Refer to the Chapter “Manual pages” for complete details on the CLI options and arguments, some of which are described in this section. For more complete details on CLI options and arguments, refer to the manual page. The commands can also 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`. If there is none, it returns as error. 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.

3.2.2 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 `rcqquery(1M)` at any time. Refer to the Chapter “Manual pages” for complete details on the CLI options and arguments.

`rcqquery(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 of nodes 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.

3.2.3 Reconfiguring quorum

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

CLI

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

When the configuration information is given to the command `rcqconfig(1M)` as arguments, it performs the transaction to CFREG to update the configuration information. The rest of the configuration procedure is the same. 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.

Examples

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.
```

Since CF only configured the cluster to consist of fuji2 and fuji3, fuji10 does not exist. The quorum set remains empty.

```
fuji2# rcqconfig -g
```

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

4 Cluster resource management

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

This chapter discusses the following:

- The Section “Overview” introduces cluster resource management.
- The Section “Kernel parameters for Resource Database” discusses the default values of the Solaris kernel which have to be modified when the Resource Database is used.
- The Section “Resource Database configuration” details how to set up the Resource Database for the first time on a new cluster.
- The Section “Registering hardware information” explains how to register hardware information in the Resource Database.
- The Section “Start up synchronization” discusses how to implement a start up synchronization procedure for the Resource Database.
- The Section “Adding a new node” describes how to add a new node to the Resource Database.

4.1 Overview

The cluster Resource Database is a dedicated database used by PRIME-CLUSTER products. It is not a general purpose database which a customer can use for their own applications.

4.2 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.



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 PRIMECLUSTER *Software Release Guide and Installation Guide* for details on meanings and methods of changing kernel parameters.



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

Table 1 shows the value of a kernel parameter required to use the resource database.

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

Table 1: Kernel parameter values

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:

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

- `shminfo_shmmax` is defined in `/etc/system`
 - If the current value is greater than *TotalResources* and also greater 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:

To take Figure 24 as an example, the following article describes how to calculate the total resources.

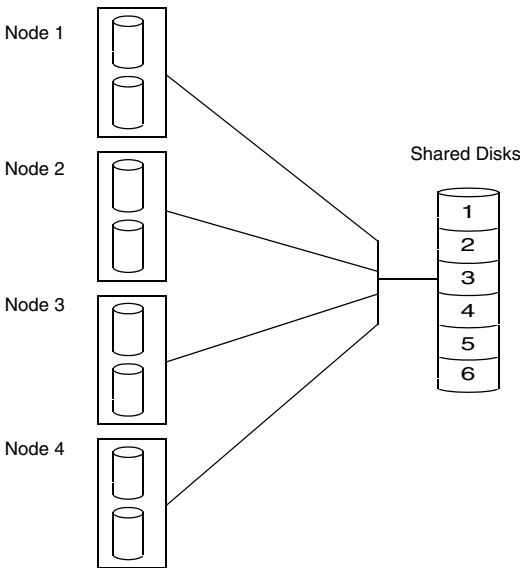


Figure 24: Cluster resource diagram

Referring to Figure 24, calculate the total resources as follows:

1. Remote resources:

`DISKS=6, NODES=4`
`remote resources = 6 x (4+1) x 2 = 60`

2. Local resources:

$$\text{Local resources} = 2 \times 4 = 8$$

3. Total resources:

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

Since 1237344 is less than 4194394, it is necessary to set 4194394 for `shminfo_shmmax`. If the value in step 3 is greater than 4194394, set the value for `shminfo_shmmax`.

4.3 Resource Database configuration

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

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 the Section “Adding a new node” for details.

Before you begin configuring the Resource Database, you must first make sure that CIP is properly configured on all 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 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 nodes are specified and they are the same across the cluster.

In general, the CIP configuration is fairly simple. You can use the Cluster Admin CF Wizard to configure a CIP subnet after you have configured CF. If you use the Wizard, then you will not need to do any additional CIP configuration. See the Section “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. You can use the `ping(1M)` command to test CIP network connectivity. The file `/etc/cip.cf` contains the CIP names that you should use in the `ping(1M)` 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:

*cfname*RMS

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 nodes.

If either of the above conditions are 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.

4.4 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:

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

The command automatically detects the information. Refer to the Chapter “Manual pages” for additional details on this command.

4.4.1 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
c1t0d16
c1t0d17
c1t0d18
c1t0d19
.....
emcpower63
emcpower64
emcpower65
emcpower66
```

Refer to the Section “Exclusive device list for Dell EMC Symmetrix” if you use the Dell EMC Symmetrix series of RAID devices (Symmetrix) in a SPARC Enterprise M-series/PRIMECLUSTER environment.

4.4.2 Exclusive device list for Dell 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 Dell EMC Symmetrix series of RAID devices (Symmetrix) is used in a SPARC Enterprise M-series/PRIMECLUSTER environment (refer to the Section "Setup exclusive device list").

You must exclude the following Dell 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 Dell 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 Dell EMC PowerPath. Then, you can perform automatic resource registration.

4.4.2.1 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 should normally set native 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.

```
emcpower<0>
(not a registration target) —┬─ c<2>t<0>d<0>
                             └─ c<3>t<0>d<0> (not a registration target)
```

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.

```
emcpower<0> —┬─ c<2>t<0>d<0> (not a registration target)
                └─ c<3>t<0>d<0> (not a registration target)
```

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

4.4.2.2 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.

4.4.2.3 VC MDB

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

4.4.2.4 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 `clmake-diskinfo`. 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 Dell EMC PowerPath installation settings. Normally the path is `/etc/powermt`.



Note:

- Dell EMC PowerPath is required to use Dell 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 the *PRIMECLUSTER Global Disk Service 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.

4.4.3 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/FJSCluster/bin/clautoconfig -r
```

3. Confirm registration.

Execute the `clgettree` command for confirmation as follows:

```
# /etc/opt/FJSCluster/bin/clgettree
```

```
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)

4.5 Start up 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 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 nodes when they eventually join the cluster.

To avoid this situation, the Resource Database implements a start up 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 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 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 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/adm/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 start up 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
60
```

This sets the `StartingWaitTime` to 300 seconds.

When you change the `StartingWaitTime` parameter, it is not necessary to stop the existing nodes. The new parameter will be effective for all nodes at the next reboot. Refer to the Chapter “Manual pages” for more details on the possible values for `StartingWaitTime`.

4.5.1 Start up synchronization and the new node

After the Resource Database has successfully been brought up in the new node, then you need to check if the `StartingWaitTime` used in start up 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.

4.6 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.

Figure 25 shows these steps as a flow chart.

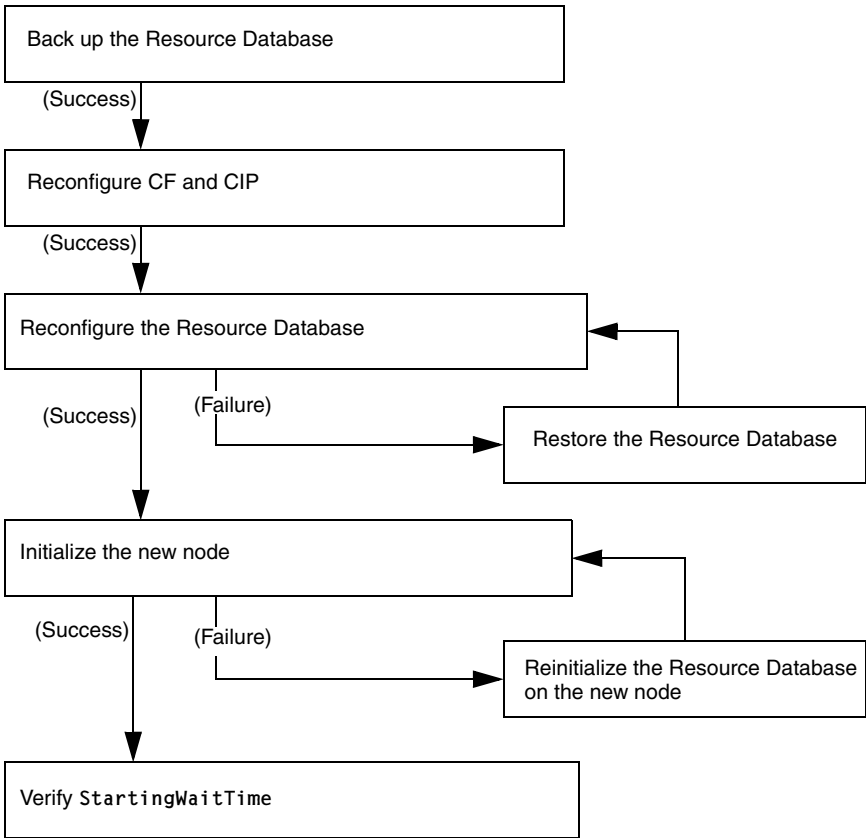


Figure 25: Adding a new node

The sections that follow describe each step in more detail.

4.6.1 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.

i 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.

i 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. The syntax is as follows:

```
/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`).

i 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.

4.6.2 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 syntax in this case is as follows:

```
/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 `fuji4` 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 fuji4 -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 the Section “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.

4.6.3 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. If you used the Cluster Admin CF Wizard to configure CF and CIP for the new node, then CIP should already be properly configured.

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.

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 the Section "Reconfiguring the Resource Database".
3. Run the command `clsetup` with the `-s` option. The syntax for this case is as follows:

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

file is the name of the backup file.

If we continue our example of adding `fujii4` 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.

4.6.4 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 the Section “Start up synchronization” for further information.

4.6.5 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
```



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

4. Mount the local file systems on each node with the following command:

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

5. Restore the Resource Database on each node with the `clrestorerdb` command. The syntax is:

```
# 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
```

5 GUI administration

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

This chapter discusses the following:

- The Section “Overview” introduces the Cluster Admin GUI.
- The Section “Starting Cluster Admin GUI and logging in” describes logging in and shows the first windows you will see.
- The Section “Main CF table” describes the features of the main table.
- The Section “CF route tracking” details the CF route tracking GUI interface.
- The Section “Node details” explains how to get detailed information.
- The Section “Displaying the topology table” discusses the topology table, which allows you to display the physical connections in the cluster.
- The Section “Starting and stopping CF” describes how to start and stop CF.
- The Section “Marking nodes DOWN” details how to mark a node DOWN.
- The Section “Using PRIMECLUSTER log viewer” explains how to use the PRIMECLUSTER log viewer, including how to view and search `syslog` messages.
- The Section “Displaying statistics” discusses how to display statistics about CF operations.
- The Section “Heartbeat monitor” describes how to monitor the percentage of heartbeats that are being received by CF.
- The Section “Adding and removing a node from CIM” describes how to add and remove a node from CIM.
- The Section “Unconfigure CF” explains how to use the GUI to unconfigure CF.
- The Section “CIM Override” discusses how to use the GUI to override CIM, which causes a node to be ignored when determining a quorum.

5.1 Overview

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

5.2 Starting Cluster Admin GUI and logging in

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

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

In this example, if `fujii2` is a management server, enter the following:

```
http://fujii2:8081/Plugin.cgi
```

This brings up the Web-Based Admin View main window (see Figure 26).

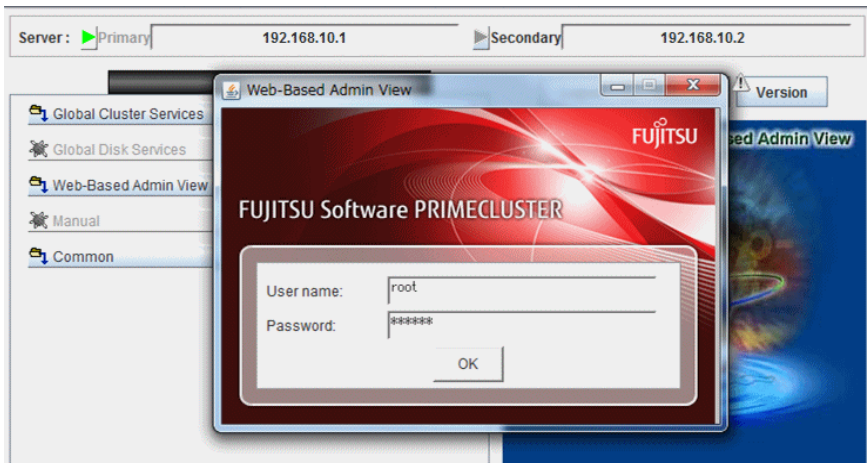



Figure 26: Main window

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. With the operator privileges, you can only perform viewing tasks.

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

Click on the *Global Cluster Services* button and the *Cluster Admin* button appears (see Figure 27).

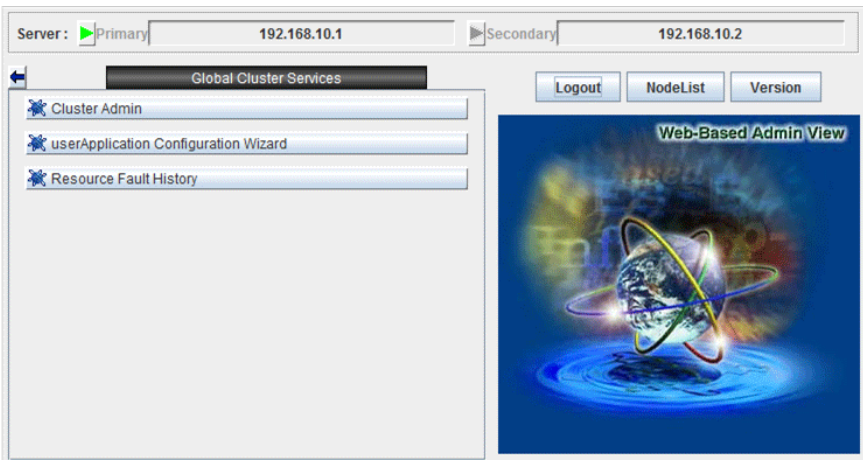


Figure 27: Cluster Admin start-up window

Click on the *Cluster Admin* button.

The *Choose a node for initial connection* window appears (see Figure 28).

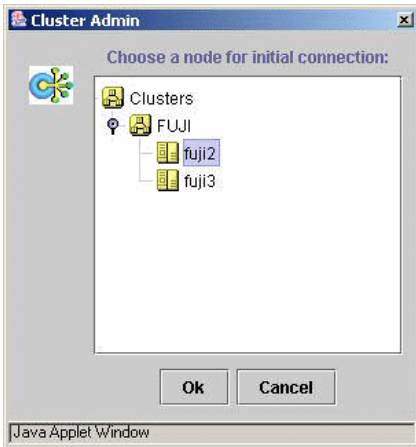


Figure 28: Initial connection choice window

Select a node and click on *Ok*.

The Cluster Admin main window appears (see Figure 29).

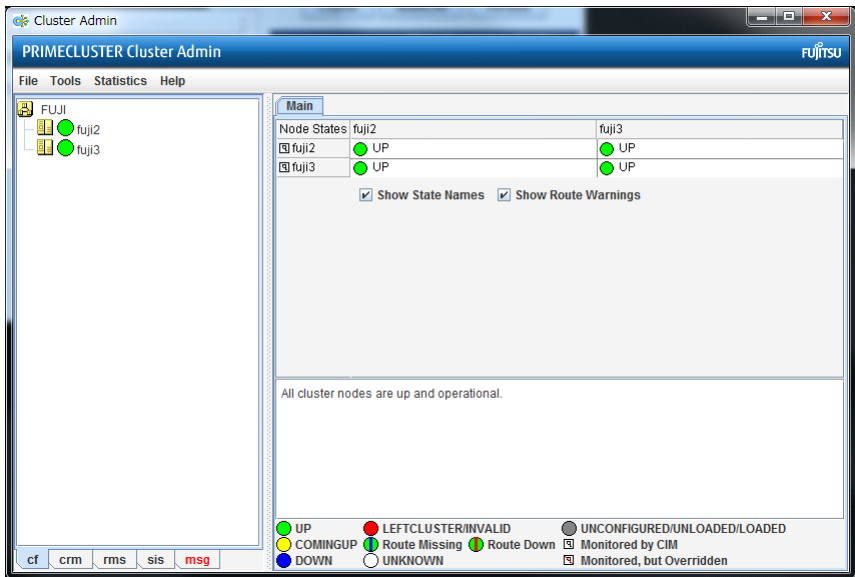


Figure 29: Cluster Admin main window

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.

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

5.3 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. An example of this display is shown in Figure 29.

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 `fujii3` as seen by the node `fujii2`.

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. Table 2 lists the local states.

CF state	Description
UNLOADED	The node does not have a CF driver loaded.
LOADED	The node has a CF driver loaded, but is not running.
COMINGUP	The node is in the process of starting and should be UP soon.
UP	The node is up and running normally.
INVALID	The node has an invalid configuration and must be reconfigured.
UNKNOWN	The GUI has no information from this node. This can be temporary, but if it persists, it probably means the GUI cannot contact that node.
UNCONFIGURED	The node is unconfigured.

Table 2: Local states

Table 3 lists the 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.

Table 3: Remote states

5.4 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 Figure 30).

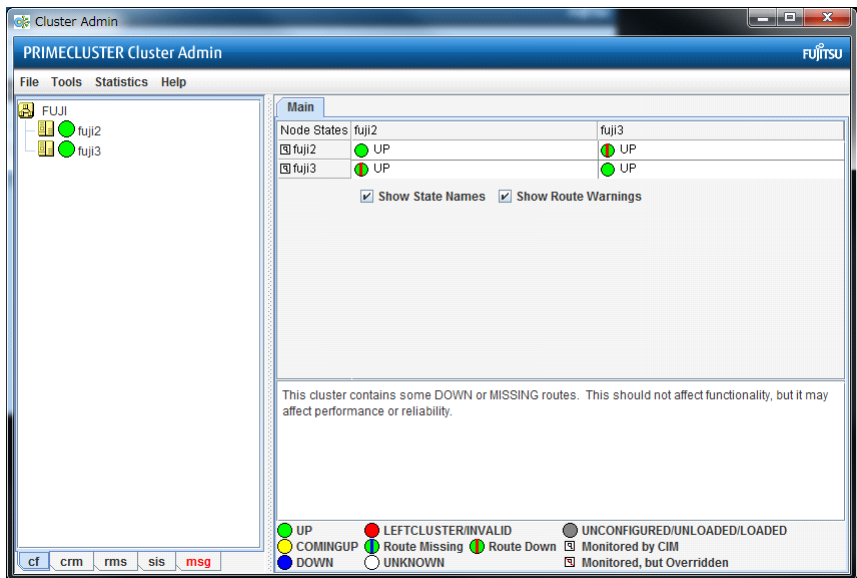


Figure 30: CF route DOWN

In this example, one of the network interfaces on `fujij2` has been unplugged. Cluster Admin, therefore, shows that a route is DOWN. Since `fujij3` cannot contact `fujij2` over that interface, it also shows that there is a route down on `fujij2`. 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 Figure 31).

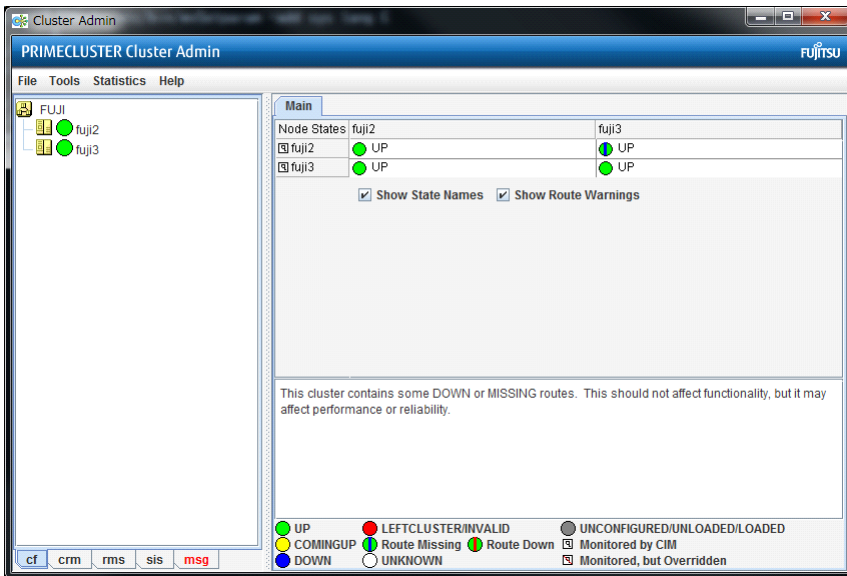


Figure 31: CF interface missing

In Figure 31, `fujij3` has a broken connection to `fujij2`, and Cluster Admin indicates that a route is missing.

In our example, clicking on `fujij2` in the left-panel tree shows that there is no route from `fujij2` to the `e1000g3` interface on `fujij3` (see Figure 32).

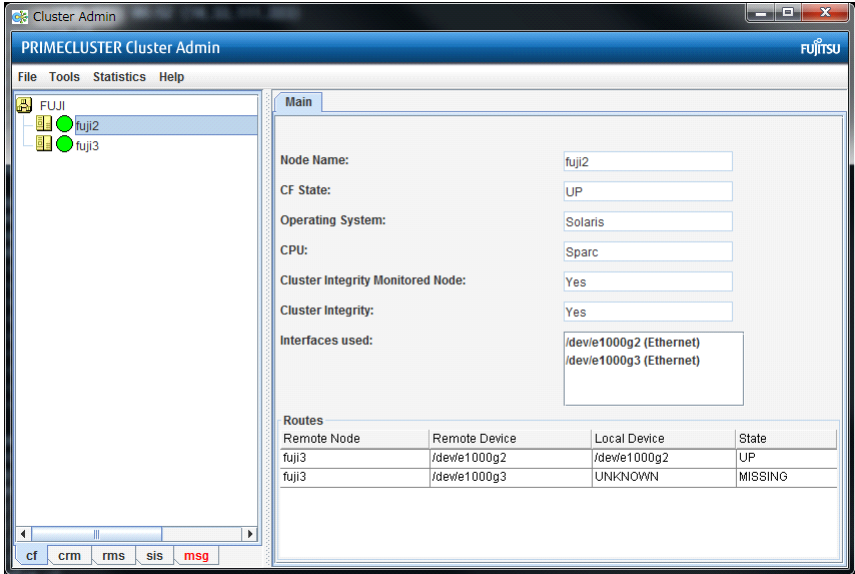


Figure 32: CF route table

5.5 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 to the display in Figure 33.

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	/dew/hme1	/dew/hme1	UP
fuji3	/dew/hme3	/dew/hme3	UP
fuji3	/dew/ip0	/dew/ip0	UP

Figure 33: CF node information

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.

5.6 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 in Figure 34.

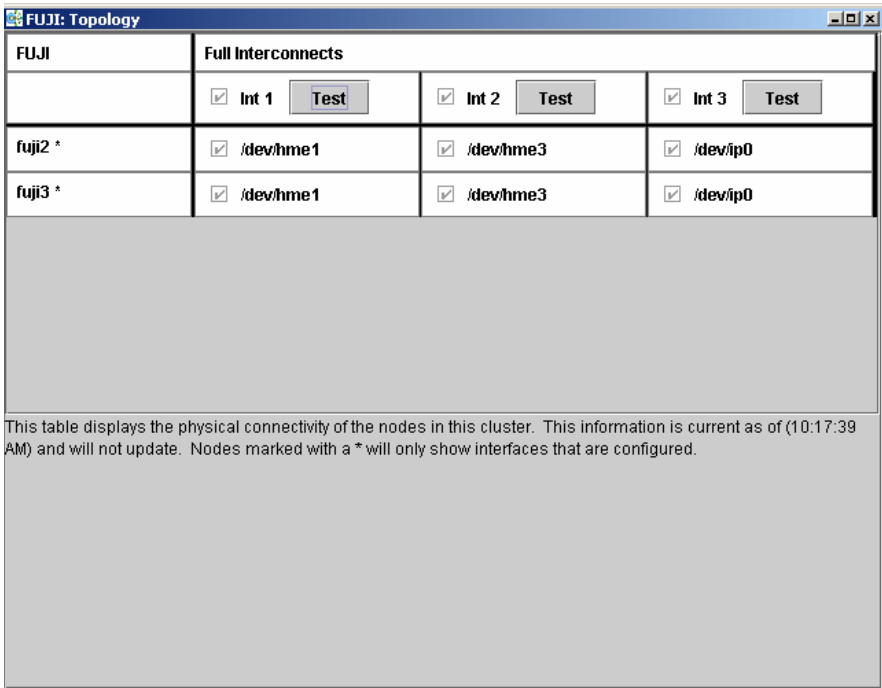


Figure 34: CF topology table

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 Figure 35).

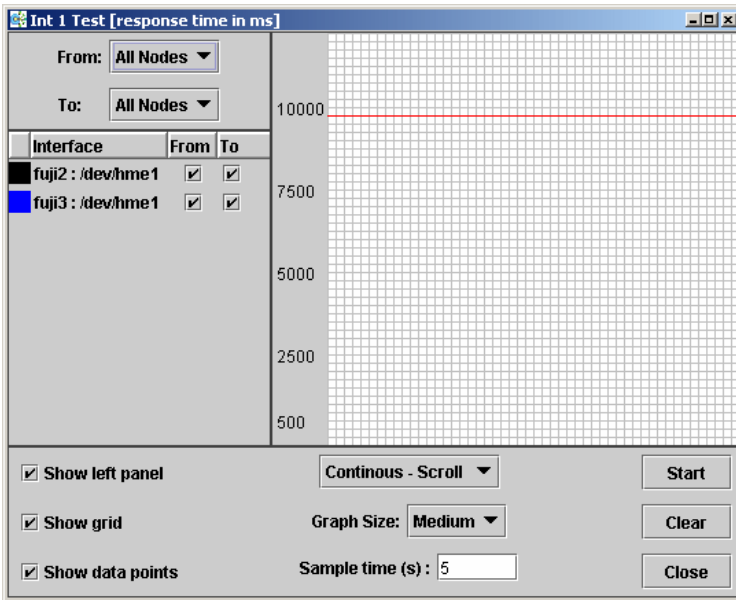


Figure 35: Response Time monitor

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.



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.

5.7 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. Figure 36 shows the content-sensitive menu pop-up when you select *Start CF*.

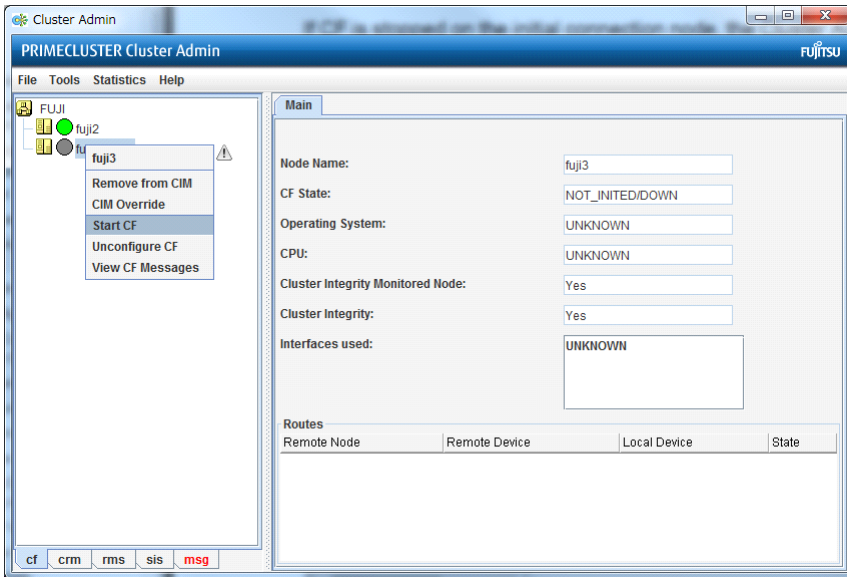


Figure 36: Starting CF

You can also go to the *Tools* 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 CF GUI gets its list of CF nodes from the node used for the initial connection window as shown in Figure 28. 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.

5.7.1 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 Figure 37). The CF state must be UNLOADED or LOADED to start CF on a node.

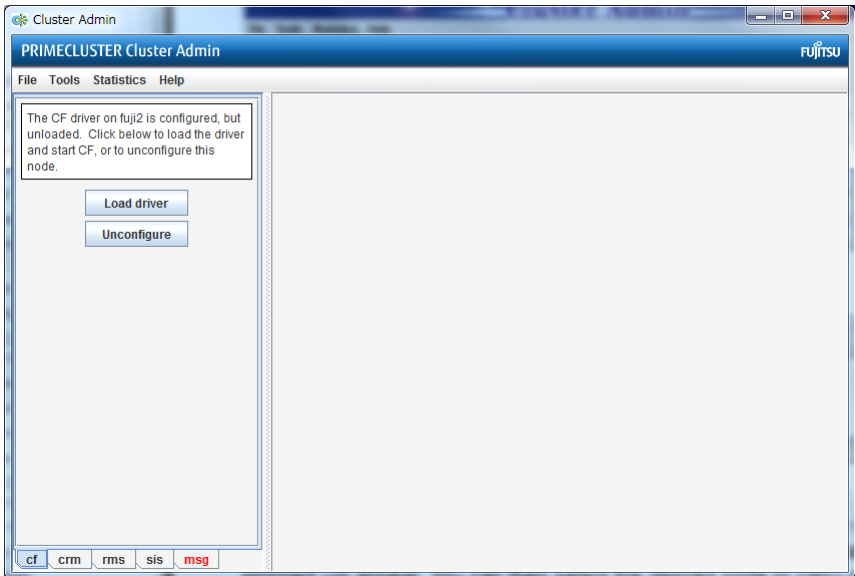


Figure 37: CF configured but not loaded

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

The Start CF services popup appears (see Figure 38). By default all CF services that have been installed on that node are selected to be started. The contents of this list may vary due according to the installed products.



Figure 38: Start CF services pop-up

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

Click on the *Ok* button and a status popup appears with the results of each service start operation (see Figure 39).

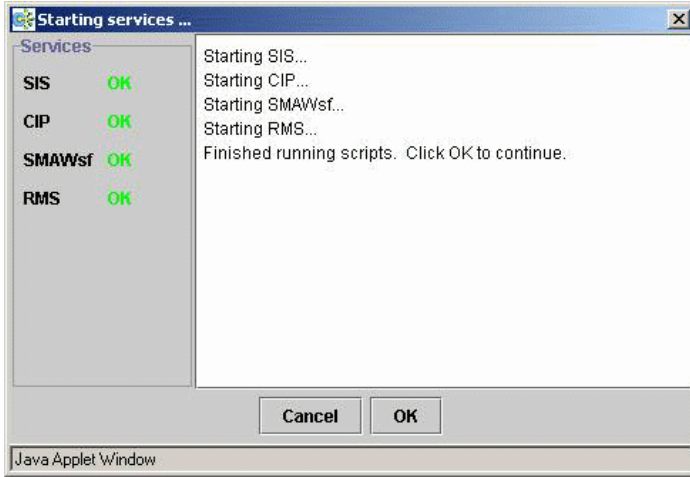


Figure 39: Start CF services status window

Click on the *Ok* button to return to the Cluster Admin main window.

5.7.2 Stopping CF

Right-click on a CF node name and select *Stop CF* (see Figure 40).

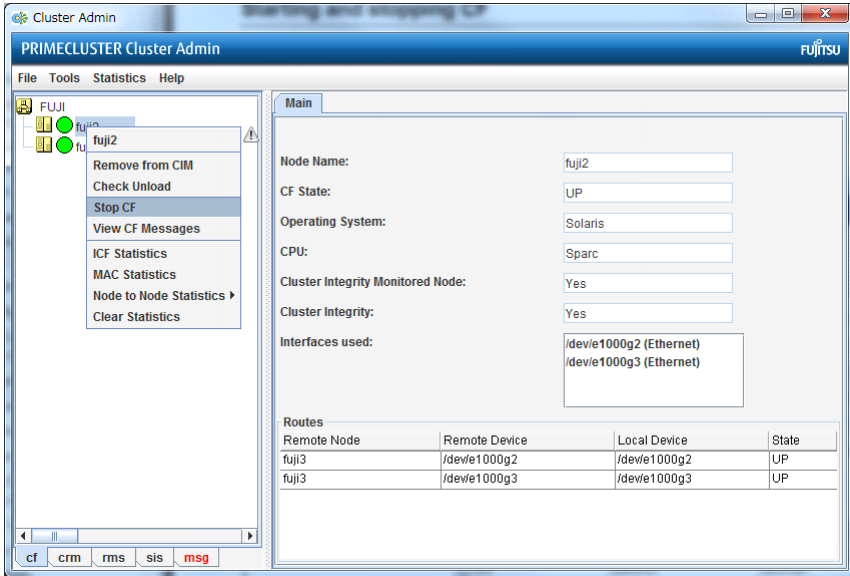


Figure 40: Stop CF

A confirmation pop-up appears (see Figure 41). Choose *Yes* to continue.

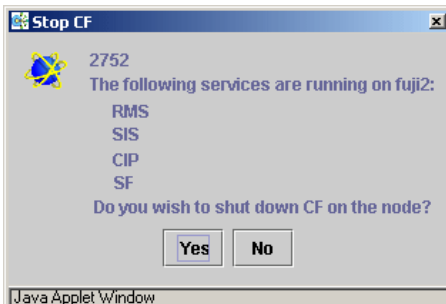


Figure 41: Stopping CF

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 any service does not shutdown, then the *Stop CF* operation will fail.

i 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.

5.8 Marking nodes DOWN

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

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

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

Refer to the Chapter “LEFTCLUSTER state” for more information on the LEFTCLUSTER state.

5.9 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 (see Figure 42).

Figure 42 shows an example of the PRIMECLUSTER log viewer.

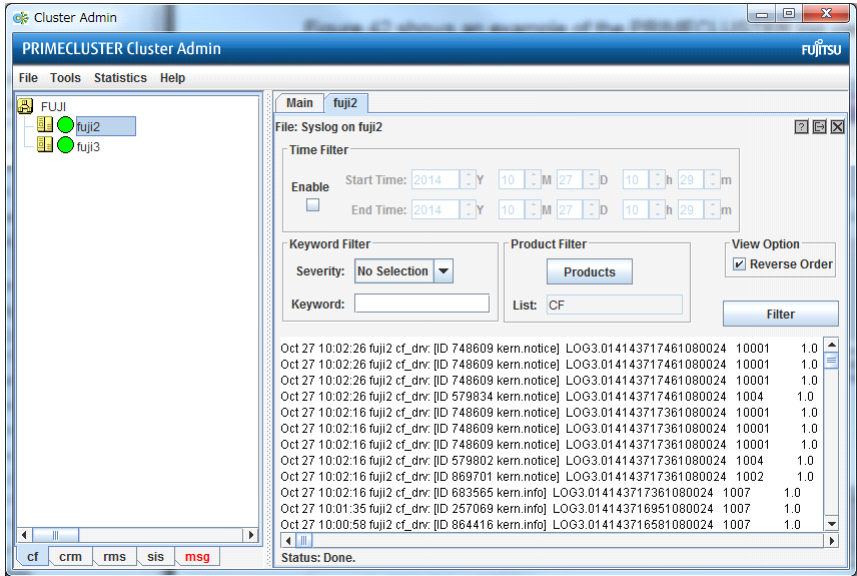


Figure 42: PRIMECLUSTER log viewer

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

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

5.9.1 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 43).

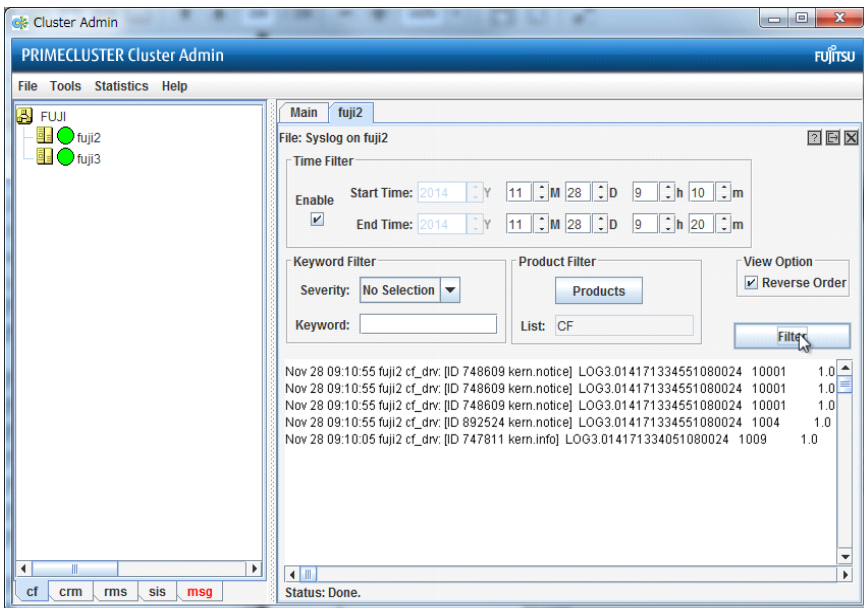


Figure 43: Search based on date/time

5.9.2 Search based on keyword

To perform a search based on a keyword, enter a keyword and click on the *Filter* button (see Figure 44).

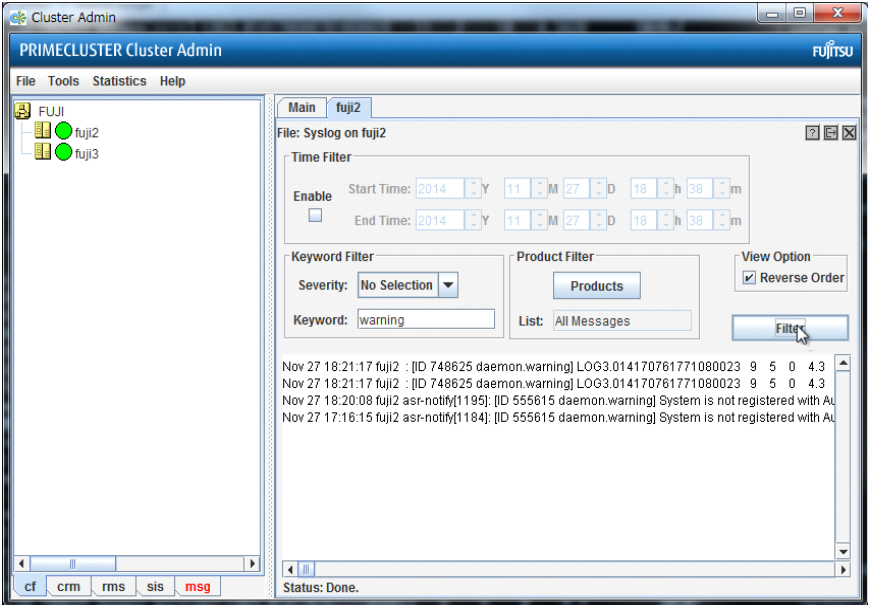


Figure 44: Search based on keyword

5.9.3 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 in Table 4 and click on the *Filter* button. Figure 45 shows the log for a search based on severity level.

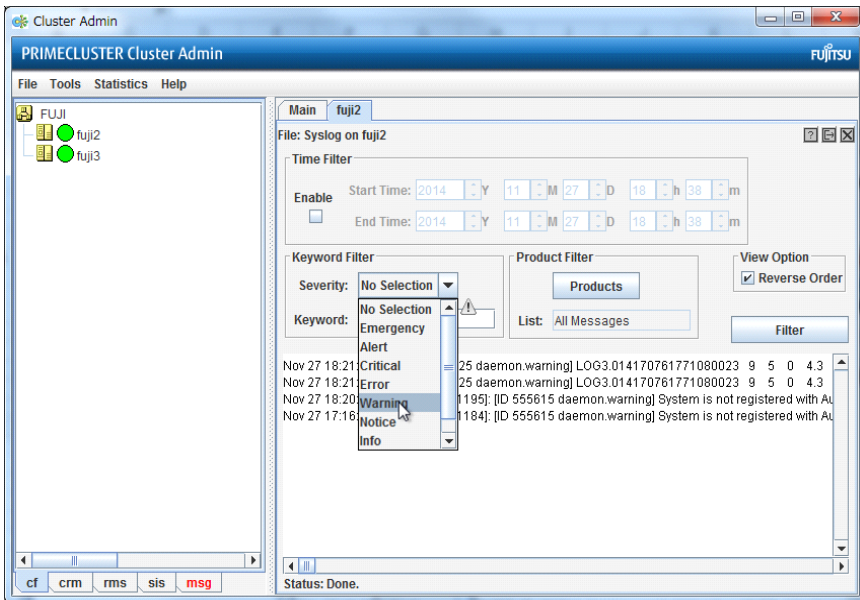


Figure 45: Search based on severity

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

Table 4: PRIMECLUSTER log viewer severity levels

5.10 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 nodes whose states are UP as viewed from the login node.

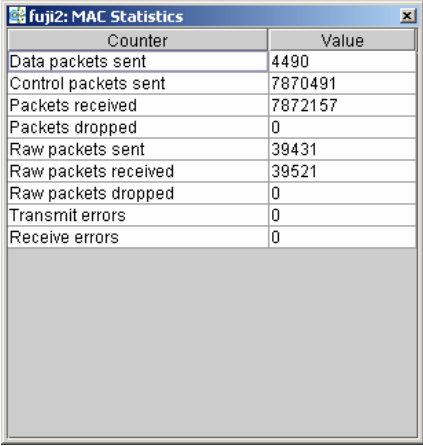
Figure 46 shows the display of ICF Statistics.

The screenshot shows a Java Applet Window titled "fuji2: ICF Statistics". The window contains a table with two columns: "Counter" and "Value". The table lists various network statistics for ICF (Inter-Cluster Forwarding) packets, including counts for transmitted (xmit) and received (rx) packets across different categories like DATA, ENQ, ACK, NACK, HTBT, SYN, SQE, ECHO, and NO_SVC. The values range from 0 to 9657. The window also has a "Java Applet Window" label at the bottom.

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 x...	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

Figure 46: ICF statistics

Figure 47 shows the display of MAC Statistics,



The screenshot shows a window titled "fuji2: MAC Statistics" with a close button in the top right corner. The window contains a table with two columns: "Counter" and "Value". The table lists various network statistics and their corresponding values.

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

Figure 47: MAC statistics

To display node to node statistics, choose *Node to Node Statistics* and click on the desired node (see Figure 48).

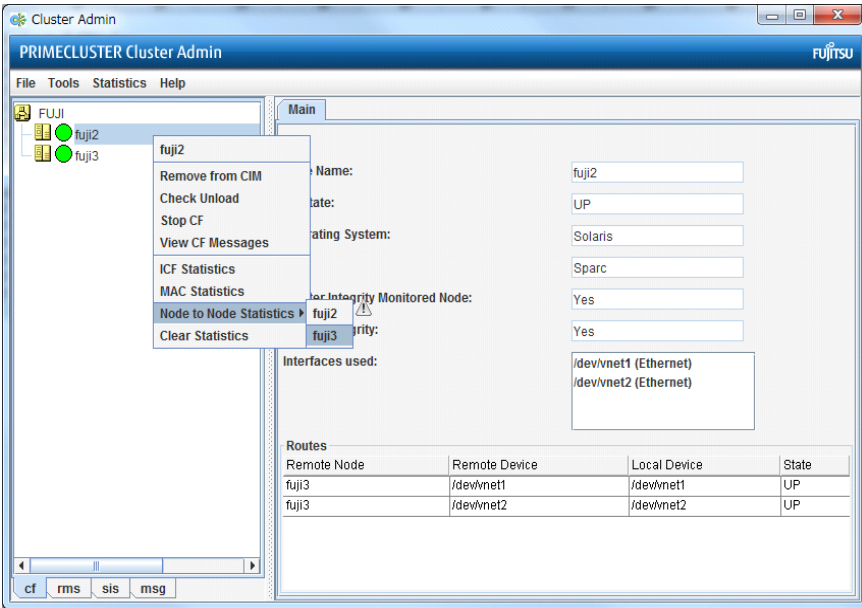


Figure 48: Selecting a node for node to node statistics

The window for Node to Node Statistics appears (see Figure 49).

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_REPLY 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_SVC 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_REPLY packets rx	11319
ICF SYN packets rx	0

Clear Statistics

Java Applet Window

Figure 49: Node to Node statistics

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.

5.11 Heartbeat monitor

To display the Heartbeat monitor, go to the *Statistics* menu and select *Heartbeat Monitor* (see Figure 50).

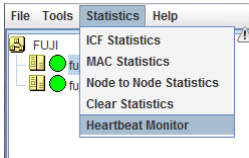


Figure 50: Selecting the Heartbeat monitor

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 Figure 50).

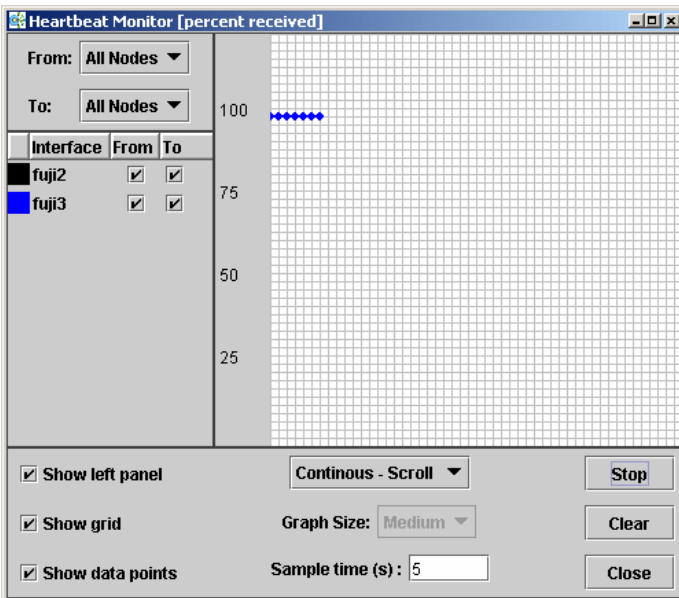


Figure 51: Heartbeat monitor

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.

5.12 Adding and removing a node from CIM

To add a node to CIM, click on the *Tools* pull-down menu. Select *Cluster Integrity* and *Add to CIM* from the expandable pull-down menu (see Figure 52).

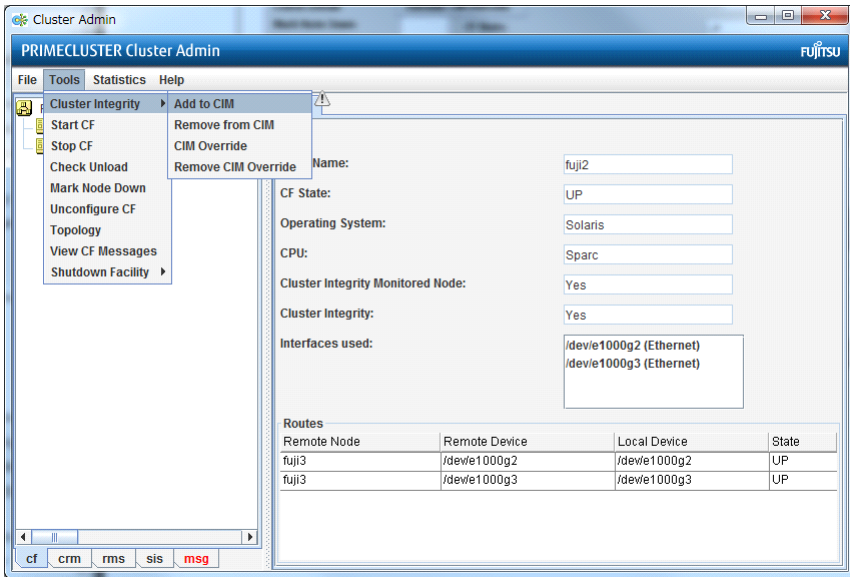


Figure 52: CIM options

The *Add to CIM* pop-up display appears. Choose the desired CF node and click on *Ok* (see Figure 53).

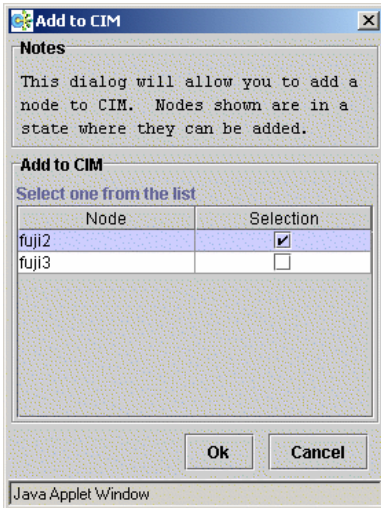


Figure 53: Add to CIM

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

Refer to the Section “Cluster Integrity Monitor” for more details on CIM.

5.13 Unconfigure CF

To unconfigure a CF node, first stop CF on that node. Then, from the *Tools* pull-down menu, click on *Unconfigure CF*.

The *Unconfigure CF* pop-up display appears. Select the check box for the CF node to unconfigure, and click on *Ok* (see Figure 51).

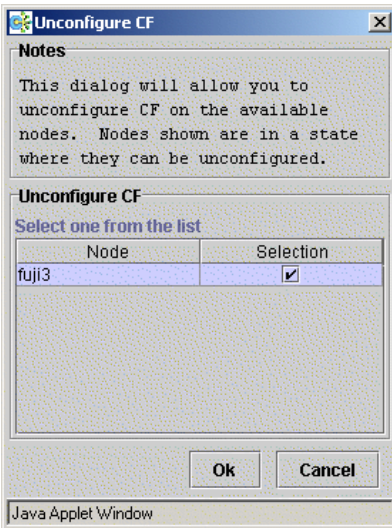


Figure 54: Unconfigure CF

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

5.14 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 Figure 55).

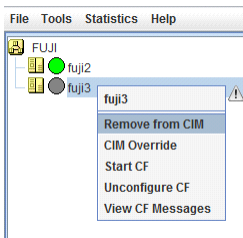


Figure 55: CIM Override

A confirmation pop-up appears (see Figure 56).

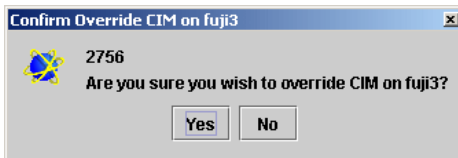


Figure 56: CIM Override confirmation

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 Figure 56).

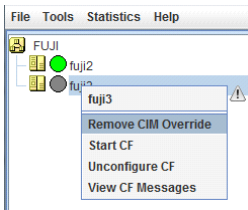


Figure 57: Remove CIM Override

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

6 LEFTCLUSTER state

This chapter defines and describes the `LEFTCLUSTER` state.

This chapter discusses the following:

- The Section “Description of the `LEFTCLUSTER` state” describes the `LEFTCLUSTER` state in relation to the other states.
- The Section “Recovering from `LEFTCLUSTER`” discusses the different ways a `LEFTCLUSTER` state is caused and how to clear it.

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.



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.

6.1 Description of the LEFTCLUSTER state

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

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

For example, consider the three-node cluster shown in Figure 58.

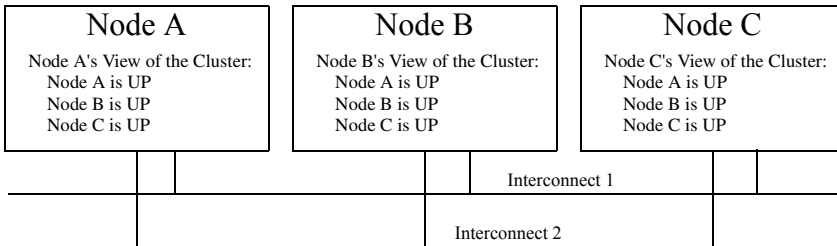


Figure 58: Three-node cluster with working connections

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 Figure 59.

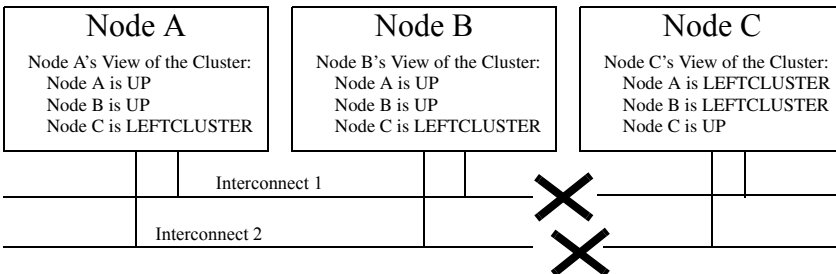


Figure 59: Three-node cluster where connection is lost

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.



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 59, each node believes that it is UP.

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

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

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

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

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

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

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



Note that a node cannot join an existing cluster when the nodes in that cluster believe that the node is in the LEFTCLUSTER state.

6.2 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.

6.2.1 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.

6.2.2 Caused by staying in the kernel debugger too long

In Figure 60, 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.

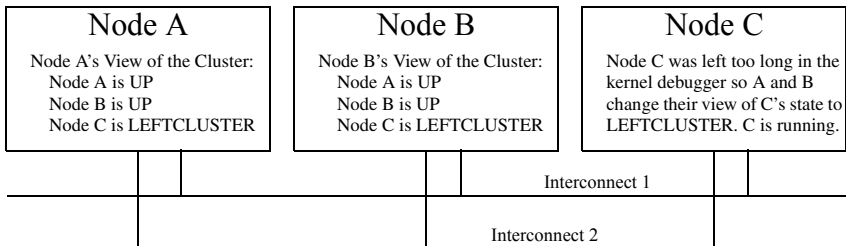


Figure 60: Node C placed in the kernel debugger too long

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.

6.2.3 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.

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 shut down 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 Figure 61.

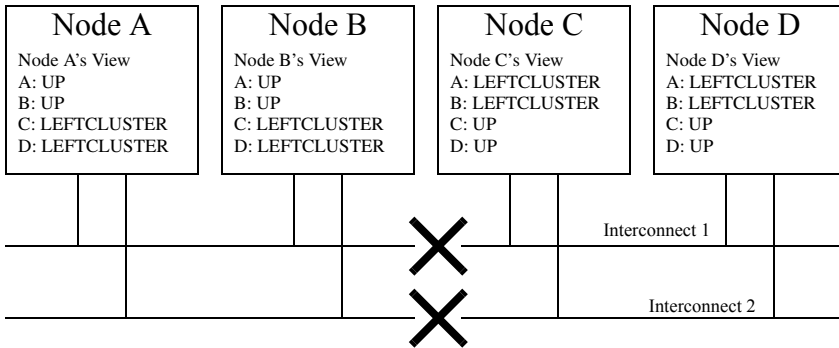


Figure 61: Four-node cluster with cluster partition

In Figure 61, 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**.
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.

6.2.4 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.

7 CF topology table

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

This chapter discusses the following:

- The Section “Basic layout” discusses the physical layout of the topology table.
- The Section “Selecting devices” discusses how the GUI actually draws the topology table.
- The Section “Examples” shows various network configurations and what their topology tables would look like.

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

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

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

Displayed devices

The topology table is generated by doing CF pings on all 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.

7.1 Basic layout

The basic layout of the topology table is shown in Table 5.

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

Table 5: Basic layout for the CF topology table

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 5 had no partial interconnects, then the table headings would list only full interconnects and unconnected devices (as well as the left-most column giving the clustername and node names).

Within the full interconnects and partial interconnects category, the devices are further sorted into separate interconnects. Each column under an *Int* number heading represents all the devices on an interconnect. (The column header *Int* is an abbreviation for *Interconnect*.) For example, in Table 5, 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 5, Table 5 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 5 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 5, `fuj3` does not have a device for the partial interconnect labeled `Int 3`.

7.2 Selecting devices

The basic layout of the topology table is shown in Table 5. 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 Table 6.

FUJI	Full interconnects		Partial interconnects		Unconnected devices
	<input checked="" type="checkbox"/> Int 1	<input checked="" type="checkbox"/> Int 2	<input type="checkbox"/> Int 3	<input type="checkbox"/> Int 4	
<code>fuj2</code>	<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
<code>fuj3</code>	<input checked="" type="checkbox"/> hme0	<input checked="" type="checkbox"/> hme2	missing	<input type="checkbox"/> hme1	
<code>fuj4</code>	<input checked="" type="checkbox"/> hme1	<input checked="" type="checkbox"/> hme2	<input type="checkbox"/> hme3	missing	<input type="checkbox"/> hme4

Table 6: Topology table with check boxes shown

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

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

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

For example, in Table 6, `fujii2` 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 `fujii2`. In the above example, if `hme2` were selected for `fujii2`, 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.

7.3 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 (see Figure 62).

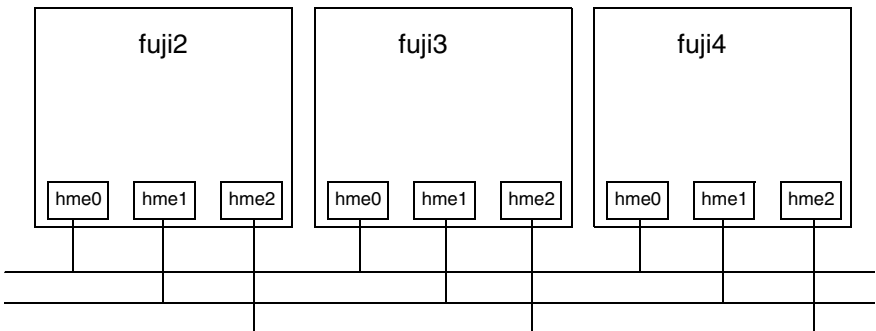


Figure 62: A three-node cluster with three full interconnects

The resulting topology table for Figure 62 is shown in Table 7.

FUJI	Full interconnects		
	Int 1	Int 2	Int 3
fujj2	hme0	hme1	hme2
fujj3	hme0	hme1	hme2
fujj4	hme0	hme1	hme2

Table 7: Topology table for 3 full interconnects

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

Example 2

In this example, fujj2's Ethernet connection for hme1 has been broken (see Figure 63).

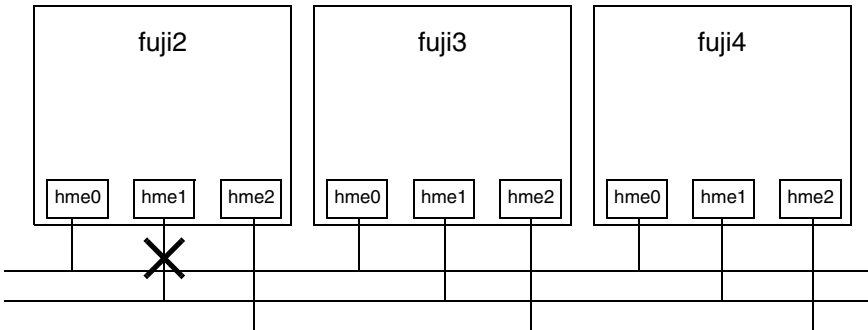


Figure 63: Broken Ethernet connection for hme1 on fujj2

The resulting topology table for Figure 63 is shown in Table 8.

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	

Table 8: Topology table with broken Ethernet connection

In Table 8, 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 7 and Table 8. In Table 7, for example, all hme1 devices were on Int 2. In Table 8, 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.

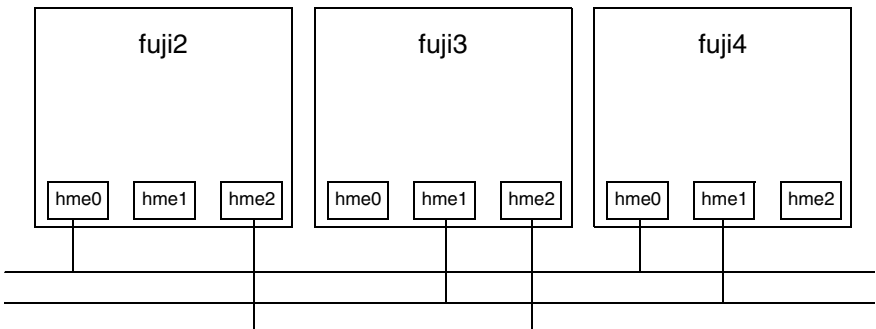


Figure 64: Cluster with no full interconnects

The resulting topology table for Figure 64 is shown in Table 9.


FUJI	Partial interconnects			Unconnected devices
	Int 1	Int 2	Int 3	
fujj2	hme0	missing	hme2	hme1
fujj3	missing	hme1	hme2	hme0
fujj4	hme0	hme1	missing	hme2

Table 9: Topology table with no full interconnects

In Table 9, 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.

8 Shutdown Facility

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

 Certain product options are region-specific. For information on the availability a specific Shutdown Agent (SA), contact your local customer-support service representative.

This chapter discusses the following:

- The Section “Overview” describes the components of SF.
- The Section “Available SAs” describes the available agents for use by the SF.
- The Section “SF split-brain handling” describes the methods for resolving split-brain situations.
- The Section “Configuring the Shutdown Facility” describes the configuration of SF and its agents.
- The Section “SF administration” provides information on administering SF.
- The Section “Logging” describes the log files used by SF and its agents.

8.1 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.

- `sdttool(1M)` command

The `sdttool(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 PRIME-CLUSTER 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.

8.2 Configuring SF

This section describes how to configure SF.

8.2.1 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.

8.2.2 Configuration file of SF



See the format of the configuration file described below as a reference. How to configure the shutdown agent is described in the Section “Configuring the Shutdown Facility”.

Create the configuration file in the `/etc/opt/SMAW/SMAWsf` directory. The configuration file name should be `rcsd.cfg`.

Here is the format of the configuration file.

```
CFName[ ,weight=weight][ ,admIP=myadmIP]:agent=SA_name,timeout=SA_timeout1: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/MAW/SMAWsf` directory. This file is the sample configuration file of the shutdown daemon configured by a dummy machine and an agent.

8.3 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
- RPDU—Remote Power Distribution Unit
- NPS—Network Power Switch (Unsupported)

SA	Name	Hardware
RCI	SA_pprcip, SA_pprcir	SPARC Enterprise M-series

Table 10: Available SAs

SA	Name	Hardware
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
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, SPARC T4, T5, T7, S7 series
RPDU	SA_rpdu	SPARC Enterprise M3000, M4000, M5000, SPARC M10-1, and M10-4

Table 10: Available SAs

8.3.1 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.



- 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.

8.3.2 XSCF

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



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 complete 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



- 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 firmwareIf 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.

8.3.3 XSCF SNMP

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

For details on XSCF, see the "SPARC M-Series System Operation and Administration Guide."

Setup and configuration

The XSCF must be configured according to the "SPARC M-Series 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_xscf SNMPg0p—panics the domain by using XSCF-LAN#0
- SA_xscf SNMPg1p—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
```



- 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.

8.3.4 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/1og/SA_sunF.1og
```



- 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.

8.3.5 ILOM

ILOM SA 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_ilmomp—panics the node through ILOM.
- SA_ilmomr—resets the node through ILOM.

The ILOM log file is as follows:

```
/var/opt/SMAWsf/log/SA_ilmomp.log  
/var/opt/SMAWsf/log/SA_ilmomr.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.

8.3.6 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 the "SPARC M10 Systems 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 "PRIME-CLUSTER 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 GuideMoreover, 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 "PRIME-CLUSTER 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.

8.3.7 RPDU

RPDU shutdown agent is the shutdown agent for SPARC Enterprise M3000, M4000, M5000, SPARC M10-1, and M10-4.

The RPDU shutdown agent controls the Remote Power Distribution Unit, and switches off the main power and the redundant power to forcibly stop the node and then, to switch the cluster system.

The use of the RPDU shutdown agent is optional.

Setup and configuration

For how to install and set up the Remote Power Distribution Unit, and how to set up the RPDU shutdown agent, see the following manuals:

- The manual of the Remote Power Distribution Unit
- "Appendix J Using Remote Power Distribution Unit" in "PRIMECLUSTER Installation and Administration Guide"
 - J.1 Design
 - J.2 Setup Procedure

Shutdown Agent

SA_rpdu - Controlling the Remote Power Distribution Unit to switch off the main power and the redundant power of the node.

RPDU log file

/var/opt/SMAwsf/log/SA_rpdu.log



- Connect only cluster node to the Remote Power Distribution Unit. When connecting peripherals including the shared disk, prepare another Remote Power Distribution Unit exclusive for controlling the peripherals.
- Connect the LAN of the Remote Power Distribution Unit to the administrative LAN.
- When using the RPDU shutdown agent with SPARC Enterprise M4000 and M5000, all the partitions of the forcibly stopped node are also stopped forcibly. Each partition is not shut down.

8.3.8 NPS

The Network Power Switch (NPS) SA is SA_wtinps. This SA provides a node shutdown function using the Western Telematic Inc. Network Power Switch (WTI NPS) unit to power-cycle selected nodes in the cluster.

Setup and configuration

The WTI NPS unit must be configured according to the directions in the manual shipped with the unit. At the very least, an IP address must be assigned to the unit and a password must be enabled. Make sure that the cluster node's power plugs are plugged into the NPS box and that the `command confirmation` setting on the NPS box is set to `on`.

It is advisable to have the NPS box on a robust LAN connected directly to the cluster nodes.

The boot delay of every configured plug in the NPS box should be set to 10 seconds.

i If you want to set the boot delay to any other value, make sure that the "timeout value" for the corresponding `SA_wtinps` agent should be set such that it is greater than this boot delay value by at least 10 seconds. To set this value, use the detailed configuration mode for SF.

i If more than a single plug is assigned to a single node (which means that more than one plug will be operated per `/on`, `/off`, `/boot` command), the "boot delay" of these plugs must be assigned to a value larger than 10 seconds, otherwise timeouts may occur. The timeout value of the corresponding `SA_wtinps` should be set as follows:

```
timeout = boot_delay + (* 2 * no of plugs) + 10
```

The NPS log file is as follows:

```
/var/opt/SMAWsf/log/SA_wtinps.log
```

8.4 SF split-brain handling

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

8.4.1 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.

8.4.2 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.

8.4.2.1 RMS ShutdownPriority attribute

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

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

8.4.2.2 Shutdown Facility weight assignment

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

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



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 the Section “Configuration notes” for help you with the configuration.

8.4.2.3 Disabling split-brain handling

Some applications require a fast failover; however, SF split-brain handling can cause a failover delay. For such applications, it is recommended that you disable the split-brain handling in the `SMAWsf` software.

To disable split-brain handling, the `/etc/opt/SMAW/SMAWsf/nsbm.cfg` file must be present consistently on all cluster hosts and readable by the root user. The contents of this file does not matter; however, it must be present or absent consistently on all cluster hosts.

8.4.3 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.

8.4.4 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 (LSS)
- Specific Hardware Survival (SHS)
- Specific Application Survival (SAS)

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, LSS 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 appli-

ation(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.

8.5 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*.

8.6 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.

8.6.1 Starting and stopping SF

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

- Manually via the CLI
- Automatically via the `rc` script interface

8.6.1.1 Starting and stopping SF manually

SF may be manually started or stopped by using the `sdt00(1M)` command. The `sdt00(1M)` command. Refer to the Chapter “Manual pages” for more information on CLI commands.

8.6.1.2 Starting and stopping SF automatically

SF can be started automatically using the `S64rcfs` RC-script available under the `/etc/rc2.d` directory. The `rc start/stop` script for SF is installed as `/etc/init.d/RC_sf`.

8.7 Logging

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

```
# sdt00l -d on
```

This will dump 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 `sdt00l -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.

9 CF over IP

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

This chapter discusses the following:

- The Section “Overview” introduces CF over IP and describes its use.
- The Section “Configuring CF over IP” details how to configure CF over IP.

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

9.1 Overview

i 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.

i 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. This is illustrated in Figure 65.

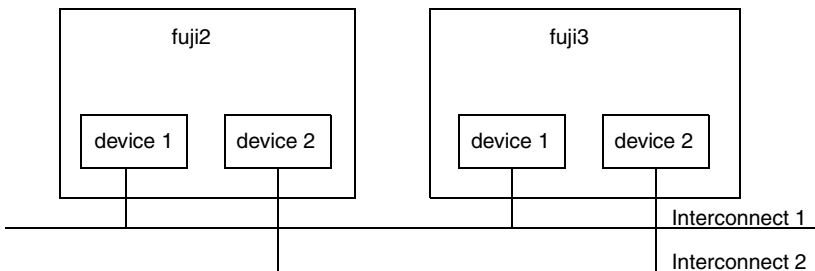


Figure 65: Conceptual view of CF interconnects

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 Figure 66.

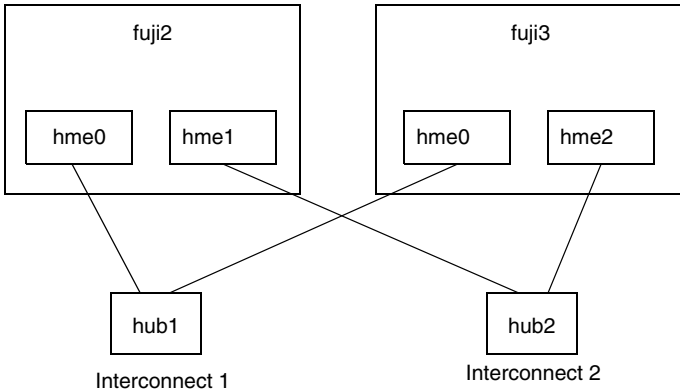


Figure 66: CF with Ethernet interconnects

When CF is run over IP, IP interfaces are the devices used to connect to the interconnect. The interconnect is an IP subnetwork. Multiple IP subnetworks may be used for the sake of redundancy. Figure 67 shows a CF over IP configuration.

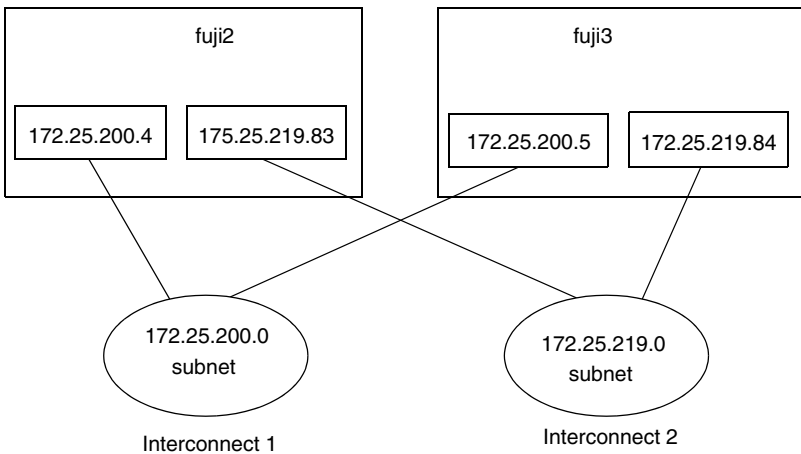


Figure 67: CF with IP interconnects

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.



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



Caution

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.

9.2 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 Section “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 67, 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`.



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.



`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.

10 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*.

This chapter discusses the following:

- The Section “Beginning the process” discusses collecting information used in the troubleshooting process.
- The Section “Symptoms and solutions” is a list of common symptoms and the solutions to the problems.
- The Section “Collecting troubleshooting information” gives steps and procedures for collecting troubleshooting information.

10.1 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* pull-down 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 PRIME-CLUSTER 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 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 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 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 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/fcp0
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 “GUI administration” for a description of the LEFTCLUSTER state 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

10.2 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.

10.2.1 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 *PRIME-CLUSTER Software Release Guide and Installation Guide* 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.



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 -e` 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
```



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 Joystate
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 a fuji4-priv a 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 a fuji4-priv a 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.



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_tmbs 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 1.0cf:ens CF: Icf Error: (service err_type
route_src route_dst). (0 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)
```

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)
```

10.3 Collecting troubleshooting information

If a failure occurs in the PRIMECLUSTER system, collect the following information required for investigations from all cluster nodes. Then, contact your local customer support.

1. Obtain the following PRIMECLUSTER investigation information:

- Use `fjsnap` to collect information required for error investigations.
- Obtaining the system information

Collect the following information that is required when a hard error, an OS error, or a panic occurs, or when you are unable to login to a node.

- Retrieve the system dump.
- XSCF log (for SPARC M10 and SPARC Enterprise M-series only)
- Retrieve the system dump.
- Collect the Java Console on the clients.
See "B.2.2 Java console" in "PRIMECLUSTER Web-Based Admin View Operation Guide."
- Collect screen shots on the clients.
See "B.2.3 Screen hard copy" in "PRIMECLUSTER Web-Based Admin View Operation Guide."

2. In case of application failures, collect such investigation material.

3. If the problem is reproducible, then include a description on how it can be reproduced.



It is essential that you collect the debugging information described in this section. Without this information, it may not be possible for customer support to debug and fix your problem.



Be sure to gather debugging information from all nodes in the cluster. It is very important to get this information (especially the `fjsnap` data) as soon as possible after the problem occurs. If too much time passes, then essential debugging information may be lost.



If a node is panicked, execute `sync` in OBP mode and take a system dump.

10.3.1 Executing the `fjsnap` command

The `fjsnap` command is a system information tool provided with the Enhanced Support Facility `FJSVsnap` package. In the event of a failure in the PRIME-CLUSTER system, the necessary error information can be collected to pinpoint the cause.

Execute the `fjsnap` command as follows:

1. Log in as root.
2. Execute one of the following `fjsnap` commands:

```
# /opt/FJSVsnap/bin/fjsnap -h output
```

```
# /opt/FJSVsnap/bin/fjsnap -a output
```

- As `-a` collects all detailed information, the data is very large. When `-h` is specified, only information relative to PRIMECLUSTER is collected.
- In *output*, specify the special file name or output file name (for example, `/dev/rmt/0`) of the output medium to which the error information collected with the `fjsnap` command is written.

For details about the `fjsnap` command, see the README file included in the FJSVsnap package.



When to run `fjsnap`:

- If an error message appears during normal operation, execute `fjsnap` immediately to collect investigation material.
- If the `fjsnap` command cannot be executed because the system hangs, collect a system dump. Then start the system in single user mode, and execute the `fjsnap` command. To collect the system dump, input the abort key sequence (for example, Break signal) to forcibly stop the node to OBP mode, and then execute `sync`. For detailed instructions on forcibly stopping the node to OBP mode, see the "System Administration Guide" of the Solaris.
- If the necessary investigation material cannot be collected because of a hang, shut down the system, and start the system in single mode. Execute the `fjsnap` command to collect information.
- If the system has rebooted automatically to multi-user mode, then execute the `fjsnap` command to collect information.

10.3.2 System dump

If the system dump is collected while the node is in panicked, retrieve the system dump as investigation material. The system dump is saved as a file during the node's startup process. For details on a system dump, see the "System Administration Guide" of the Solaris.

10.3.3 XSCF log

Collect the XSCF log when the following event occurs.

- The message 7240 or 7241 is output in the Shutdown Facility in SPARC M10 environment.

- How to collect

For how to collect the XSCF log, see "Saving a log to a local USB device" in "SPARC M10 Systems System Operation and Administration Guide."

- The message 7040, 7042, or 7203 is output in the Shutdown Facility in SPARC Enterprise M-series environment.

- How to collect

For how to collect the XSCF log, see "Using the snapshot Tool" in "SPARC Enterprise M3000/M4000/M5000/M8000/M9000 Servers Administration Guide."

11 Manual pages

This chapter lists the online manual pages for CCBR, CF, CFS, CIP, CPAT, Monitoring Agent, PAS, PCS, Resource Database, RMS, RMS Wizards, SF, and Web-Based Admin View.

To display a manual page, type the following command:

```
$ man man_page_name
```

11.1 CCBR

System administration

```
cfbackup
    save the cluster configuration information for a PRIMECLUSTER node

cfrestore
    restore saved cluster configuration formation on a PRIMECLUSTER
    node
```

11.2 CF

System administration

```
cfconfig
    configure or unconfigure a node for a PRIMECLUSTER cluster

cfregd
    CF registry synchronization daemon

cfset
    apply or modify /etc/default/cluster.config entries into the CF
    module

cftool
    print node communications status for a node or the cluster

rcqconfig
    configure or start quorum

rcquery
    get quorum state of the cluster
```

11.3 CFS

- fsck_rcfs
file system consistency check and interactive repair
- mount_rcfs
mount RCFS file systems
- rcfs_fumount
force unmount RCFS mounted file system
- rcfs_list
list status of RCFS mounted file systems
- rcfs_switch
manual switchover or failover of a RCFS file system
- ngadmin
node group administration utility
- cfsmntd
cfs mount daemon for RCFS

11.4 CIP

System administration

- cipconfig
start or stop CIP 2.0
- ciptool
retrieve CIP information about local and remote nodes in the cluster

File format

- cip.cf
CIP configuration file format

11.5 CPAT

System administration

`cluster_uninstall`
remove PRIMECLUSTER software from a system

11.6 Monitoring Agent

System administration

`clrcimonctl`
Start, stop or restart of the RCI monitoring agent daemon, and display of daemon presence

`cldevparam`
changes or displays the tunable parameter of the RCI/RCCU Monitoring Agent

`clrccumonctl`
Start, stop or restart of the console monitoring agent daemon, and display of daemon presence

`clrccusetup`
registers, changes, deletes, or displays console information

11.7 PAS

System administration

`mipcstat`
MIPC statistics

`clmstat`
CLM statistics

11.8 PCS

System administration

`pcstool`

Modifies PCS configurations from the command line

`pcscui`

Character-based interface for PCS

`pcs_reinstall`

Utility for re-integrating PCS with dependent products

11.9 Resource Database



To display a Resource Database manual page, add `/etc/opt/FJsvcluster/man` to the environment variable `MANPATH`.

System administration

`clautoconfig`

execute of the automatic resource registration

`clbackuprdb`

save the resource database

`clexec`

execute the remote command

`cldeldevice`

delete resource registered by automatic resource registration

`clinitreset`

reset the resource database

`clrestorerdb`

restore the resource database

`clsetparam`

display and change the resource database operational environment

`clsetup`

set up the resource database

`clstartresc`
resource activation

`clstopresc`
resource deactivation

`clsyncfile`
distribute a file between cluster nodes

User command

`clgettree`
display the tree information of the resource database

11.10 RMS

System administration

`hvassert`
assert (test for) an RMS resource state

`hvcm`
start the RMS configuration monitor

`hvconfig`
display or save the RMS configuration file

`hvdisp`
display RMS resource information

`hvdump`
collect debugging information about RMS

`hvlogclean`
clean RMS log files

`hvlogcontrol`
control volume of a log disk

`hvsetenv`
manipulate RMS rc start or AutoStartUp

`hvshut`
shut down RMS

hvswitch

switch control of an RMS user application resource to another node

hvutil

manipulate availability of an RMS resource

File formats

config.us

RMS configuration file

hvenv.local

RMS local environment variables file

hvgdstartup

RMS generic detector startup file

11.11 RMS Wizards

RMS Wizards and RMS Application Wizards

RMS Wizards are documented as html pages in the SMAWRhvd0 package on the DVD. After installing this package, the documentation is available in the following directory:

`/usr/opt/reliant/htdocs.solaris/wizards.en`

11.12 SF

System administration

rcsd

Shutdown Daemon of the Shutdown Facility

sdtool

interface tool for the Shutdown Daemon

File formats

rcsd.cfg

configuration file for the Shutdown Daemon

SA_pprci.cfg

configuration file for RCI Shutdown Agent

- SA_rccu.cfg
configuration file for XSCF Shutdown Agent
- SA_sspint.cfg
configuration file for Sun E10000 Shutdown Agent
- SA_sunF.cfg
configuration file for sunF system controller Shutdown Agent
- SA_wtinps.cfg
configuration file for WTI NPS Shutdown Agent

11.13 Web-Based Admin View

System administration

- fjswvbs
stop Web-Based Admin View
- fjswvcnf
start, stop, or restart the web server for Web-Based Admin View
- wvCntl
start, stop, or get debugging information for Web-Based Admin View
- wvGetparam
display Web-Based Admin View's environment variable
- wvSetparam
set Web-Based Admin View environment variable
- wvstat
display the operating status of Web-Based Admin View

12 Release information

This chapter explains primary changes in this manual.

No	VL	Edition	Section	Description
1	4.3A20	December 2012	Section "CF, CIP, and CIM configuration" Section "Differences between CIP and CF over IP" Section "Example of creating a cluster" Section "CIP configuration file"	Added descriptions of IPv6.
2	4.3A20	December 2012	Section "CF security"	Deleted descriptions about the cluster interconnect.
3	4.3A20	December 2012	Section "CF security" Section "Overview"	Stated the range of CF over IP support.
4	4.3A20	December 2012	Section "CIP configuration file"	Changed descriptions when adding the definition of CIP.
5	4.3A20	December 2012	Section "CIP configuration file"	Changed the condition to stop CIP.
6	4.3A20	December 2012	Section "XSCF SNMP"	Added "XSCF SNMP" to Shutdown Agent.
7	4.3A20	December 2012	All	Deleted the description of SIS.
8	4.3A20	February 2013	Section "Example of creating a cluster" Section "Starting Cluster Admin GUI and logging in"	Changed GUIs.

Table 11: Release information

No	VL	Edition	Section	Description
9	4.3A20	February 2013	Section "System dump"	Changed the description for collecting the system dump.
10	4.3A40	June 2015	Section "CF security" Section "RMS"	Deleted the description of the following commands: - hvattr - hvdist - hvgdmake - hvrclev - hvreset - hvthrottle
11	4.3A40	June 2015	Section "Example of creating a cluster" Section "Starting Cluster Admin GUI and logging in" Section "CF route tracking" Section "Starting and stopping CF" Section "Starting CF" Section "Stopping CF" Section "Using PRIME-CLUSTER log viewer" Section "Heartbeat monitor" Section "Adding and removing a node from CIM" Section "CIM Override"	Changed GUI screenshots.
12	4.3A40	June 2015	Section "Cluster Integrity Monitor"	Deleted the description of the RCI method.

Table 11: Release information

No	VL	Edition	Section	Description
13	4.3A40	June 2015	Section "Setting procedure before configuring SF" Section "Available SAs" Section "ILOM" Section "KZONE"	Added the description of Kernel Zones.
14	4.3A40	June 2015	Section "ALOM" Section "ILOM"	Added the description that explains the IP address should belong to the same segment as the Administrative LAN.
15	4.3A40	June 2015	Section "ALOM" Section "ILOM"	Added the description of the time to detect the configuration error.
16	4.3A40	June 2015	Section "ILOM"	Added the description when the No.7004 error message is output.
17	4.3A40	June 2015	Section "RPDU"	Added "RPDU" to the shutdown agent.
18	4.3A40	June 2015	Section "Administrative LAN"	Changed the description of the administrative LAN.
19	4.3A40	June 2015	Section "Collecting troubleshooting information" Section "XSCF log"	Delete the description of SCF dump. Added the description of XSCF log.
20	4.3A40	June 2015	"11 CF messages and codes"	Deleted the whole chapter.

Table 11: Release information

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 hme0 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)*.

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)*.

Glossary

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 a SPARC Enterprise M3000 machine, and another node is a SPARC Enterprise M4000 machine. The models are divided into several groups, which are represented by the SPARC M10-1/M10-4/M10-4S machines, SPARC S7-2/S7-2L machines, SPARC T7-1/T7-2/T7-4 machines, SPARC T5-2/T5-4/T5-8 machines, SPARC T4-1/T4-2/T4-4 machines, SPARC T3-1/T3-2/T3-4, SPARC Enterprise T1000/T2000, SPARC Enterprise T5120/T5220/T5140/T5240/T5440, and the SPARC Enterprise M3000/M4000/M5000/M8000/M9000 machines.

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.

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 GLS.

There is a total of five switching mode types: fast switching mode, NIC switching mode, GS/SURE linkage mode, multipath mode, and multilink Ethernet mode.

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 PRIME-CLUSTER. 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.

Abbreviations

AC

Access Client

API

application program interface

bm

base monitor

CCBR

Cluster Configuration Backup/Restore

CF

Cluster Foundation or Cluster Framework

CIM

Cluster Integrity Monitor

CIP

Cluster Interconnect Protocol

CLI

command-line interface

CRM

Cluster Resource Management

DLPI

Data Link Provider Interface

ENS

Event Notification Services

GDS

Global Disk Services

GFS

Global File Services

Abbreviations

GLS	Global Link Services
GUI	graphical user interface
HA	high availability
ICF	Internode Communication Facility
I/O	input/output
JOIN	cluster join services module
LAN	local area network
MDS	Meta Data Server
MIB	Management Information Base
NIC	network interface card
NSM	Node State Monitor
OE	operating environment
OPS	Oracle Parallel Server
OSD	operating system dependant

PAS	Parallel Application Services
PCS	PRIMECLUSTER Configuration Services
RCI	Remote Cabinet Interface
RMS	Reliant Monitor Services
RTP	Reliant Telco Product
SA	Shutdown Agent
SAN	Storage Area Network
SD	Shutdown Daemon
SF	Shutdown Facility
SIS	Scalable Internet Services
VIP	Virtual Interface Provider

Figures

Figure 1:	CIP diagram	12
Figure 2:	CF over IP diagram	13
Figure 3:	Login pop-up	16
Figure 4:	Main Web-Based Admin View window after login	17
Figure 5:	Global Cluster Services window in Web-Based Admin View	18
Figure 6:	Initial connection pop-up	18
Figure 7:	CF is unconfigured and unloaded	19
Figure 8:	CF loaded but not configured	20
Figure 9:	Scanning for clusters	21
Figure 10:	Creating or joining a cluster	22
Figure 11:	Selecting cluster nodes and the cluster name	23
Figure 12:	CF loads and pings	24
Figure 13:	Edit CF node names	25
Figure 14:	CF topology and connection table	26
Figure 15:	CF over IP window	28
Figure 16:	CIP Wizard window	29
Figure 17:	CIP Wizard window (IPv6)	30
Figure 18:	CIM configuration window	33
Figure 19:	Summary window	35
Figure 20:	Configuration processing window	36
Figure 21:	Configuration completion pop-up	36
Figure 22:	Configuration window after completion	37
Figure 23:	Main CF window	38
Figure 24:	Cluster resource diagram	56

Figures

Figure 25:	Adding a new node	68
Figure 26:	Main window	76
Figure 27:	Cluster Admin start-up window	77
Figure 28:	Initial connection choice window	78
Figure 29:	Cluster Admin main window	79
Figure 30:	CF route DOWN	81
Figure 31:	CF interface missing	82
Figure 32:	CF route table	83
Figure 33:	CF node information	84
Figure 34:	CF topology table	85
Figure 35:	Response Time monitor	86
Figure 36:	Starting CF	88
Figure 37:	CF configured but not loaded	89
Figure 38:	Start CF services pop-up	90
Figure 39:	Start CF services status window	91
Figure 40:	Stop CF	92
Figure 41:	Stopping CF	93
Figure 42:	PRIMECLUSTER log viewer	95
Figure 43:	Search based on date/time	96
Figure 44:	Search based on keyword	97
Figure 45:	Search based on severity	98
Figure 46:	ICF statistics	100
Figure 47:	MAC statistics	101
Figure 48:	Selecting a node for node to node statistics	102
Figure 49:	Node to Node statistics	103
Figure 50:	Selecting the Heartbeat monitor	104

Figure 51:	Heartbeat monitor	104
Figure 52:	CIM options	106
Figure 53:	Add to CIM	107
Figure 54:	Unconfigure CF	108
Figure 55:	CIM Override	109
Figure 56:	CIM Override confirmation	109
Figure 57:	Remove CIM Override	109
Figure 58:	Three-node cluster with working connections	112
Figure 59:	Three-node cluster where connection is lost	112
Figure 60:	Node C placed in the kernel debugger too long	115
Figure 61:	Four-node cluster with cluster partition	116
Figure 62:	A three-node cluster with three full interconnects	123
Figure 63:	Broken Ethernet connection for hme1 on fuji2	124
Figure 64:	Cluster with no full interconnects	125
Figure 65:	Conceptual view of CF interconnects	149
Figure 66:	CF with Ethernet interconnects	150
Figure 67:	CF with IP interconnects	150

Tables

Table 1: Kernel parameter values	54
Table 2: Local states	80
Table 3: Remote states	81
Table 4: PRIMECLUSTER log viewer severity levels	98
Table 5: Basic layout for the CF topology table	121
Table 6: Topology table with check boxes shown	122
Table 7: Topology table for 3 full interconnects	124
Table 8: Topology table with broken Ethernet connection	125
Table 9: Topology table with no full interconnects	126
Table 10: Available SAs	130
Table 11: Release information	179

Index

- /etc/cip.cf 57
- /etc/hosts
 - CIP configuration 10
 - CIP Wizard 32
- /etc/system 53
- /usr/sbin/shutdown 72, 73
- /var/opt/SMAWsf/log/SA_xscfsnmp0r.
 - log 136
- /var/opt/SMAWsf/log/SA_xscfsnmp1r.
 - log 136
- /var/opt/SMAWsf/log/SA_xscfsnmpg0
 - p.log 136
- /var/opt/SMAWsf/log/SA_xscfsnmpg0
 - r.log 136
- /var/opt/SMAWsf/log/SA_xscfsnmpg1
 - p.log 136
- /var/opt/SMAWsf/log/SA_xscfsnmpg1
 - r.log 136
- A**
 - adding
 - new node 57
 - nodes 23
 - to CIM 107
 - ALOM
 - configuration 135, 137
 - SA_sunF 135, 136, 137
 - Shutdown Agent 135, 137
 - automatic resource registration 63
- B**
 - backing up
 - configuration 41
 - Resource Database 69
 - broadcast messages 12
 - broken interconnects 111
- C**
 - CCBR
 - See* Cluster Configuration Backup and Restore
 - CCBR commands
 - cfbackup 171
 - cfrestore 171
 - CCBRHOME directory 44
 - CF
 - See also* Cluster Foundation
 - CF commands
 - cfconfig 171
 - cfregd 171
 - cfset 171
 - cftool 171
 - rcqconfig 171
 - rcquery 171
 - CF driver 19
 - CF over IP 11, 149
 - broadcast mask 149
 - CF Wizard 151
 - cftool -d 152
 - configure 151
 - devices 152
 - mixed configurations 151
 - scenarios 12
 - unique IP address 149
 - CF Registry
 - cfregd 47
 - user-level daemon 47
 - CF Remote Services 34
 - CF Wizard
 - bringing up 20
 - CF driver 120
 - CF over IP 151
 - displaying interconnects 28
 - edit node names 25
 - error message 37
 - new cluster 22
 - new node on existing cluster 120
 - running 39
 - scanning for clusters 21
 - summary window 35

- CF/CIP Wizard, starting 11
- cfbackup 41, 171
- cfconfig 171
- cfconfig -l 119
- CFCP 14
- cfcp 15, 34
- CFReg 50
- cfrestore 41, 171
- CFS commands
 - cfsmtd 172
 - fsck_rcfs 172
 - mount_rcfs 172
 - ngadmin 172
 - rcfs_fumount 172
 - rcfs_list 172
 - rcfs_switch 172
- cfset 13, 171
 - CFCP 14
 - CFSH 14
 - CLUSTER_TIMEOUT 14
 - maximum entries 14
 - options 14
 - tune timeout 15
- CFSH 14
- cfsh 15, 34
- cfsmtd 172
- cftool 171
- cftool -d 152
- cftool -n 111
- CIM
 - See* Cluster Integrity Monitor
- CIP
 - See* Cluster Interconnect Protocol
- CIP commands
 - cip.cf 172
 - cipconfig 172
 - ciptool 172
- CIP Wizard
 - /etc/hosts 32
 - CIP interface 30
 - CIP names 32
 - Cluster Admin 10
 - configuration file 32
 - numbering 31
 - screen 29
 - starting 11
 - cip.cf 39, 172
 - cipconfig 172
 - ciptool 172
 - clautoconfig 59, 174
 - clbackuprdb 69, 174
 - cldeidevice 174
 - cldevparam 173
 - clexec 174
 - clgettree 58, 64, 71, 72, 175
 - output 58
 - verify configuration 59
 - clinittest 58, 59, 72, 174
 - clmtest 173
 - clrcumonctl 173
 - clrcusetup 173
 - clrcimonctl 173
 - clrestorerdb 73, 174
 - clroot 17
 - clsetparam 66, 174
 - clsetup 58, 59, 70, 71, 72, 174
 - clstartsrc 175
 - clstopsrc 175
 - clsynfile 175
 - cluster
 - additional node 54
 - avoiding single point of failure 9
 - CF states 80
 - CIP traffic 9
 - data file 47
 - interfaces 8
 - name 7
 - node in consistent state 48
 - number of interconnects 9
 - partition 115
- Cluster Admin 76, 77
 - administration 75
 - login window 18
 - main CF table 82
 - routes 81
 - starting 18, 76
 - starting CF 87
 - stopping CF 87

- Cluster Configuration Backup and Restore 41
 - ccbr.conf 43
 - CCBRHOME directory 44
 - cfbackup 41
 - cfrestore 41
 - configuration file 43, 45
 - OS files 45
 - root files 45
- Cluster Foundation
 - administration 75
 - configuration 7
 - connection table 26
 - dependency scripts 93
 - device driver 158
 - devices 121
 - driver load time 119
 - Heartbeat monitor 104
 - interface 8
 - IP interfaces 8
 - loading driver 19
 - log viewer 94
 - main table 79
 - node information 84
 - node name 8, 57
 - quorum set 34
 - remote services 34
 - Response Time monitor 85
 - route tracking 81
 - security 15
 - topology table 26, 85, 119
 - unconfigure 108
- Cluster Integrity Monitor 48
 - adding a node 105
 - CF quorum set 34
 - cfcp 34
 - cfsh 34
 - configuration window 34
 - node state 48
 - options 106
 - override 109
 - override confirmation 109
 - quorum state 49
 - rcqconfig 49
- Cluster Interconnect Protocol
 - /etc/cip.cf 57
 - /etc/hosts 10
 - CF Wizard 57
 - configuration 9
 - configuration error 71
 - configuration file 39
 - configuration reset 72
 - configuration verification 71
 - defining 9
 - file format 39
 - interfaces 9
 - name 57, 58
 - ping command 58
 - properly configured 57
 - subnetwork 58
 - syntax 40
- CLUSTER_TIMEOUT 14
- cluster_uninstall 173
- collecting troubleshooting information 167
- COMINGUP state 93
- commands
 - CCBR 171
 - CF 171
 - CFS 172
 - CIP 172
 - cluster_uninstall 173
 - CPAT 173
 - MA 173
 - PAS 173
 - Resource Database 174
- config.us 176
- configuration
 - changing 50
 - hardware 69
 - restore 67
 - verify 72
 - See also* configuring 72
- configuring
 - CF 10
 - CF driver 20

- CF over IP 151
- CIM 48
- CIP 9, 10, 30, 39
- CIP with CF Wizard 57
- NPS 139, 141
- RCI 131
- resource database 57
- XSCF 134
 - See also* configuration 72
- connection table 27
- contents, manual 1
- corrupt data 113
- CPAT command 173
- creating
 - cluster, example 16
 - new cluster 22
- D**
- data, corrupt 113
- debugging 148
- default values, Solaris kernel 53
- defining virtual CIP interfaces 9
- devices
 - displayed 119
 - Ethernet 122
 - unconnected 26
- diagnostics 153
- disabling split-brain handling 144
- display statistics 99
- displayed devices 119
- documentation, related 2
- DOWN routes 81
- DOWN state 93, 112, 113
- E**
- editing
 - CF node names 25
 - cip.cf file 39
 - cluster.config file 13
- errors, CIP configuration 71
- Ethernet
 - CF drivers 37
 - CF networking 8
 - CF over IP 150
 - devices 122
 - example 124
 - Gigabit 8, 152
 - topology table 119
- eXtended System Control
 - Facility 130
 - different types 134
 - log files 134
 - SA_rccu 134
 - SA_rccux 134
 - SA_xscfp 134
 - SA_xscfr 134
- F**
- fjsnap 167
- fjsvvubs 177
- fjsvvcnf 177
- fsck_rcfs 172
- full interconnect 26, 121
- G**
- Gigabit Ethernet 152
- Global Disk Services 187
- Global File Services 187
- Global Link Services 188
- GUI
 - See* Cluster Admin
- H**
- Heartbeat monitor 104
- hvassert 175
- hvcn 175
- hvconfig 175
- hvdisp 175
- hvdump 175
- hvenv.local 176
- hvgdstartup 176
- hvlogclean 175
- hvlogcontrol 175
- hvsetenv 175
- hvshut 175
- hvswitch 176
- hvutil 176

I

ICF statistics 100
ILOM
 configuration 137
 SA_ilomp 138
 SA_ilomr 138
 Shutdown Agent 138
init command 111
Initial Connection Choice window 78
interconnects
 CF 8
 CF over IP 149
 Ethernet 122
 full 26
 IP 29
 IP subnetwork 150
 number of 9
 partial 26
 topology table 121
interfaces 8
 CIP 11
 missing 82
 network 82
Internet Protocol address
 CIP interface 30
INVALID state 93
IP interfaces 8
IP name, CIP interface 31
IP over CF 11
IP subnetwork 150

J

join problems 159
joining a running cluster 65

K

kernel parameters 53
keyword, search based on 97

L

Largest Sub-cluster Survival 145
LEFTCLUSTER 189
LEFTCLUSTER state 111, 114, 116,
 187, 189

cluster partition 115
description 112
displaying 111
in kernel debugger too long 114
lost communications 113
node state 191
panic/hung node 114
purpose 113
recovering from 114
shutdown agent 113
troubleshooting 165

LOADED state 89

loading

 CF driver 19
 CF driver with CF Wizard 24
 CF duration 24

local file systems, mount 73

local states 80

login

 password 16
 window 18

low latency 8

M

MA

See Monitoring Agents

MA commands

 cldevparam 173
 clrcumonctl 173
 clrcusetup 173
 clrcimonctl 173

MAC statistics 101

main CF table 79, 82

manual contents 1

manual pages

 display 171
 listing 171

marking down nodes 93

mipcstat 173

mirror virtual disks 190

Monitoring Agents 128

 delay detecting hardware
 failures 135

 RCI daemon 133

Index

mount_rcfs 172
mountall 73
Multi-path automatic generation 60

N

names

- CCBR 43
- CCBRHOME directory 44
- CF 84
- CF cluster 121
- cfname 10, 70
- CIP 71
- cluster 7, 23, 84
- configuration file 7
- connections table 27
- IP 31
- plug-ins 42
- tupple entries 14
- user 16
- Web-Based Admin View 8
- with asterisk 120

network

- interfaces 82
- outages 8

Network Power Switch

- configuration 139, 141
- setup 139, 141

ngadmin 172

Node to Node statistics 103

nodes

- adding 23
- adding a new 67
- details 84
- in kernel debugger 111
- joining a running cluster 65
- marking down 93
- panicked 111
- shut down 93

O

operating system files 45

P

panicked nodes 111

panics 135
partial interconnects 26, 121

PAS commands

- clmtest 173
- mipcstat 173

passwords 16

PCS commands

- pcs_reinstall 174
- pcscui 174
- pcstool 174

pcs_reinstall 174

pcscui 174

pcstool 174

plumb-up state 63

privileged user ID 16

pseudo device driver 198

Q

quorum

- CF 34
- CIM override 109
- reconfiguring 50
- state 49

R

RAID 197

rc scripts 119

RC_sf 148

rc2.d directory 148

rcfs_fumount 172

rcfs_list 172

rcfs_switch 172

rcqconfig 48, 49

RC-script 148

rcsd 176

rcsd log 148

rcsd.cfg 176

rdb.tar.Z 69, 70

reboot command 111

rebooting

- after cfrestore command 42
- clusterwide 47
- reboot command 111
- shut down CF 111

- reconfiguring Resource Database 70
- registering hardware 59
- Remote Cabinet Interface 130
 - configuration 131
 - hardware setup 131
 - log file 132
 - setup 131
- remote states 80
- resets 136
- Resource Database 57
 - adding new node 67
 - backing up 69
 - clgettree 58
 - clsetup 70
 - configure on new node 71
 - initializing 67
 - kernel parameters 53
 - new node 67
 - plumb-up state 63
 - reconfiguring 67, 70
 - registering hardware 59, 63
 - restoring 72, 73
 - start up synchronization 65
 - StartingWaitTime 65
- Resource Database commands
 - clautoconfig 174
 - clbackuprdb 174
 - cldeldevice 174
 - clexec 174
 - clgettree 175
 - clinitreset 174
 - clrestorerdb 174
 - clsetparam 174
 - clsetup 174
 - clstartpsc 175
 - clstoppsc 175
 - clsynfile 175
- Response Time monitor 85
- restoring
 - cluster configuration 171
 - Resource Database 72, 73
- RFC 1918 10
- RMS commands
 - config.us 176
 - hvassert 175
 - hvcm 175
 - hvconfig 175
 - hvdisk 175
 - hvdump 175
 - hvenv.local 176
 - hvgdstartup 176
 - hvlogclean 175
 - hvsetenv 175
 - hvshut 175
 - hvswitch 176
 - hvutil 176
- RMS Wizard Tools 146
- root 17
- root files 45
- route tracking 81
- S**
- SA
 - See* Shutdown Agents
 - SA specific log files 148
 - SA_ilomp 138
 - SA_ilomr 138
 - SA_pprci.cfg 176
 - SA_rccu
 - XSCF 134
 - SA_rccu.cfg 177
 - SA_rccux
 - XSCF 134
 - SA_sspint.cfg 177
 - SA_sunF 135, 136, 137
 - SA_sunF.cfg 177
 - SA_wtinps.cfg 177
 - SA_xscfp 134
 - SA_xscfr 134
- saving
 - cfbackup command 42, 171
 - PRIMECLUSTER
 - configuration 41
- SD
 - See* Shutdown Daemon
- sdtool 176
- sdtool command 128, 148
- search

Index

- keyword 97
 - severity levels 98
 - time filter 96
 - security
 - CF 15
 - selecting devices 122
 - setting up
 - RCI 131
 - SF
 - See* Shutdown Facility
 - SF commands
 - rcsd 176
 - sdtool 176
 - SF Wizard 7
 - starting 37
 - shutdown 73
 - Shutdown Agents 127
 - with LEFTCLUSTER 113
 - shutdown command 111
 - Shutdown Daemon 127
 - Shutdown Facility 7, 127
 - node weight 146
 - RMS Wizard Tools 146
 - split-brain handling 142
 - starting and stopping 148
 - starting automatically 148
 - starting manually 148
 - stopping automatically 148
 - stopping manually 148
 - weight assignment 144
 - ShutdownPriority attribute 143
 - simple virtual disks 197
 - single user mode 69
 - SMAWcf 42
 - special priority interfaces 9
 - Specific Application Survival 145
 - Specific Hardware Survival 145
 - split-brain
 - handling 144
 - LSS 145
 - SAS 145
 - SHS 145
 - start up synchronization 65
 - new node 67
 - StartingWaitTime 72
 - starting
 - CF 87, 88
 - CF Wizard 20
 - Cluster Admin 11
 - GUI 18
 - SF Wizard 37
 - Web-Based Admin View 16
 - StartingWaitTime 65, 67
 - default value 65
 - value 66
 - start-up window 76
 - states
 - COMINGUP 80, 93
 - DOWN 81, 112, 113
 - INVALID 80, 93
 - LEFTCLUSTER 81, 111, 114, 116
 - LOADED 80, 89
 - table of 112
 - UNCONFIGURED 80
 - UNKNOWN 80, 81
 - UNLOADED 80
 - UP 80, 81, 112
 - statistics, display CF 99
 - stopping
 - CF 87, 88
 - CF, third-party products 93
 - SF automatically 148
 - SF manually 148
 - valid CF states 93
 - subnet mask, CIP interface 31
 - synchronization phase 65
 - synchronization, start up 65
 - syslog window 95
 - system dump 167, 169
- ## T
- table of states 112
 - third-party product, shut down 93
 - time filter, search 96
 - timeout, tune 15
 - timestamp 66
 - top window 77

- topology table 119
 - basic layout 121
 - CF 26, 85
 - CF cluster name 121
 - CF driver 120
 - displayed devices 119
 - displaying 85
 - examples 123
 - flexibility 27
 - interconnects 121
 - Response Time monitor 85
 - selecting devices 122
- troubleshooting 153
 - beginning 153
 - collecting information 167
 - diagnostics 153
 - join-related problems 158
 - symptoms and solutions 157
- tunable parameters 13
- tune timeout 15
- tuple entries
 - name 14
 - value 14
- U**
 - unconfigure CF 108
 - unconnected devices 26, 121
 - UNKNOWN state 80
 - UNLOADED state 89
 - UP state 112
 - updating CFReg 50
 - user ID 16
 - user name 17
- V**
 - virtual disks
 - mirror 190
 - simple 197
- W**
 - Web-Based Admin View
 - known nodes 39
 - node list 19
 - starting 16
 - Web-Based Admin View commands
 - fjsvwvbs 177
 - fjsvwvcnf 177
 - wvCntl 177
 - wvGetparam 177
 - wvSetparam 177
 - wvstat 177
 - wvCntl 177
 - wvGetparam 177
 - wvroot 17
 - wvSetparam 177
 - wvstat 177
- X**
 - XSCF
 - configuration 134
 - See* eXtended System Control Facility
- Z**
 - zfs mount 73

