FUJITSU

PRIMECLUSTER

DR/PCI Hot Plug User's Guide

Oracle Solaris

J2S2-1596-01ENZ0(02)
January 2015

Preface

The PRIMECLUSTER DR/PCI Hot Plug User's Guide contains information on how to operate the SPARC M10 and SPARC Enterprise Dynamic Reconfiguration, and PCI Hot Plug functions on PRIMECLUSTER.

- Single-node system

  For a single-node system where cluster software is not used, but a redundant disk or network is used, the following operations are supported: (For further details, see "1.3 DR standard operation modes" and "1.4 PCI Hot Plug standard operation modes."

  - Hot swap of an I/O-connected system board
  - Addition, deletion, and move of an I/O-connected system board (only for SPARC M10)
  - Hot swap, hot expansion, and inter-partition move of a system board (*) on which CPU/memory boards are equipped
  - Hot swap and hot expansion of a PCI card

- Cluster system

  For a system that uses cluster software (PRIMECLUSTER), the following operations are supported:

  - For SPARC M10
    - Addition, deletion, replacement, and move of an I/O-connected system board
    - Addition, deletion, replacement, and move of a system board (*) on which CPU/memory boards are equipped
    - Hot swap and hot expansion of a PCI card
  - For SPARC Enterprise
    - Hot swap, hot expansion, and inter-partition move of a system board (*) on which CPU/memory boards are equipped
    - Hot swap and hot expansion of a PCI card

(*) You can operate hot swap, hot expansion, and inter-partition move of I/O-equipped system boards for SPARC Enterprise by disabling I/O. See "4.2.1 Prerequisite".

Target models

The document describes SPARC M10-4/M10-4S and SPARC Enterprise M4000/M5000/M8000/M9000.

Product models

The document covers the following products:

- PRIMECLUSTER GDS (hereafter GDS)
- PRIMECLUSTER GLS (hereafter GLS)
- PRIMECLUSTER

Target Readers

This manual is intended for all users who use PRIMECLUSTER and perform operations for SPARC M10 and SPARC Enterprise Dynamic Reconfiguration, and PCI Hot Plug functions.

Configuration of This Documentation

The contents of each part are described below.

Title	Description
Chapter 1 Overview	This chapter explains the overview of DR and PCI Hot Plug.
Chapter 2 GDS	This chapter explains how to use the DR/PCI Hot Plug functions using GDS.
Chapter 3 GLS	This chapter explains how to use the DR/PCI Hot Plug functions using GLS.

Title	Description
Chapter 4 PRIMECLUSTER	This chapter explains how to use the DR/PCI Hot Plug functions using PRIMECLUSTER.

Related Documentation

- Fujitsu M10/SPARC M10 Systems Domain Configuration Guide
- SPARC Enterprise M4000/M5000/M8000/M9000 Server Dynamic Reconfiguration (DR) User's Guide (Hereafter "DR User's Guide.")
- PRIMECLUSTER Concepts Guide
- PRIMECLUSTER Installation and Administration Guide
- PRIMECLUSTER Cluster Foundation (CF) Configuration and Administration Guide
- PRIMECLUSTER Reliant Monitor Services (RMS) with Wizard Tools Configuration and Administration Guide
- PRIMECLUSTER Reliant Monitor Services (RMS) Reference Guide
- PRIMECLUSTER Software Release Guide
- PRIMECLUSTER Installation Guide
- PRIMECLUSTER Web-Based Admin View Operation Guide
- PRIMECLUSTER Global Disk Services Configuration and Administration Guide
- PRIMECLUSTER Global File Services Configuration and Administration Guide
- PRIMECLUSTER Global Link Services Configuration and Administration Guide : Redundant Line Control Function
- PRIMECLUSTER Global Link Services Configuration and Administration Guide: Multipath Function
- PRIMECLUSTER Messages

Conventions

Notation

Prompts

Command line examples that require system administrator (or root) rights 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 ($).

Manual page section numbers

References to the UNIX operating system commands are followed by their manual page section numbers in parentheses - for example, cp(1)

The keyboard

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

Typefaces

The following typefaces highlight specific elements in this manual.

Typeface	Usage
Constant Width	Computer output and program listings; commands, file names, manual page names and other literal programming elements in the main body of text.
Italic, <*Italic*>	Variables that you must replace with an actual value.
<Constant Width>	Variables that you must replace with an actual displayed value.
Bold	Items in a command line that you must type exactly as shown.

Typeface	Usage
"Constant Width"	The title, documentation, screen, and etc of lookup destination.
[Constant Width]	Tool bar name, menu name, command name, button name, and icon names.

Example 1

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

```
root:x:0:1:0000-Admin(0000):/:
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):/:
```

Example 2

To use the cat(1) command to display the contents of a file, enter the following command line:

```
$ cat file name
```

Notation symbols

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

Point

Text that requires special attention.

Note

Information that users should be cautious of.

Example

Describes operation using an example.

Information

Information that users can refer to.

See

Manuals users find workable.

Abbreviations

- Oracle Solaris is abbreviated as Solaris, Solaris Operating System or Solaris OS.

Export Controls

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

Trademarks

UNIX is a registered trademark of the Open Group in the United States and other countries.

Date of publication and edition

February 2012, First edition
July 2014, 1.1 edition
January 2015, 1.2 edition

Copyright notice

Revision History

Revision	Location	Edition
Added the descriptions about DR (Dynamic Reconfiguration) function in SPARC M10.	All	1.1
Added the descriptions about PCI Hot Plug function in SPARC M10.	1.2 PCI Hot Plug 1.4 PCI Hot Plug standard operation modes 4.3 PCI Hot Plug operation in SPARC M10/SPARC Enterprise	1.2

Contents

Chapter 1 Overview

This chapter explains the overview of the DR and PCI Hot Plug.

1.1 DR

The document covers the following products:

- SPARC M10-4S and SPARC Enterprise M4000/M5000/M8000/M9000
- Solaris 10 and Solaris 11
- PRIMECLUSTER Enterprise Edition and HA Server 4.3A20 or later (SPARC M10-4S),
  PRIMECLUSTER Enterprise Edition and HA Server 4.3A10 or later
  (SPARC Enterprise M4000/M5000/M8000/M9000)

To use the DR function in the PRIMECLUSTER environment, perform the following operations.

- OS patch application

  For Solaris 10, apply the required patch: 118833-17 or later.

- Tuning for kernel parameters

  Delete the following line set for /etc/system during installation of PRIMECLUSTER and enable the kernel cage function.

```
set kernel_cage_enable=0
```

  To check if the setting is enabled, refer to "2.3.2 Settings of Kernel Cage Memory" of the "DR User's Guide."

Note

For details on conditions and notes to perform DR function, see "Fujitsu M10/SPARC M10 Systems Domain Configuration Guide" or "DR User's Guide."

1.2 PCI Hot Plug

The document covers the following products.

- SPARC M10-4/M10-4S and SPARC Enterprise M4000/M5000/M8000/M9000
- Solaris 10 and Solaris 11
- PRIMECLUSTER Enterprise Edition and HA Server 4.3A20 or later (SPARC M10-4/M10-4S),
  PRIMECLUSTER Enterprise Edition and HA Server 4.3A10 or later
  (SPARC Enterprise M4000/M5000/M8000/M9000)

For the operational environment of the PCI Hot Plug function, refer to "1.2 PCI Hot Plug Requirements" of the "PCI Hot Plug User's Guide."

1.3 DR standard operation modes

SPARC M10 or the SPARC Enterprise DR (Dynamic Reconfiguration) is used to add or remove a system board without disrupting ongoing operations in partitions. This section describes operation modes for DR.

1.3.1 SPARC M10 operation modes

- Addition of hardware resource

  Operation to add a system board to a physical partition without stopping the logical domains on the physical partition to deal with the business expansion of system load increase.
  It is necessary to set the CPU core activation as necessary.

- Reduction of hardware resource

  Operation to delete a system board from a physical partition without stopping the running operation to utilize the hardware resource effectively by the change of the operation volume.

- Move of hardware resource

  Operation to delete a system board in the other physical partition temporarily, and then add the once deleted system board to the physical partition where hardware resource needs to be added. This operation is to deal with the business expansion of system load increase. The hardware resource can be moved while logical domains on two physical partitions are operating. It is a flexible system construction for the system load fluctuation.

- Hot swap when parts fail

  Hot swap (*) of a system board for replacing a failed I/O device, memory, or CPU

  (*) When the system board deleted once is added again, the types of I/O controller and device, and the location of the implemented slot must be the same as before it was deleted.

Note

When adding or deleting I/O resources to/from the existing domain physically, you need to reboot the following domains where the I/O resources are changed: the control domain, I/O root domain, and I/O domain.

1.3.2 SPARC Enterprise operation modes

- Failed component hot swap

  Hot swap (*) of a system board for replacing a failed I/O device, memory, or CPU

  (*) When the system board deleted once is added again, the types of I/O controller and device, and the location of the implemented slot must be the same as before it was deleted.

- Hardware resource hot expansion

  Addition of a system board for adding the CPU, memory, or I/O device

- Hardware resource inter-partition move and system board pool

  Inter-partition move of a system board. The system board pool function is used to reserve specific system boards that do not belong to any one of the partitions.

Note

You can even operate hot swap of I/O-equipped system boards by disabling I/O (See "4.2.1 Prerequisite").

1.4 PCI Hot Plug standard operation modes

The SPARC M10 or the SPARC Enterprise PCI Hot Plug is used to add or remove a PCI card in an Oracle Solaris environment without rebooting a system. The system resource on the PCI card is an I/O device. This PCI card can be added or removed with PCI Hot Plug. PCI Hot Plug can be used for the following operations:

- PCI card hot swap

  Replacing or removing a failed PCI card without stopping operations

- PCI card hot expansion

  Adding a new PCI card without stopping operations

1.5 Supported DR operations on PRIMECLUSTER

This section describes the range of the supported DR operations on each product of PRIMECLUSTER.

1.5.1 Range of SPARC M10 support

The range of support when using DR in PRIMECLUSTER is as follows.

Product	Addition	Deletion	Replacement	Move
GDS	Available (*1)	Available (*2)	Available (*3)	Available (*2)
GLS	Available (*4)	Available (*5)	Available (*6)	Available (*5)
GFS(*7)	Not available	Not available	Not available	Not available
PRIMECLUSTER(*8)	Available	Available	Available	Available

(*1) The procedure when the new disk, connected to a system board to be added, is managed by GDS, see "Chapter 2 GDS."

(*2) For disks that are connected to a system board to be removed or moved, change the GDS configuration in advance so that GDS is not used, and then perform the DR operation. The procedure to change the configuration is the same as when DR is not used.

(*3) For a system board that is connected to a GDS mirror disk, you must confirm the prerequisites for GDS operations, and also perform GDS specific operations. For necessary conditions and the procedure, see "Chapter 2 GDS."

(*4) The procedure when using NIC that was added by the DR operation in GLS is the same as when DR is not used.

(*5) For NIC that is connected to a system board to be removed or moved, change the GLS configuration in advance so that GLS is not used, and then perform the DR operation. The procedure to change the configuration is the same as when DR is not used.

(*6) For a system board where a redundant NIC with GLS is used, you must confirm the prerequisites for GLS operations, and also perform GLS specific operations. For necessary conditions and the procedure, see "Chapter 3 GLS."

(*7) DR cannot be operated in an environment where the GFS Shared File System is used.

(*8) When DR is operated while PRIMECLUSTER is in operation, there are notes and the procedure. For necessary conditions and the procedure, see "Chapter 4 PRIMECLUSTER."

1.5.2 Range of SPARC Enterprise support

The range of support varies depending on whether I/O is equipped on the system board when using DR in PRIMECLUSTER.

Product	Hot swap (*1)		Hot expansion		Transfer/SB pool	
	Without I/O	With I/O	Without I/O	With I/O	Without I/O	With I/O
GDS	Available	Available (*2)	Available	Not available	Available	Not available
GLS	Available	Available (*3)	Available	Not available	Available	Not available
GFS	Available	Not available	Available	Not available	Available	Not available
PRIMECLUSTER	Available (*4)(*5)	Not available	Available (*4)	Not available	Available (*4)(*5)	Not available

(*1) If the system board uses Extended System Boards (XSB) and configures multiple partitions, it is necessary to remove all the Extended System Boards from the partitions.

(*2) For a system board that is connected to a GDS mirror disk, you must confirm the prerequisites for GDS operations. For details, see "Chapter 2 GDS."

(*3) For a system board where a redundant Network Interface Card (NIC) with GLS is used, you must confirm the prerequisites for GLS operations. For details, see "Chapter 3 GLS."

(*4) Observe the warnings and cautions in "Chapter 4 PRIMECLUSTER."

(*5) The kernel memory board cannot be removed (replaced or moved) on any of the node models.

1.6 Supported PCI Hot Plug operations on PRIMECLUSTER

PRIMECLUSTER supports PCI Hot Plug based replacement and addition of a PCI card that is connected to any of the following devices.

		Node type	
		Single node system	Nodes in a cluster system
Devices	RAID device	Available	Available
	Public LAN	Available	Available (*1)(*2)
	Interconnect	-	Available (*1)(*2)(*3)

(*1) CF configuration must be complete.

(*2) NIC cannot be replaced with PCI Hot Plug if Scalable Internet Services (SIS) is used (SMAWdtcp is installed).

(*3) Only hot swap will be supported.

Chapter 2 GDS

This chapter explains how to use the function of the DR/PCI Hot Plug in GDS.

GDS allows you to operate DR-based hot swap and hot expansion of a system board by configuring mirrored disks that are connected to different system boards without disrupting ongoing operations. Similarly, GDS allows you to operate PCI Hot Plug based hot swap and hot expansion of a PCI card without disrupting ongoing operations.

This section describes the prerequisites and procedures for the following operations:

- Hot swap of a system board using DR (only root or local class for SPARC Enterprise)
- Addition of a system board by using DR (SPARC M10)
- Hot swap of a PCI card using PCI Hot Plug
- Hot expansion of a PCI card using PCI Hot Plug

2.1 Prerequisites

To enable hot swap and hot expansion with DR/PCI Hot Plug on a GDS disk, the disk must satisfy the following conditions.

2.1.1 Prerequisites for DR and PCI Hot Plug

- All parts excluding the failed portion must be in the same configuration as that used before the removal.

  To add the system board that is removed with DR. The GDS disk and the PCI card must be on the same mounted slot locations as those used before the removal.

  Similarly, not the GDS disk, but the PCI card must be replaced with PCI Hot Plug. For corrective action against disk failures, see the "PRIMECLUSTER Global Disk Services Configuration and Administration Guide"

- The mirrored volumes must be synchronized.

  Hot swap with DR/PCI Hot Plug is not allowed in the following conditions:

  a. If there's a volume that is not synchronized in the mirror disk group. The disk that is connected to a system board or PCI card belongs to this mirror disk group, and

  b. If there's an ACTIVE slice on the disk in the 1) volume.

  You can presume the following causes if the volumes are not synchronized. Check the slice state before initiating hot swap using the "sdxinfo(1)" command.

  - If you detach the source disk during synchronization copy.
  - If there's only one ACTIVE slice during TEMP slice detachment.
  - When you will sequentially remove the system board (or PCI card) to which each mirrored disk is connected using DR (or PCI Hot Plug).

- The disks must be mirrored between the system boards (or PCI cards).

  If the mirrored disks are connected to the same system board, you are not allowed to remove the system board using DR.

  If the mirrored disks are connected to the same PCI card, you are not allowed to operate hot swap of the PCI card using PCI Hot Plug.

- The disk should be non-GDS, and the I/O path to the disk must not be redundant.

  If the disk is controlled in a redundant I/O configuration with the Multipath Disk Control, it is necessary to follow the DR/PCI Hot Plug procedure for the Multipath Disk Control. Also, if you are using non-Fujitsu redundant I/O software (for example, EMC PowerPath), check how DR/PCI Hot Plug is supported for each software product, then follow the procedure.

- The class attribute (scope) must be set to access the shared class disk on all the nodes.

  You can check the class attribute (scope) on the Class Configuration screen of the GDS Management View.

2.1.2 Prerequisite specific to DR

- DR must be in a single partition environment

  Because PRIMECLUSTER supports Dynamic Reconfiguration (DR) only for the I/O-disabled CPU or memory, DR must be in a single partition environment.

2.2 Replacing the system board with DR

This section explains how to operate DR-based hot swap in an environment where GDS is used.

2.2.1 Detaching the disk

1. Prior check

   Check if the GDS mirror disk that is connected to the system board using the following steps. In the following example, the "sd" driver is used to control the disk.

   1. Check the name of the disk device that is connected to the system board. For details, see the following each document.

      For details, see "Fujitsu M10/SPARC M10 Systems Domain Configuration Guide" or "DR User's Guide."

   2. Check the disk is under the GDS control, then find the SDX disk name.

      ```
      # sdxinfo -D | grep c1t1d0
      disk   disk01  mirror  class01    grp01  c1t1d0   8493876  *        ENABLE
      ```

      If the disk is not under the GDS control, remove the system board using DR after checking whether the I/O path is redundant for non-GDS software, or if the disk is not used.

   3. Check if there are ACTIVE slices configuring a mirror volume on the mirrored disk other than the disk that is connected to the system board.

      ```
      # sdxinfo  -S  -o disk01
      OBJ    CLASS   GROUP   DISK    VOLUME  STATUS
      ------ ------- ------- ------- ------- --------
      slice  class01 grp01   disk01  v01     ACTIVE
      slice  class01 grp01   disk01  v02     ACTIVE
      slice  class01 grp01   disk02  v01     ACTIVE
      slice  class01 grp01   disk02  v02     ACTIVE
      ```

      If there's no ACTIVE slice, you cannot remove the system board using DR.

2. Before the system board is removed

   1. When the target disk is SPARC M10 internal disk (other than ZFS mirroring)

      1. Detaching the disk from the mirror disk group

         Detach the target disk from the mirror disk group.

         ```
         # sdxdisk -D  -c class01  -g grp01  -d disk01
         ```

      2. Removing the disk from the disk class

         Remove the target disk from the disk class

         ```
         # sdxdisk -R  -c class01  -d disk01
         ```

   2. When the target disk is not SPARC M10 internal disk (other than ZFS mirroring)

      1. Detach the target disk with the sdxswap command.

         ```
         # sdxswap  -O  -c class01  -d disk01
         ```

2. Check if the target disk is correctly detached (changed to the SWAP state).

```
# sdxinfo  -D  -o disk01
 OBJ     NAME     TYPE    CLASS   GROUP   DEVNUM  DEVBLKS  DEVCONNECT STATUS
 ------ ------- ------ ------- ------- ------ -------- ---------- ------
 disk   disk01   mirror  class01  grp01  c1t1d0  8493876    *          SWAP
```

3. When the target disk is a system disk and it is being mirrored (ZFS mirroring)

    1. Check the status of the mirror configuration. Check if all STATEs are ONLINEs.

```
# zpool status rpool
       pool: rpool
      state: ONLINE
       scan: resilvered 17.8G in 0h3m with 0 errors on Wed Apr 16 13:06:34 2014
     config:

            NAME                                STATE     READ WRITE CKSUM
            rpool                               ONLINE       0     0     0
              mirror-0                          ONLINE       0     0     0
                /dev/sfdsk/System/dsk/Volume2   ONLINE       0     0     0
                /dev/sfdsk/System/dsk/Volume1   ONLINE       0     0     0

     errors: No known data errors
```

    2. Detach the GDS volume consisted by the target disk from the ZFS mirroring.

```
# zpool detach rpool /dev/sfdsk/System/dsk/Volume2
```

    3. Stop and remove the detached volume.

```
# sdxvolume -F -c System -v Volume2
# sdxvolume -R -c System -v Volume2
```

    4. Detach the target disk from the disk group.

```
# sdxdisk -D -c System -g Group2 -d Root2
```

    5. Remove the target disk from the disk class.

```
# sdxdisk -R -c System -d Root2
```

2.2.2 Replacing the system board with DR

Perform a hot swap of a system board using DR (Detachment and addition). For details, see "Fujitsu M10/SPARC M10 Systems Domain Configuration Guide" or "DR User's Guide."

2.2.3 Re-attaching the disk

1. Check the name of the disk device that is connected to the system board. For details, see "Fujitsu M10/SPARC M10 Systems Domain Configuration Guide" or "DR User's Guide."
2. When the target disk is the SPARC M10 internal disk (other than ZFS mirroring)

    1. Registering a disk in the disk class

    Register the target disk in the disk class again.

```
# sdxdisk -M  -c class01  -d c1t1d0=disk01
```

    You can check if the disk has been registered using the following command.

```
# sdxinfo  -D
 OBJ    NAME    TYPE   CLASS   GROUP   DEVNAM  DEVBLKS  DEVCONNECT       STATUS
```

```
 ------ ------- ------ ------- ------- ------- -------- ---------------- -------
 disk   disk01   undef  class01  *     c1t1d0  8493876   *               ENABLE
 disk   disk02   mirror class01  grp01 c1t2d0  8493876   *               ENABLE
 #
```

2. Adding a disk to the mirror disk croup

   Add the target disk to the mirror disk group again.

```
# sdxdisk -C  -c class01  -g grp01  -d disk01
```

   Check if the disk has been registered using the following command.

```
# sdxinfo  -D
 OBJ    NAME    TYPE   CLASS   GROUP   DEVNAM  DEVBLKS  DEVCONNECT       STATUS
 ------ ------- ------ ------- ------- ------- -------- ---------------- -------
 disk   disk01   mirror class01  grp01 c1t1d0  8493876   *               ENABLE
 disk   disk02   mirror class01  grp01 c1t2d0  8493876   *               ENABLE
 #
```

3. When the target disk is not the SPARC M10 internal disk (other than ZFS mirroring)

   1. Check the disk class and the SDX disk name after confirming that the disk is under the GDS control and in the SWAP state.

```
# sdxinfo  -D | grep c1t1d0
disk   disk01  mirror class01    grp01 c1t1d0   8493876 *           SWAP
```

   2. Attach the disk again using the "sdxswap" command.

```
# sdxswap  -I  -c class01  -d disk01
```

   3. Check if the disk has been attached (whether the disk has entered ENABLE).

```
# sdxinfo  -D  -o disk01
OBJ     NAME     TYPE    CLASS   GROUP   DEVNUM  DEVBLKS  DEVCONNECT STATUS
  ------ ------- ------ ------- ------- ------ -------- ---------- ------
disk    disk01   mirror class01  grp01  c1t1d0  8493876    *         ENABLE
```

4. When system disk was mirrored by using GDS (ZFS mirroring)

   1. Register the target disk in the disk class again.

```
# sdxdisk -M -c System -a type=root -d c0t1d0=Root2:keep
```

   2. Attach the target disk to the disk group again.

```
# sdxdisk -C -c System -g Group2 -d Root2 -v 0=Volume2:on
```

   3. Add the GDS volume to the ZFS mirroring configuration.

```
# zpool attach rpool /dev/sfdsk/System/dsk/Volume1 /dev/sfdsk/System/dsk/Volume2
```

   4. Check the status of the mirror configuration. Check if all STATEs are ONLINEs.

```
# zpool status rpool
  pool: rpool
 state: ONLINE
  scan: resilvered 17.8G in 0h3m with 0 errors on Wed Apr 16 14:05:30 2014
config:

        NAME                              STATE     READ WRITE CKSUM
        rpool                             ONLINE       0     0     0
          mirror-0                        ONLINE       0     0     0
            /dev/sfdsk/System/dsk/Volume2 ONLINE       0     0     0
            /dev/sfdsk/System/dsk/Volume1 ONLINE       0     0     0
```

```
   errors: No known data errors
```

5. Install the boot block to the GDS volume.

   For the following environments, do not perform this procedure:

   - Using Solaris 10 and the applying the kernel patch 144500-19 or later
   - Solaris 11 11/11 or later

```
# installboot -F zfs /usr/platform/`uname -i`/lib/fs/zfs/bootblk /dev/sfdsk/System/rdsk/
Volume2
```

2.3 Adding the system board with DR (SPARC M10)

This section describes a procedure when a new disk, connected to a system board added with DR in the SPARC M10 environment, is managed by GDS.

1. Adding the system board with DR

   Add the system board by using DR. For details, see "Fujitsu M10/SPARC M10 Systems Domain Configuration Guide."

2. Setting GDS

   1. For a shared class

      See "9.3.1 Adding a Shared Disk Device" in "PRIMECLUSTER Installation and Administration Guide 4.3."

   2. For a root class or a local class

      Select [Update Now] in [View] in the main menu of GDS Management View, then the status will be updated.

      After that you can set GDS with a normal procedure. For details, see "PRIMECLUSTER Global Disk Services Configuration and Administration Guide."

2.4 Replacing the PCI card with PCI Hot Plug

This section explains how to replace a PCI card using PCI Hot Plug.

2.4.1 Detaching the disk

1. Stop the hardware monitoring.

   Stop the hardware monitoring function of the Machine Administration. For details, see the manual of the Machine Administration.

2. Specify the PCI card.

   Specify the PCI card, the path interface name of the I/O device, and the physical disk. For details, see the "Oracle Solaris Administration: Devices and File Systems."

3. Search the SDX object.

   Search the SDX object for the disk that is connected to the PCI card specified at step 2).

```
# sdxinfo -D
OBJ    NAME    TYPE   CLASS   GROUP   DEVNAM  DEVBLKS  DEVCONNECT       STATUS
------ ------- ------ ------- ------- ------- -------- ---------------- -------
disk   disk01  mirror class01 grp01  c2t10d0 17682084 sdxst3:sdxst4    ENABLE
disk   disk02  mirror class01 grp01  c2t11d0 17596416 sdxst3:sdxst4    ENABLE
#
```

4. Detach the disk.

   Detach the disk from the mirror disk group.

```
# sdxdisk -D  -c class01  -g grp01  -d disk02
```

5. Delete the disk from the disk class.

   Delete the disk from the disk class.

```
# sdxdisk -R  -c class01  -d disk02
```

2.4.2 Removing, replacing and adding the PCI card

For details, see the "Oracle Solaris Administration: Devices and File Systems."

2.4.3 Adding the disk again

Edit the resource information

1. If the system is in a cluster environment, edit the resource information using the following procedure.

   If it is not in a cluster environment, then skip step 1) and start from step 2).

   1. Search the node ID.

```
# /etc/opt/FJSVlcluster/bin/clgetnode
RID 3
KEY sdxst3
RNAME sdxst3
NODEID 0
#
```

   2. Search the Resource ID.

```
# /etc/opt/FJSVcluster/bin/clgetrid  -c DISK  -k c2t11d0  -s 0
87
#
```

   3. Delete the resource information.

```
# /etc/opt/FJSVcluster/bin/cldelrsc -r 87
```

   4. Execute the automatic resource registration, and then the following command on all the nodes:

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

2. Register the disk

   Register the disk in the disk class again.

```
# sdxdisk -M  -c class01  -d c1t1d0=disk02
```

   You can check if the disk has been registered using the following command:

```
# sdxinfo -D
OBJ    NAME    TYPE   CLASS   GROUP   DEVNAM  DEVBLKS  DEVCONNECT       STATUS
------ ------- ------ ------- ------- ------- -------- ---------------- -------
disk   disk01   mirror class01  grp01 c2t10d0 17682084 sdxst3:sdxst4    ENABLE
disk   disk02   undef  class01  *      c2t11d0  17596416 sdxst3:sdxst4    ENABLE
#
```

3. Add the disk

   Add the disk to the mirror disk group.

```
# sdxdisk -C  -c class01  -g grp01  -d disk01
```

   Check if the disk has been registered using the following command:

```
# sdxinfo -D
OBJ    NAME    TYPE   CLASS   GROUP   DEVNAM  DEVBLKS  DEVCONNECT       STATUS
```

```
------ ------- ------ ------- ------- ------- -------- ---------------- -------
disk   disk01   mirror class01  grp01 c2t10d0 17682084 sdxst3:sdxst4    ENABLE
disk   disk02   mirror class01  grp01 c2t11d0 17596416 sdxst3:sdxst4    ENABLE
#
```

2.5 Adding the PCI card with PCI Hot Plug

1. Add a PCI card.

   For information on how to add a PCI card using PCI Hot Plug, see the "Oracle Solaris Administration: Devices and File Systems."

2. Execute the automatic resource registration

   If the system is in a cluster, register the disk resource using the "clautoconfig(1M)" with the "-r" and "-d" options, or register the disk resource by using the CRM main view of ClusterAdmin (GUI).

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

3. Check the additional LU

   Check if the additional LU has been registered in the resource using the "clgettree(1) command or the CRM main view of Cluster Admin (GUI).

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

   If you check the additional LU using the GDS management view, it might not be displayed immediately. Display the latest information choosing "Update information" from the main menu "Display".

4. Create the GDS volume.

   You can create a GDS volume using the normal procedure. For details, see the "PRIMECLUSTER Global Disk Services Configuration and Administration Guide."

Chapter 3 GLS

This chapter explains how to use the function of the DR/PCI Hot Plug in GLS.

You can replace, remove or add the system board with DR by configuring a redundant LAN between the NIC (Network Interface Cards) that are connected to different system boards using Global Link Services (GLS). Similarly, you can replace or add the NIC using PCI Hot Plug without disrupting ongoing operations.

3.1 Prerequisites specific to DR

To enable DR-based hot swap of a system board, the GLS-configured NIC must satisfy the following conditions.

- NIC must be redundant between the system boards

  If the redundant NIC is connected to the same system board, you are not allowed to remove the system board using DR.

- All parts excluding the failed portion must be in the same configuration as that used before the removal.

  If you remove the system board that has previously been replaced, the NIC slot location must be the same as that used before the removal.

3.2 Replacement of the system board using the DR

This section explains the replacement of the system board using the DR.

3.2.1 When using the DR function of ESF

See "4.5.1 Executing DR command of ESF" of "PRIMECLUSTER Global Link Services Configuration and Administration Guide 4.3 (Redundant Line Control Function)".

When executing the DR command of ESF, the automatic disconnection/embedding is executed for a virtual interface of GLS by DR associated script in a redundant line control function.

Moreover, the DR associated script does not support the disconnecting/embedding of an NIC under a redundant virtual interface comprised of an IPv6 virtual interface and a tagged VLAN Interface.

3.2.2 When using the DR function of XSCF

See the following manual for the hot swap of a system board by using DR in XSCF

- Redundant Line Control Function (NIC switching mode, Fast switching mode, and GS/SURE Linkage mode)

  "4.5.2.1 Replacing the system board using the DR of XSCF (SPARC M10)" or "4.5.2.2 Replacing the system board using the DR of XSCF (SPARC Enterprise M Series)" in "PRIMECLUSTER Global Link Services Configuration and Administration Guide 4.3 (Redundant Line Control Function)"

- Redundant Line Control Function (Virtual NIC Mode)

  "6.1 When using the DR function" in "PRIMECLUSTER Global Link Services Configuration and Administration Guide 4.3"

See the following manuals when disconnecting the system board from non-redundant cluster configuration

- Redundant Line Control Function (NIC switching mode, Fast switching mode, and GS/SURE Linkage mode)

  "2) Disconnecting the system board from non-redundant cluster configuration" of "4.5.2.1 Replacing the system board using the DR of XSCF (SPARC M10)" or "2) Disconnecting the system board from non-redundant cluster configuration" of "4.5.2.2 Replacing the system board using the DR of XSCF (SPARC Enterprise M Series)" in "PRIMECLUSTER Global Link Services Configuration and Administration Guide 4.3"

- Redundant Line Control Function (Virtual NIC Mode)

  "1. Reconnecting a NIC to the virtual NIC" of "6.1.2 Replacing the system board" in "PRIMECLUSTER Global Link Services Configuration and Administration Guide 4.3

See the following manuals when embedding NIC to non-redundant cluster configuration

- Redundant Line Control Function (NIC switching mode, Fast switching mode, and GS/SURE Linkage mode)

  "4) Embedding NIC to non-redundant configuration" of "4.5.2.1 Replacing the system board using the DR of XSCF (SPARC M10)" or "4) Embedding NIC to non-redundant configuration" of "4.5.2.2 Replacing the system board using the DR of XSCF (SPARC Enterprise M Series)" in "PRIMECLUSTER Global Link Services Configuration and Administration Guide 4.3"

- Redundant Line Control Function (Virtual NIC Mode)

  "2. Reconnecting a NIC to the virtual NIC" in "6.1.2 Replacing the system board" in "PRIMECLUSTER Global Link Services Configuration and Administration Guide 4.3

3.3 NIC hot swap and hot expansion with PCI Hot Plug

See the following for the replacement and expansion NIC by PCI Hot Plug:

- Redundant Line Control function (NIC switching mode, fast switching mode, and GS/SURE switching mode)

  See "4.5.3 Replacement/Expansion PHP (PCI Hot Plug)" of "PRIMECLUSTER Global Link Services Configuration and Administration Guide 4.3 (Redundant Line Control Function)".

- Redundant Line Control function (Virtual NIC mode)

  See "6.2 When using the PHP function" of "PRIMECLUSTER Global Link Services Configuration and Administration Guide 4.3 (Redundant Line Control Function for Virtual NIC Mode)".

Chapter 4 PRIMECLUSTER

This chapter explains how to use the DR/PCI Hot Plug in PRIMECLUSTER.

4.1 DR operation in SPARC M10

You can operate DR-based hot swap or hot expansion in the following cases:

- When an insignificant failure (such as 1 bit intermittent error) that does not disrupt ongoing operations occurs in the CPU or memory.
- When the system board must be added, removed, or moved due to business expansion or reduction.

Note

- To replace, move, or remove the system board, it is necessary to verify whether the reduced CPU or memory does not affect the operations in advance.
- Estimate in advance how long the system will be suspended due to the DR operation.
  If you perform DR when the operations are being suspended, the system will be stopped. To avoid the system stop, the failover or failback of the operations must be done to make them continuous on the node where DR is not performed.
- When performing DR on an operation node, you need to change the timeout value of CF. If DR is not completed within this changed timeout value, the node state becomes LEFTCLUSTER. In this case, the node is not switched automatically.
- If you have a plan to use DR, make sure to perform the following DR process test in the stage of building cluster system to verify that DR does not cause any problems.

4.1.1 DR operation procedure

Performing DR will suspend the system. This might affect the PRIMECLUSTER node monitoring facility, and the node may be stopped forcibly. Stop the node monitoring facility of the cluster before performing DR.

In Oracle VM Server for SPARC environment, perform the operation before and after DR on all the guest domains and control domains on which PRIMECLUSTER is built.

Note

- If you perform DR when the operations are being suspended, the system will be stopped. To avoid the system stop, the failover or failback of the operations must be done to make them continuous on the node where DR is not performed. Then, perform DR on the standby node.
- Estimate in advance how long the system will be suspended due to the DR operation.
- In the cluster system, operations such as the unmanned DR operation with time scheduling are not supported.

Take the following steps to perform DR.

1. Check the device name for the cluster interconnect.
   When removing or replacing the device for the cluster interconnect, check that multiple devices that are not to be removed or replaced exist, and the state of these devices are "UP."

```
# cftool -d
Number Device    Type Speed   Mtu     State  Configured  Address
1      /dev/igb1  4    100    1432     UP     YES        00.00.0e.25.1a.38
2      /dev/igb7  4    100    1432     UP     YES        00.00.0e.25.1a.38
```

2. When deleting or replacing the device for the cluster interconnect, remove the device from the cluster interconnect.

```
# cfrecon -d <device name>
```

3. When performing hot swap in the configuration where GDS is used, see "2.2.1 Detaching the disk" to remove a disk from GDS class.

4. When performing hot swap in the configuration where GLS is used, see "3.2 Replacement of the system board using the DR" to disconnect an NIC from the multiplex configuration.

5. When deleting, replacing, or moving the system board, take the prior steps to remove the system board from the physical partition. When the domain should be restarted as the system board is deleted, replaced, or moved, restart the domain before the monitoring facility of PRIMECLUSTER is shut down or changed as follows. For details on deleting, replacing, or moving steps of the system board, see "SPARC M10 System Domain Configuration Guide."

6. Check the PRIMECLUSTER configuration file name by executing the "hvdisp -n" command on any of the nodes where RMS is running. In the example below, the RMS configuration file name is "config.us."

```
# hvdisp -n
/opt/SMAW/SMAWRrms/build/config.us
#
```

7. Stop PRIMECLUSTER RMS by executing the "hvshut" command on all the nodes. When you answer "yes", PRIMECLUSTER RMS will stop, however the applications defined on the cluster applications remain running.

```
# hvshut -L
                            WARNING
                            -------
The '-L' option of the hvshut command will shut down the RMS
software without bringing down any of the applications.
In this situation, it would be possible to bring up the same
application on another node in the cluster which *may* cause
data corruption.

Do you wish to proceed ? (yes = shut down RMS / no = leave RMS running).
yes

NOTICE: User has been warned of 'hvshut -L' and has elected to proceed.
```

Add the following line in the "/opt/SMAW/SMAWRrms/bin/hvenv.local" file on all the nodes.

```
export HV_RCSTART=0
```

The above procedure is necessary to prevent RMS start automatically right after the OS startup.

8. Stop PRIMECLUSTER SF by executing the "sdtool" command on all the nodes as follows.

```
# sdtool -e
(SMAWsf, 30, 11) : RCSD returned a successful exit code for this command
```

9. Change the timeout value of PRIMECLUSTER CF heartbeat monitoring. Perform the following operation on all the nodes:

   - Check the set timeout value. This value is used to restore the settings.

```
# cfset -g CLUSTER_TIMEOUT
>From cfset configuration in CF module:
Value for key: CLUSTER_TIMEOUT --->10
#
```

     When the following message is displayed, the timeout value is 10 seconds (default value).

```
# cfset -g CLUSTER_TIMEOUT
cfset: No matching key found in CF Module
#
```

   - Add the following setting to /etc/default/cluster.config.

```
CLUSTER_TIMEOUT "timeout"
```

timeout(second) = time while the system is suspended due to DR + DR operation time

- Execute the following command.

```
# cfset -r
```

- Check whether or not the CF timeout value is valid.

```
# cfset -g CLUSTER_TIMEOUT
>From cfset configuration in CF module:
Value for key: CLUSTER_TIMEOUT --->timeout
#
```

10. Add, remove, replace, or move the system board.
    To add the system board, take the prior steps to add the physical partition and to check the operation status of the logical domains.
    To remove the system board, detach the system board from the physical partition.
    To replace the system board, take the prior steps to detach the system board from the physical partition, to add the replaced system board to the physical partition, and then to check the operation status of the logical domains.
    To move the system board, take the steps to add and remove the system board.
    For details on adding, deleting, replacing, or moving steps of the system board, see "SPARC M10 System Domain Configuration Guide."

11. Return the CF heartbeat timeout to the default value on all the nodes as follows:

    - Change the CLUSTER_TIMEOUT in /etc/default/cluster.config to the timeout value that is checked in step 9.

      Before change:

```
CLUSTER_TIMEOUT "timeout"
```

      timeout(second) = timeout time that is set in step 9

      After change: (when the original timeout value is 10)

```
CLUSTER_TIMEOUT "10"
```

    - Execute the following command.

```
# cfset -r
```

    - Check whether the timeout value is changed correctly.

```
# cfset -g CLUSTER_TIMEOUT
>From cfset configuration in CF module:
Value for key: CLUSTER_TIMEOUT --->10
#
```

12. When the building block configuration is added from 1BB to multiple BB, see "5.1.2.1.3 Using the Shutdown Configuration Wizard" in "PRIMECLUSTER in Installation and Administration Guide "to reconfigure the takeover IP address of XSCF for the XSCF IP address registered to SF.

13. Start PRIMECLUSTER SF. Execute the sdtool command on all nodes as follows.

```
# sdtool -b
```

14. Check if PRIMECLUSTER SF is running.

    Select the [Tools]-[Shutdown Facility]-[Show Status] menu from the CF main window of Cluster Admin, then check the "Test State" field on each node.

15. Start PRIMECLUSTER RMS by executing the "hvcm" command as follows on all the nodes. Specify the RMS configuration file name, which is checked in step 1, for the "-c" option. For example, if the name is "config.us", "config" will be specified.

```
# hvcm -c config
Starting Reliant Monitor Services now
```

16. PRIMECLUSTER RMS must be running on all the nodes. Check if each icon indicating the node state is green (Online) in the RMS main window of Cluster Admin.

    Remove the following line from "/opt/SMAW/SMAWRrms/bin/hvenv.local" on all the nodes.

    ```
    export HV_RCSTART=0
    ```

17. When adding, replacing, or moving the system board, take the steps after restarting the use of I/O devices. For details on adding, removing, replacing, or moving steps of the system board, see "SPARC M10 System Domain Configuration Guide."

18. When the NIC is disconnected from the multiplex configuration in step 4, see "3.2 Replacement of the system board using the DR" to embed the NIC to the multiplex configuration.

19. When the disk is detached in step 3, see "2.2.3 Re-attaching the disk" to add the detached disk again.

20. To add or replace the device for the cluster interconnect, add the device to the cluster interconnect.

    ```
    # cfrecon -a <device name>
    ```

Note

When a node ends abnormally (panic or reset) or hangs up due to hardware failure, or the node state becomes LEFTCLUSTER due to the timeout of CF while monitoring of RMS by PRIMECLUSTER is suspended, you should start the cluster applications on a standby node.

- When a node ends abnormally (panic or reset) or hangs up, shut down the node forcibly. After that, wait until the failed node becomes LEFTCLUSTER. You need to wait for the timeout value that was changed in the above step 9.

  1. Return the CF heartbeat timeout value by using the above step 11.
  2. Start PRIMECLUSTER SF by using the above step 13.
  3. When the state of the failed node does not become DOWN, execute the "sdtool -k <CF node name of the host node>" command so that the state of the failed node becomes DOWN.

     ```
     # cftool -n
     Node    Number State       Os      Cpu
     node0   1      UP          Solaris Sparc
     node1   2      LEFTCLUSTER Solaris Sparc
     # sdtool -k node1
     LOG3.013944205091080028 20 6 30 4.3A20 SMAWsf : RCSD returned a successful exit code for
     this command(sdtool -k node1)
     # cftool -n
     Node    Number State       Os      Cpu
     node0   1      UP          Solaris Sparc
     node1   2      DOWN        Solaris Sparc
     #
     ```

     If the failed node remains in the UP state, the "sdtool -k" command fails.

     Wait until the failed node becomes LEFTCLUSTER.

  4. Start PRIMECLUSTER RMS by using the above step 15.
  5. For the operational and standby cluster applications, execute the "hvswitch -f" command to start the cluster applications forcibly.

     ```
     # hvswitch -f <userApplication>
     The use of the -f (force) flag could cause your data to be corrupted and could cause your
     node to be killed. Do not continue if the result of this forced command is not clear.

     The use of force flag of hvswitch overrides the RMS internal security mechanism. In particular
     RMS does no longer prevent resources, which have been marked as "ClusterExclusive", from
     coming Online on more than one host in the cluster. It is recommended to double check the
     state of all affected resources before continuing.

     IMPORTANT: This command may kill nodes on which RMS is not running in order to reduce the
     risk of data corruption!
     ```

```
Ensure that RMS is running on all other nodes. Or shut down OS of the node on which RMS
is not running.

Do you wish to proceed ? (default: no) [yes, no]:yes
```

6. Remove the following line from the /opt/SMAW/SMAWRrms/bin/hvenv.local file.

```
export HV_RCSTART=0
```

- If the node becomes LEFTCLUSTER, take the following steps to clear LEFTCLUSTER.

1. See "6.2.3 Caused by a cluster partition" in "PRIMECLUSTER CF Configuration and Administration Guide" to clear LEFTCLUSTER manually.
2. Return the timeout value of PRIMECLUSTER CF by using the above step 11.
3. Start PRIMECLUSTER SF by using the above step 13.
4. Start PRIMECLUSTER RMS by using the above step 15.
5. For the operational and standby cluster applications, execute the "hvswitch -f" command to start the cluster applications forcibly.

```
# hvswitch -f <userApplication>
The use of the -f (force) flag could cause your data to be corrupted and could cause your
node to be killed. Do not continue if the result of this forced command is not clear.

The use of force flag of hvswitch overrides the RMS internal security mechanism. In particular
RMS does no longer prevent resources, which have been marked as "ClusterExclusive", from
coming Online on more than one host in the cluster. It is recommended to double check the
state of all affected resources before continuing.

IMPORTANT: This command may kill nodes on which RMS is not running in order to reduce the
risk of data corruption!
Ensure that RMS is running on all other nodes. Or shut down OS of the node on which RMS
is not running.

Do you wish to proceed ? (default: no) [yes, no]:yes
```

6. Remove the following line from the /opt/SMAW/SMAWRrms/bin/hvenv.local file.

```
export HV_RCSTART=0
```

4.1.2 Notes

4.1.2.1 Messages to the real time process

In PRIMECLUSTER, a part of the processes work in a real time class. If you add or remove the system board using DR, the following message will appear:

```
Dec 25 21:12:41 Real time processes[pid= 4038 4218 4216 4286 4286 4286 4286 4046 4220 4134 4134 4134
4134 4134 4214 4221 4228 4287 4256 4291 4290 4288 4289 5350 12946 12946 12946 12946 12946 12946 12946
12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946
12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946
12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946] is running.
Do you continue DR ? [YES]/[NO]
```

Check if the displayed process IDs are for the PRIMECLUSTER daemons, then enter "yes" to continue with DR. (The same process ID will be repeatedly displayed for a multithread process.)

The real time processes that are started with PRIMECLUSTER are as follows:

- PRIMECLUSTER

```
rcsd, bm, hvdet_*(Note 1)
```

(Note 1) The detector will be started up from the process bm as a real time process according to an application configuration. The process, in which the parent process is bm, and the process name starts with "hvdet_", is the PRIMECLUSTER detector.

If you frequently use DR, or if you want to make the DR process automatic, you can disable the message by creating an associated script and executing the "dr_conf" command. For details, see the "SPARC M10 System Domain Configuration Guide."

4.1.2.2 Output message during DR

RMS might be suspended during DR. If this occurs, the following warning message might be output:

```
(SYS, 88): WARNING: No heartbeat from cluster host node0RMS within the last 10 seconds.
This may be a temporary problem caused by high system load. RMS will react if this problem persists
for 590 seconds more.
```

This message indicates that the RMS heartbeat monitoring is suspended temporarily. If this message appears during DR, you do not have to take any action.

4.2 DR operation in SPARC Enterprise

You can operate DR-based hot swap or hot expansion in the following cases:

- When an insignificant failure (1 bit intermittent error) that does not disrupt ongoing operations occurs in the CPU or memory.
- When the CPU or memory must be added or reduced due to business expansion or reduction.

- You cannot use DR for an I/O-equipped system board in which the I/O setting is not disabled. To add or reduce the CPU or memory, it is necessary to verify whether the reduced CPU or memory does not affect the operations in advance.
- The kernel memory board cannot be replaced, removed or, moved on all the node models.

4.2.1 Prerequisite

It is necessary to disable the I/O setting in advance for the system board.

For details, see "DR User's Guide."

4.2.2 Notes

4.2.2.1 Messages to the real time process

PRIMECLUSTER makes part of the process to operate in a real time class. If you add or remove the system board using DR, the following message will appear:

```
Dec 25 21:12:41 Real time processes[pid= 4038 4218 4216 4286 4286 4286 4286 4046 4220 4134
 4134 4134 4134 4134 4214 4221 4228 4287 4256 4291 4290 4288 4289 5350 12946 12946 12946
 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946
 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946
 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946 12946
 12946 12946 12946 12946 12946 12946] is running.
 Do you continue DR ? [YES]/[NO]
```

Check if the displayed process IDs are for the PRIMECLUSTER daemons, then enter "yes" to continue with DR. (The same process ID will be repeatedly displayed for a multithread process.)

The real time processes that are started with PRIMECLUSTER are as follows:

- PRIMECLUSTER

```
rcsd, bm, hvdet_* (Note 1)
```

Note 1:

The detector will be started up from the process bm as a real time process according to an application configuration. The process, in which the parent process is bm, and the process name starts with "hvdet_", is the PRIMECLUSTER detector.

If you frequently use DR, or if you want to make the DR process automatic, you can disable the message by creating an associated script and executing the "dr_conf" command. For details, see the "DR Users' Guide".

4.2.2.2 Output message during DR use

RMS might be suspended during DR use. If this occurs, the following warning message might be output:

```
(SYS, 88): WARNING: No heartbeat from cluster host node0RMS within the last 10 seconds.
This may be a temporary problem caused by high system load. RMS will react if this problem persists
for 590 seconds more.
```

This message indicates that the RMS heartbeat monitoring is suspended. If this appears during DR use, you do not have to take any action.

4.2.2.3 Using DR in an overloaded system

If a system board is added, replaced, or inter-partitioned with DR in a cluster system overloaded by cluster applications, this might affect the PRIMECLUSTER node monitoring facility resulting in node elimination.

Stop the node monitoring facility by using the following steps, so you can continue operation.

1. Check the PRIMECLUSTER configuration file name by executing the "hvdisp -n" command on any of the nodes where RMS is running. In the example below, the RMS configuration file name is "config.us."

   ```
   # hvdisp -n
   /opt/SMAW/SMAWRrms/build/config.us
   #
   ```

2. Stop RMS by executing the "hvshut" command on all the nodes. Answer "yes", then only RMS will stop.

   ```
   # hvshut -L
                              WARNING
                              -------
   The '-L' option of the hvshut command will shut down the RMS
   software without bringing down any of the applications.
   In this situation, it would be possible to bring up the same
   application on another node in the cluster which *may* cause
   data corruption.

   Do you wish to proceed ? (yes = shut down RMS / no = leave RMS running).
   yes

   NOTICE: User has been warned of 'hvshut -L' and has elected to proceed.
   ```

   Add the following line in the "/opt/SMAW/SMAWRrms/bin/hvenv.local" file on all the nodes. Then, RMS will not automatically be started.

   ```
   export HV_RCSTART=0
   ```

3. Stop SF by executing the "sdtool" command on all the nodes.

   ```
   # sdtool -e
   (SMAWsf, 30, 11) : RCSD returned a successful exit code for this command
   ```

4. Change the timeout value of CF heartbeat monitoring on all the nodes as follows:

   - Add the following line in the "/etc/default/cluster.config" file on all the nodes, so the CF heartbeat timeout will be 600 seconds.

     ```
     CLUSTER_TIMEOUT "600"
     ```

- Execute the following command on all the nodes.

```
# cfset -r
```

- Check whether or not the CF timeout value is valid.

```
# cfset -g CLUSTER_TIMEOUT
>From cfset configuration in CF module:
Value for key: CLUSTER_TIMEOUT --->600
#
```

5. Use DR.

6. Return the CF heartbeat timeout to the default value on all the nodes as follows:

   - Change the CLUSTER_TIMEOUT in /etc/default/cluster.config to 10.

     Before change

     ```
     CLUSTER_TIMEOUT "600"
     ```

     After change

     ```
     CLUSTER_TIMEOUT "10"
     ```

   - Execute a following command on each nodes.

     ```
     # cfset -r
     ```

   - Check whether or not the CF timeout is valid using the following command on all the nodes.

     ```
     # cfset -g CLUSTER_TIMEOUT
     >From cfset configuration in CF module:
     Value for key: CLUSTER_TIMEOUT --->10
     #
     ```

7. Start SF by executing the "sdtool" command on all the nodes.

   ```
   # sdtool -b
   ```

8. Check if the SF is running.

   Select the [Tools]-[Shutdown Facility]-[Show Status] menu from the CF main window of Cluster Admin, then check the "Test State" field on each node.

9. Start RMS by executing the "hvcm" command on all the nodes. Specify the RMS configuration file name, which is checked at step 1, for the "-c" option. For example, if the name is "/opt/SMAW/SMAWRrms/build/config.us", "config" will be specified.

   ```
   # hvcm -c config
   Starting Reliant Monitor Services now
   ```

10. RMS must be running on all the nodes. Check if each icon indicating the node state is green (Online) in the RMS main window of Cluster Admin.

    Remove the following line from "/opt/SMAW/SMAWRrms/bin/hvenv.local" on all the nodes. Then, RMS startup will automatically be enabled.

    ```
    export HV_RCSTART=0
    ```

- Be sure to verify a cluster system during cluster configuration using the above steps.

- If a node failure such as a node panic or reset occurs during step 3 through 7 or the state of a node becomes LEFTCLUSTER due to the timeout of CF, cluster applications on a standby node must be started.

  - When a node ends abnormally (panic or reset) or hangs up, shut down the node forcibly. After that, wait until the failed node becomes LEFTCLUSTER. You need to wait for the timeout value that was changed in the above step 4.

    Check that the failed not is not running, and then perform the following procedure:

    1. Return the timeout value of PRIMECLUSTER CF by using the above step 6.
    2. Start PRIMECLUSTER SF by using the above step 7.
    3. When the state of the failed node does not become DOWN, execute the "sdtool -k <CF node name of the host node>" command so that the state of the host node becomes DOWN.

       ```
       # cftool -n
       Node    Number State       Os      Cpu
       node0   1      UP          Solaris Sparc
       node1   2      LEFTCLUSTER Solaris Sparc
       # sdtool -k node1
       LOG3.013944205091080028 20 6 30 4.3A20 SMAWsf : RCSD returned a successful exit code for
       this command(sdtool -k node1)
       # cftool -n
       Node    Number State       Os      Cpu
       node0   1      UP          Solaris Sparc
       node1   2      DOWN        Solaris Sparc
       #
       ```

       If the failed node remains in the UP state, the "sdtool -k" command fails.

       Wait until the failed node becomes LEFTCLUSTER.

    4. Start PRIMECLUSTER RMS by using the steps from 6 through 9.
    5. For the operational and standby cluster applications, execute the "hvswitch -f" command to start the cluster applications forcibly.

       ```
       # hvswitch -f <userApplication>
       The use of the -f (force) flag could cause your data to be corrupted and could cause your
       node to be killed. Do not continue if the result of this forced command is not clear.

       The use of force flag of hvswitch overrides the RMS internal security mechanism. In
       particular RMS does no longer prevent resources, which have been marked as
       "ClusterExclusive", from coming Online on more than one host in the cluster. It is
       recommended to double check the state of all affected resources before continuing.

       IMPORTANT: This command may kill nodes on which RMS is not running in order to reduce the
       risk of data corruption!
       Ensure that RMS is running on all other nodes. Or shut down OS of the node on which RMS
       is not running.

       Do you wish to proceed ? (default: no) [yes, no]:yes
       ```

    6. Remove the following line from the /opt/SMAW/SMAWRrms/bin/hvenv.local file.

       ```
       export HV_RCSTART=0
       ```

  - If the node becomes LEFTCLUSTER, take the following steps to clear LEFTCLUSTER.

    1. See "6.2.3 Caused by a cluster partition" in "PRIMECLUSTER CF Configuration and Administration Guide" to clear LEFTCLUSTER manually.
    2. Return the timeout value of PRIMECLUSTER CF by using the above step 6.
    3. Start PRIMECLUSTER SF by using the above step 7.
    4. Start PRIMECLUSTER RMS by using the above step 9.

5. For the operational and standby cluster applications, execute the "hvswitch -f" command to start the cluster applications forcibly.

```
# hvswitch -f <userApplication>
The use of the -f (force) flag could cause your data to be corrupted and could cause your
node to be killed. Do not continue if the result of this forced command is not clear.

The use of force flag of hvswitch overrides the RMS internal security mechanism. In
particular RMS does no longer prevent resources, which have been marked as
"ClusterExclusive", from coming Online on more than one host in the cluster. It is
recommended to double check the state of all affected resources before continuing.

IMPORTANT: This command may kill nodes on which RMS is not running in order to reduce the
risk of data corruption!
Ensure that RMS is running on all other nodes. Or shut down OS of the node on which RMS
is not running.

Do you wish to proceed ? (default: no) [yes, no]:yes
```

6. Remove the following line from the /opt/SMAW/SMAWRrms/bin/hvenv.local file.

```
export HV_RCSTART=0
```

- PRIMECLUSTER do not support the unmanned DR operation with time scheduling.

4.3 PCI Hot Plug operation in SPARC M10/SPARC Enterprise

4.3.1 Replacing and adding the NIC with PCI Hot Plug

To enable NIC hot swap and hot expansion with PCI Hot Plug, the NC must be redundant.

Note

- To replace the NIC using PCI Hot Plug, the network interface must be active. Stop the node and replace the NIC if the following public LAN is used:

  * The interface attribute of the public LAN and takeover IP address is set to "base", and

  * The public LAN uses PRIMECLUSTER GLS (Multipath) in which the network interface is inactive on a standby node or a failed node.

- You cannot replace the NIC using PCI Hot Plug if the CF configuration is not complete. If you are aware of it after PRIMECLUSTER installation, stop the node and replace the NIC. For information on how to configure CF, see the "PRIMECLUSTER Cluster Foundation Configuration and Administration Guide."

- You cannot replace the NIC using PCI Hot Plug if Scalable Internet Services (SIS) is used (SMAWdtcp is installed). If you want to replace the NIC, it is necessary to uninstall SMAWdtcp in single user mode in advance. If you want to use the Scalable Internet Services (SIS), stop the node and replace the NIC.

4.3.1.1 Replacing the public LAN

If GLS is used, see "3.3 NIC hot swap and hot expansion with PCI Hot Plug."

If GLS is not used, see "Oracle Solaris Administration: Devices and File Systems."

4.3.1.2 Adding the public LAN

If GLS is used, see "3.3 NIC hot swap and hot expansion with PCI Hot Plug."

If GLS is not used, see "Oracle Solaris Administration: Devices and File Systems."

4.3.1.3 Replacing the cluster interconnect

If a cluster interconnect failure occurs, stop the node and replace the cluster interconnect.

4.3.2 Replacing or adding the PCI card connected to the shared disk

To enable the PCI Hot Plug based hot swap and hot expansion of the PCI card that is connected to a shared disk, the shared disk must be configured as follows:

- A RAID unit must be redundant with multipath driver
- An SCSI device must be redundant with GDS (for swapping only)

4.3.2.1 Replacing or adding the PCI card connected to the redundant RAID

For information on how to replace a PCI card that is connected to the redundant RAID device with multipath driver, and how to add a path along with the addition of a PCI card, see the "Oracle Solaris Administration: Devices and File Systems."

4.3.2.2 Replacing the PCI card connected to the redundant SCSI device

For information on how to replace a PCI card that is connected to the redundant SCSI device with GDS, see "Chapter 2 GDS".

4.3.3 Replacing the cluster interconnect (Ethernet)

The procedure for replacing the PCI card can vary depending on whether the PCI card is only connected to the cluster interconnect or connected to the cluster interconnect and GLS Public LAN.

4.3.3.1 The PCI card is only connected to the cluster interconnect

1. Save hardware configuration information in the file before replacing the cluster interconnect.

```
# /usr/sbin/prtpicl -v > pre_prtpicl.log
```

2. Check the location of the PCI slot on which the PCI card is installed. Check the PCI slot by blinking the ALARM LED of the PCI slot. Stop the blinking LED after the slot is checked.

```
# cftool -d
Number Device    Type Speed    Mtu     State  Configured  Address
1      /dev/fjgi1  4    100     1432     UP    YES       00.00.0e.25.1a.38
2      /dev/fjgi7  4    100     1432     UP    YES       00.00.0e.25.1a.38
# /usr/sbin/FJSVmadm/inst2comp fjgi7
pcipsy8:R0B00-PCI#slot05
# cfgadm -x led=fault,mode=blink pcipsy8:R0B00-PCI#slot05
# cfgadm -x led=fault,mode=off pcipsy8:R0B00-PCI#slot05
```

3. Disconnect the device from the cluster interconnect. Then, check if "Receptacle" is "disconnected" and "Occupant" is "unconfigured" respectively.

```
# cfrecon -d /dev/fjgi7

Delete Cluster Interconnect:

device "/dev/fjgi7" successfully removed from node configuration
# cfgadm -c disconnect pcipsy8:R0B00-PCI#slot05
# cfgadm pcipsy8:R0B00-PCI#slot05
Ap_Id                          Type       Receptacle    Occupant    Condition
pcipsy8:R0B00-PCI#slot05     unknown    disconnected  unconfigured  unknown
```

4. CE removes the PCI card that is installed on the PCI slot.
5. CE installs a new PCI card on the PCI slot.

6. Determine the value that indicates the PCI slot location.

   See the "ap_id" format described in "1.3.1 Displaying the PCI card" of the "Oracle Solaris Administration: Devices and File Systems.".

7. Enable the PCI card on the Solaris OE. Specify the value at step 6 after "cfgadm -c configure". Then, check if "Receptacle" is "connected" and "Occupant" is "configured" respectively.

```
# cfgadm -c configure pcipsy18:R0B02-PCI#slot02
# cfgadm pcipsy18:R0B02-PCI#slot02
Ap_Id                         Type       Receptacle    Occupant    Condition
pcipsy18:R0B02-PCI#slot02     bridge/hp    connected    configured    ok
```

8. Save hardware configuration information in a different file than that previously used.

```
# /usr/sbin/prtpicl -v > post_prtpicl.log
```

9. Compare the files and specify the interface name and number of the PCI card.

```
# diff -c pre_prtpicl.log post_prtpicl.log | more
(snip)
!                :devfs-path    /pci@9c,4000/SUNW,hme@3,1
!                :driver-name    hme
!                :binding-name    SUNW,hme
!                :bus-addr     3,1
!                :instance         9
!                :_class         obp-device
!                :name        SUNW,hme
 (The rest is omitted)
```

10. Add the PCI card in the cluster interconnect using the interface name and number.

```
# cfrecon -a /dev/hme9

Add Cluster Interconnect:

device "/dev/hme9" successfully added to node configuration
```

11. Check that the following message is output to the console of each node.

```
cf:eventlog     CF: Route recovery on /dev/hme9 to node node0. (#0000 1 2 2 2)
```

12. Confirm that the PCI card has successfully been added in the cluster interconnect.

```
# cftool -d
Number Device     Type   Speed  Mtu   State  Configured  Address
1      /dev/fjgi1  4     100    1432  UP     YES         00.00.0e.25.1a.38
2      /dev/hme9   4     100    1432  UP     YES         00.00.0e.25.1a.38
```

4.3.3.2 When the cluster interconnect and GLS public LAN are connected to the PCI card

Note

In this section, the cluster interconnect and GLS public LAN are connected to each port of the PCI card.

You cannot use a different device name before or after the PCI card replacement with PCI Hot Plug. Be sure to install the same PCI card into the same PCI slot as before its replacement.

1. Delete the GLS physical interface using the following command. Change the interface name in the arguments of the following command according to a GLS configuration.

   - Fast switching mode

```
# /opt/FJSVhanet/usr/sbin/hanetnic delete -n sha0 -i fjgi6
# /usr/sbin/ifconfig fjgi6 unplumb
```

   - NIC switching mode

```
# /opt/FJSVhanet/usr/sbin/hanetpoll off
# /opt/FJSVhanet/usr/sbin/stpptl -n sha1
# /opt/FJSVhanet/usr/sbin/hanetnic change -n sha0
```

2. Check the location of the PCI slot on which the PCI card is installed. Check the PCI slot by blinking the ALARM LED of the PCI slot (*1). Stop the blinking LED (*2) after the slot is checked.

```
# cftool -d
Number Device    Type  Speed  Mtu   State Configured  Address
1      /dev/fjgi1  4     100   1432  UP     YES          00.00.0e.25.1a.38
2      /dev/fjgi7  4     100   1432  UP     YES          00.00.0e.25.1a.38
# /usr/sbin/FJSVmadm/inst2comp fjgi7
pcipsy8:R0B00-PCI#slot05
# cfgadm -x led=fault,mode=blink pcipsy8:R0B00-PCI#slot05          (*1)
# cfgadm -x led=fault,mode=off pcipsy8:R0B00-PCI#slot05            (*2)
```

3. Disconnect the device from the cluster interconnect by using the device name and PCI slot number checked in step 2 . Then, check if "Receptacle" is "disconnected" and "Occupant" is "unconfigured" respectively.

```
# cfrecon -d /dev/fjgi7

Delete Cluster Interconnect:

device "/dev/fjgi7" successfully removed from node configuration
# cfgadm -c disconnect pcipsy8:R0B00-PCI#slot05
# cfgadm pcipsy8:R0B00-PCI#slot05
Ap_Id                              Type       Receptacle    Occupant    Condition
pcipsy8:R0B00-PCI#slot05     unknown    disconnected  unconfigured  unknown
```

4. CE replaces the PCI card that is installed on the PCI slot.

5. Enable the replaced PCI card on the Solaris OE. Check if "Receptacle" is "connected" and "Occupant" is "configured" respectively.

```
# cfgadm -c configure pcipsy8:R0B00-PCI#slot05
# cfgadm pcipsy8:R0B00-PCI#slot05
Ap_Id                       Type       Receptacle   Occupant   Condition
pcipsy8:R0B00-PCI#slot05      bridge/hp    connected   configured   ok
```

6. Add the PCI card to the GLS settings using the following command. Change the interface name and IP address in the arguments of the following command according to a GLS configuration.

   - Fast switching mode

```
# /usr/sbin/ifconfig fjgi6 plumb
# /usr/sbin/ifconfig fjgi6 192.168.10.10 netmask + broadcast + -trailers up
# /opt/FJSVhanet/usr/sbin/hanetnic add -n sha0 -i fjgi6
```

   - NIC switching mode

```
# /opt/FJSVhanet/usr/sbin/hanetnic change -n sha0
# /opt/FJSVhanet/usr/sbin/strptl -n sha1
# /opt/FJSVhanet/usr/sbin/hanetpoll on
```

7. Add the replaced PCI card to the cluster interconnect.

```
# cfrecon -a /dev/fjgi7

Add Cluster Interconnect:

device "/dev/fjgi7" successfully added to node configuration
```

8. Check that the following message is output to the console of each node

```
cf:eventlog     CF: Route recovery on /dev/fjgi7 to node node0. (#0000 1 2 2 2)
```

9. Confirm that the PCI card has successfully been added in the cluster interconnect.

```
# cftool -d
Number Device   Type  Speed  Mtu   State  Configured  Address
1      /dev/fjgi1  4    100   1432   UP     YES          00.00.0e.25.1a.38
2      /dev/fjgi7  4    100   1432   UP     YES          00.00.0e.25.1a.38
```

Index