PRIMECLUSTER
DR/PCI Hot Plug Users' Guide

- SPARC Solaris -

Version 4.4

January, 2005

Introduction

The *PRIMECLUSTER DR/PCI Hot Plug User's Guide* contains information on how to operate the PRIMEPOWER Dynamic Reconfiguration (DR) 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:
  - Hot swap of an I/O-connected system board
  - Hot swap, hot expansion, and inter-partition movement of a system board* on which CPU/memory boards are mounted
  - Hot swap and hot expansion of a PCI card

  For further details, see "1.3 DR standard operation modes" and "1.4 PCI Hot Plug standard operation modes".

- ***Cluster system***
  For a cluster system that uses cluster software like PRIMECLUSTER, the following operations are supported:
  - Hot swap, hot expansion, and inter-partition movement of a system board* on which CPU/memory boards are mounted
  - Hot swap and hot expansion of a PCI card

  * You can operate hot swap, hot expansion, and inter-partition movement of I/O-mounted boards by disabling I/O. See "5.1.1 Prerequisites".

Product models

The document describes PRIMEPOWER450 (pedestal/rack mount type 7U), 800, 900, 1000, 1500, 2000, and 2500. In this document, the pedestal/rack mount (7U) models are referred to as PRIMEPOWER450, and the SPARC64 V mounted models are referred to as PRIMEPOWER1500.

Products

The document covers the following products:

PRIMECLUSTER GDS (hereafter GDS)
PRIMECLUSTER GFS (hereafter GFS)
PRIMECLUSTER GLS (hereafter GLS)
PRIMECLUSTER

Reference

"Dynamic Reconfiguration Users' Guide" (Hereafter "DR User's Guide.")
"Dynamic Reconfiguration Users' Guide I/O device edition" (Hereafter "DR Users' Guide I/O device edition.")
"PCI Hot Plug Users' Guide"
"PCI Hot Plug Users' Guide I/O device edition"

Trademarks

Contents

1. Overview

1.1 DR

The document covers the following PRIMECLUSTER products and versions that operate in a Solaris 8 or Solaris 9 environment of PRIMEPOWER800/900/1000/1500/2000/2500:

- PRIMECLUSTER Storage Suite (PRIMECLUSTER GDS/GFS), PRIMECLUSTER GLS 4.1A10 or later
- PRIMECLUSTER Enterprise Edition, HA Server 4.1A10 or later

1.2 PCI Hot Plug

The document covers the following PRIMECLUSTER products and versions that operate in a Solaris 8 or Solaris 9 environment on PRIMEPOWER450 (pedestal /rack-mount type (7U))/900/1500 (with SPARC64 V) /2500:

- PRIMECLUSTER Storage Suite (PRIMECLUSTER GDS/GFS), PRIMECLUSTER GLS 4.1A10 or later
- PRIMECLUSTER Enterprise Edition, HA Server 4.1A10 or later

1.3 DR standard operation modes

The PRIMEPOWER DR (Dynamic Reconfiguration) is used to add or remove a system board without disrupting ongoing operations in partitions. DR can be used for the following operations:

- Failed component hot swap
  Hot swap* of a system board for replacing a failed I/O device, memory, or CPU

- Hardware resource hot expansion
  Addition of a system board for adding the CPU, memory, or I/O device

- Hardware resource inter-partition movement and system board pool
  Inter-partition movement 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.

You can even operate hot swap of I/O mounted boards by disabling I/O (See “5.1 Prerequisites”).

* If you add the system board that has previously been removed, the I/O settings such as I/O controller, I/O device, and mounted slot locations must be the same as those used before the removal.

1.4 PCI Hot Plug standard operation modes

The PRIMEPOWER PCI Hot Plug is used to add or remove a PCI card in a Solaris™ operating 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 PRIMECLSTER

PRIMECLUSTER supports the following DR operation modes. The range of support varies depending on whether I/O is mounted on the system board.

	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 (*7)	Available (*3)(*7)	Available	Not available	Available (*7)	Not available
GFS	Available	See below (*4)	Available	Not available	Available	Not available
PRIMECLUSTER	Available (*5)(*6)	Not available	Available (*5)	Not available	Available (*5)(*6)	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.

(*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.

(*4) For a system board that is connected to a non-redundant I/O disk with GFS, you must stop (umount) GFS. For details, see Chapter 4.

(*5) Observe the warnings and cautions in Chapter 5.

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

(*7) For a system where the GLS Redundant Line Control Function is used, the kernel memory board cannot be removed (replaced or moved) on PRIMEPOWER 900/1500/2500. If you want to make this possible, it is necessary to apply the patch 912419-04 or later for the GLS Multipath function.

1.6 Supported PCI Hot Plug operations on PRIMECLSTER

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)(*4)(*5)

(*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) PCL4.1A20 or later is required.

(*4) The Patch 901146-10 or later (for Solaris 8) or 901147-10 or later (for Solaris 9) is required on PCL4.1A20.

(*5) Only hot swap will be supported.

2. 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 for root or local class)
- 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

1. Required patch

   It is necessary to apply the Patch 912704-03 or later to operate DR or PCI Hot Plug based hot swap or hot expansion of a system board or PCI card that is connected to a GDS disk.

2. 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 4.1."

3. The mirrored volumes must be synchronized.

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

   1) 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

   2) 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).

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

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

   2) 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.

5. 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.

6. 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

1. 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 a system board with DR

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

2.2.1 Removing a system board

(1) Prior check (checkremove or preremove of an associated script)

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.

1-1) Check the name of the disk device that is connected to the system board using the “drcstat“ command.

Example: When sb01 will be removed

```
# drcstat -device sb01 | grep sd
01-PCI#0B  "/pci@81,4000/scsi@2/sd@1,0" 1 "sd"
```

Note: Use the “dr_info“ command to make the system board removal automatic with an associated script.

1-2) Check the logical device name (c#t#d#) of the disk.

```
# ls -l /dev/dsk | grep /pci@81,4000/scsi@2/sd@1,0:c   <--(Note) Suffix ":c".
lrwxrwxrwx   1 root     root         55 Sep  6  1999 c1t1d0s2-> ../..
/devices/pci@81,4000/scsi@2/sd@1,0:c
```

1-3) 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.

1-4) 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 (preremove of an associated script)

2-1) Prior check and name confirmation

Confirm if you can remove the system board using DR according to step 2.2.1 (1), then check the class name and the SDX disk name at step 1-3) and 1-4).

2-2) Remove the disk using the “sdxswap” command.

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

2-3) Confirm that the disk has been removed correctly (whether the disk has entered SWAP).

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

(3) Remove the system board with DR.

Remove the system board using DR. For details, see the “DR Users’ Guide”.

```
# drc -disconnect sb01 -keep
```

2.2.2 Adding the system board

(1) Add the system board with DR.

Add the system board again using DR. For details, see the “DR Users’ Guide”.

```
# drc -connect sb01
```

(2) After addition (postadd)

2-1) Check the device name of the disk that is connected to the system board using the “drcstat“ command.

```
# drcstat -device sb01 | grep sd
01-PCI#0B  "/pci@81,4000/scsi@2/sd@1,0" 1 "sd"
```

2-2) Check the logical device name (c#t#d#) of the disk.

```
# ls -l /dev/dsk | grep /pci@81,4000/scsi@2/sd@1,0:c  <-(Note) Suffix ":c"
lrwxrwxrwx   1 root     root          55 Sep  6  1999 c1t1d0s2 -> ../../devices/pci@81,4000/scsi@2/sd@1,0:c
```

2-3) 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
```

2-4) Attach the disk again using the “sdxswap” command.

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

2-5) 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
```

2-6) Wait until the synchronization copy to the disk is done.

```
# sdxinfo -S -o disk01
OBJ    CLASS   GROUP   DISK    VOLUME  STATUS
------ ------- ------- ------- ------- --------
slice  class01 grp01   disk01  v01     COPY   <-Wait until the state changes to ACTIVE.
```

```
slice  class01 grp01   disk01  v02    COPY
slice  class01 grp01   disk02  v01    ACTIVE
slice  class01 grp01   disk02  v02    ACTIVE
```

2.3 Replacing the PCI card with PCI Hot Plug

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

2.3.1 Detaching the disk

1) Stop the hardware monitoring.

   Stop the hardware monitoring facility of the machine management. For details see the “PCI Hot Plug Users’ Guide I/O device”.

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 “PCI Hot Plug Users’ Guide I/O device”.

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
```

2.3.2 Removing, replacing and adding the PCI card

For details, see the “PCI Hot Plug Users’ Guide I/O device”.

2.3.3 Adding the disk again

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).

1-1) Search the node ID.

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

1-2) Search the Resource ID.

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

```
#
```

1-3) Delete the resource information.

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

1-4) Execute the automatic resource registration, and then the following command on all the nodes:

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

1) 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
#
```

2) 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
#
```

2.4 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 “PCI Hot Plug Users’ Guide I/O device”.

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.”

3. 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.

1. 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.

2. 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 DR associated script

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

GLS automatically uses DR using the DR associated script, so you do not have to stop the NIC. For details on the GLS DR associated function, see the following chapters in the "PRIMECLUSTER Global Link Services Configuration and Administration Guide (Redundant Line Control Function)":

- "2.2.14 DR (Dynamic Reconfiguration) associated function"
- "3.3.13 Setting DR (Dynamic Reconfiguration)"
- "4.5.6 Recover the transmission route"

Multipath

Similarly, GLS automatically uses DR using the DR associated script, so you do not have to stop the NIC. For details on the GLS DR associated function, see the following chapter in the online manual, "PRIMECLUSTER Global Link Services Configuration and Administration Guide (Multipath)".

- "Chapter 8 DR (Dynamic Reconfiguration)"

3.3 NIC hot swap and hot expansion with PCI Hot Plug

See the "PCI Hot Plug User's Guide I/O device". Note that CF configuration must be complete if the system is in a cluster environment. For details, see "5.2.1 Replacing and adding the NIC with PCI Hot Plug" of this guide.

4. GFS local file system

The procedure for DR-based hot swap and hot expansion for a disk that is using a GFS local file system is fundamentally the same as that for UFS. It is necessary to stop (umount) the file system in the event a device is not in a redundant I/O configuration on the system board and is used for the GFS local file system. However, it is not necessary to stop the file system if the disk is in a redundant I/O configuration, and can be removed from the file system.

Similarly, it is necessary to stop (umount) the file system if the device is not in a redundant I/O configuration on the PCI board and is used for the GFS local file system. However, it is not necessary to stop the file system if the disk is in a redundant I/O configuration and can be removed from the file system.

Be aware that the GFS local file system supports the integrated management of multiple slices as a single file system. If you are using this multi-slice configuration, you must execute the command specific to the GFS local file system to check the mounted slices.

This document describes the procedure for removing a system board with DR and replacing a system board using PCI Hot Plug when the GFS local file system is created in a multi-slice configuration and mounted on a single partition.

4.1 Removing the system board with DR

This section explains how to stop a failed file system.

a) Unmount the failed file system in which an I/O error is detected using the “df” command.

b) Find the file system that might affect the PCI card removal using the “sfxinfo“ command, then unmount it.

4.1.1 Taking corrective action against I/O errors

Check the file system in which an I/O error is detected using the “df” command, then unmount it.

1) Execute the “df” command as specified by the GFS local file system.

You can check whether or not the GFS local file system can be mounted by specifying -F sfxfs to the “df” command.

```
# df -F sfxfs
/DOC          (/dev/dsk/c1t9d0s1 ):  209088 blocks   113211 files
/PKG          (/dev/dsk/c1t10d0s1):  377200 blocks   172394 files
df: cannot statvfs /sfxfs3: I/O error
```

2) Remove the failed file system.

Unmount any file systems in which I/O errors are detected.

If there are any active processes or dirty data, it is necessary to execute the “umount” command with –f.

```
# umount -f /sfxfs3
```

4.1.2 Checking the system board and physical disk slice name

Check the system board and the disk slice name that might be affected by the system board removal according to the “DR Users’ Guide”.

Note:

If the device is used for the file system and is a logical device, it cannot be used depending on specifications of the logical device driver.

4.1.3 Checking the multi-slice configuration

You can check a slice configuration and mount points of the GFS local file system using the “sfxinfo” command without arguments. If the slice affects the system board removal, unmount the file system based on the mount point name.

```
# sfxinfo
ID      special                          size      mount     Type
0       /dev/dsk/c1t9d0s1(800041)       101763    /sfxfs1    META
0       /dev/dsk/c1t9d0s1(800041)       938080    /sfxfs1    DATA
0       /dev/dsk/c1t9d0s1(800041)        10503    /sfxfs1    LOG
ID      special                          size      mount     Type
0       /dev/dsk/c1t10d0s1(800049)      203085    /sfxfs2    META
0       /dev/dsk/c1t10d0s1(800049)     1874840    /sfxfs2    DATA
0       /dev/dsk/c1t10d0s1(800049)       20989    /sfxfs2    LOG
1       /dev/dsk/c1t11d0s1(800051)     1874840    /sfxfs2    DATA
```

Note:

The “umount” command fails if it is executed while an application is using the file system. It is necessary to stop the application, or unmount the file system using the “umount” command with –f.

4.2 Replacing the PCI Card with PCI Hot Plug

This section explains how to stop a failed file system.

a) Unmount the failed file system in which an I/O error is detected using the “df” command.

b) Find the file system that might affect the PCI card removal using the “sfxinfo“ command, then unmount it.

4.2.1 Taking corrective action against I/O errors

Check the file system in which an I/O error is detected using the “df” command, then unmount it.

1) Execute the “df“ command

You can check whether or not the GFS local file system can be mounted by specifying -F sfxfs to the “df” command.

```
# df -F sfxfs
/DOC          (/dev/dsk/c1t9d0s1 ):  209088 blocks   113211 files
/PKG          (/dev/dsk/c1t10d0s1):  377200 blocks   172394 files
df: cannot statvfs /sfxfs3: I/O error
```

2) Remove the failed file system.

Unmount any file systems in which I/O errors are detected.

If there are any active processes or dirty data, it is necessary to execute the “umount” command with –f.

```
# umount -f /sfxfs3
```

4.2.2 Checking the physical disk slice name

Check the PCI card and the physical disk slice name according to the “DR Users’ Guide”.

Note:

If the device is used for the file system and is a logical device, it might not be used depending on specifications of the logical device driver.

4.2.3 Checking the multi-slice configuration

You can check a slice configuration and mount points of the GFS local file system using the “sfxinfo” command without arguments. If the slice affects the PCI card removal, unmount the file system based on the mount point name.

```
# sfxinfo
ID      special                        size      mount     Type
0       /dev/dsk/c1t9d0s1(800041)      101763    /sfxfs1    META
0       /dev/dsk/c1t9d0s1(800041)      938080    /sfxfs1    DATA
0       /dev/dsk/c1t9d0s1(800041)       10503    /sfxfs1    LOG
ID      special                        size      mount      Type
0       /dev/dsk/c1t10d0s1(800049)     203085    /sfxfs2    META
0       /dev/dsk/c1t10d0s1(800049)    1874840   /sfxfs2     DATA
0       /dev/dsk/c1t10d0s1(800049)      20989    /sfxfs2    LOG
1       /dev/dsk/c1t11d0s1(800051)    1874840   /sfxfs2     DATA
```

Note:

The “umount” command fails if it is executed when an application is using the file system. It is necessary to stop the application, or unmount the file system using the “umount” command with –f.

5. PRIMECLUSTER

5.1 Using DR

In a PRIMECLUSTER system, it is recommended that you replace, add or remove the system board after stopping a failed node by generating a failover without disrupting ongoing operations.

You can also 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.

Note:

You cannot use DR for an I/O-mounted system board in which the OBP 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.

5.1.1 Prerequisite

It is necessary to disable the OBP I/O setting in advance for the system board. Specify the number of the system boards in the OBP variable, **no-io-sb-c*X*** (***X*** indicates a unit number).

For details, see the “DR Users’ Guide”.

5.1.2 Notes

5.1.2.1 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] 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.

- PRIMECLUSTER GFS shared file system

  sfcfsmg

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”.

5.1.2.2 Output messages during DR use

RMS might be suspended during DR use. If this occurs, the following error and warning messages might be displayed:

(BM, 65): NOTICE: Base monitor heartbeat between this host and host <coll0e12RMS> is off.
(NOD, 27): NOTICE: Detector for cluster host coll0e12RMS has reported offline.
(WRP, 34): ERROR: Cluster host node is no longer in time sync with local node. Sane operation of RMS can no longer be guaranteed. Further out-of-sync messages will appear in the syslog.
(SYS, 88): WARNING: Cluster hostcoll0e12RMS transitioned from online to offline unexpectedly. Allowing 45 more seconds to expire before reacting.
(BM, 66): NOTICE: Base monitor heartbeat between this host and host <coll0e12RMS> is on.
(NOD, 27): NOTICE: Detector for cluster host coll0e12RMS has reported online.
(US, 11): NOTICE: Cluster host coll0e12RMS has recovered and reported online once again.

These messages indicate that the RMS heartbeat monitoring is suspended. If they appear during DR use, you do not have to take any action.

5.1.2.3 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 that the NULLDETECTOR attribute is not being used in the RMS configuration by executing the following command on any of the nodes where RMS is running. Be aware that you cannot use DR if "NullDetector" is output.

NULLDETECTOR attribute is used in the RMS configuration.

```
# hvconfig -l|grep -i nulldetect
        NullDetector = scripts; // RMS_IgnoreCheckSum
        host * NullDetector = "scripts"; // RMS_IgnoreCheckSum
        PreCheckScript = "hvexec -p app1 nulldetect";
#
```

NULLDETECTOR attribute is not used in the RMS configuration.

```
# hvconfig -l|grep -i nulldetect
#
```

2) Check the RMS configuration file name by executing the "hvdisp" command on any of the nodes where RMS is running.

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

3) 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
```

4) 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
```

5) 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
     #
     ```

6) Use DR.

7) 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: `CLUSTER_TIMEOUT "600"`
     after : `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
     #
     ```

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

```
# sdtool -b
```

9) 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.

10) Start RMS by executing the "hvcm" command on all the nodes. Specify the RMS configuration file name, which is checked at step 2, 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
```

11) 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
```

Cautions:

- 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 4 through 8, cluster applications on a standby node must be started. Stop the failed node manually. Check that the node is stopped, then follow the procedure below.
  - Mark the node DOWN by executing the "cftool" command and specifying the node number and CF node name.

```
---------------------------------------------------------------------------------------------------
# cftool -n
Node  Number State        Os      Cpu
node0 1      UP           Solaris Sparc
node1 2      LEFTCLUSTER Solaris Sparc
# 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: 2
Enter name for node #2: node1
cftool(down): declaring node #2 (node1) down
cftool(down): node node1 is down
# cftool -n
Node  Number State       Os       Cpu
node0 1      UP          Solaris  Sparc
node1 2      DOWN        Solaris  Sparc
#
---------------------------------------------------------------------------------------------------
```

  - Start RMS using the steps from 7 through 10.
  - Answer "yes" for the following question when you start RMS in standby operation. For details of the message, see the "PRIMECLUSTER Installation and Administration Guide".

```
---------------------------------------------------------------------------------------------------
1421 The userApplication " userApplication " did not start
automatically because not all of the nodes where it can run are
online. Do you want to force the userApplication online on the
SysNode " SysNode "?
---------------------------------------------------------------------------------------------------
```

    Note) The operator intervention request does not function by default. You can set up the function by referring to "5.4 Setting Up Fault Resource Identification and Operator Intervention Request" of the "PRIMECLUSTER Installation and Administration Guide". If you do not set it up, it is necessary to use the "hvswitch" command. For details, see the "-f" option of the online manual.

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

5.1.2.4 Restrictions during DR use

Pay careful attention to the following restrictions when a cluster-interconnect or node failure occurs during DR use:

[Restrictions]

The PRIMEPOWER SCF driver is suspended during DR use, so you are not allowed to use SCF/RCI with the PRIMECLUSTER inter-node monitoring facility. Be aware of the following:

1. In an environment where a single cluster interconnect is used, operation is stopped if an error occurs during DR use. Do not use DR if the cluster interconnect is not redundant.
2. The RCI monitoring agent cannot quickly detect a panic of the other node during DR use. The heartbeat monitoring will detect the node panic after a delay of approximately 10 seconds.

5.2 Using PCI Hot Plug

5.2.1 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.

5.2.1.1 Replacing the public LAN

Follow the public LAN replacement procedure described in “Redundant configuration” of the “PCI Hot Plug User’s Guide I/O device”.

5.2.1.2 Adding the public LAN

For information on how to add a path along with the addition of a PCI card, follow the public LAN addition procedure described in the “PCI Hot Plug User’s Guide I/O device”.

5.2.1.3 Replacing the cluster interconnect

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

5.2.2 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 MPHD/GRMPD.
- An SCSI device must be redundant with GDS (for swapping only)

5.2.2.1 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 MPHD/GRMPD, and how to add a path along with the addition of a PCI card, see the “PCI Hot Plug Users’ Guide I/O device”.

5.2.2.2 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 “2. GDS”.

5.2.3 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.

5.2.3.1 The PCI card is only connected to the cluster interconnect

1. Stop the hardware monitoring facility of the machine management.

```
# /usr/sbin/FJSVmadm/prephp
```

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

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

3. 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
```

4. 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
```

5. CE removes the PCI card that is installed on the PCI slot.
6. CE installs a new PCI card on the PCI lost.
7. 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 "PCI Hot Plug Users' Guide".
8. Enable the PCI card on the Solaris OE. Specify the value at step 7 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
```

9. Restart the hardware monitoring facility of the machine management.

```
# /usr/sbin/FJSVmadm/postphp
```

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

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

11. 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)

12. 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
```

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

cf:eventlog CF: Route recovery on /dev/hme9 to node arteco00. (#0000 1 2 2 2)

14. 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
```

5.2.3.2 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. Stop the hardware monitoring facility of the machine management.

```
# /usr/sbin/FJSVmadm/prephp
```

2. 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
```

3. 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
```

4. 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
```

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

6. 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
```

7. 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**

8. Restart the hardware monitoring facility of the machine management.

```
# /usr/sbin/FJSVmadm/postphp
```

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

```
# cfrecon -a /dev/fjgi7

Add Cluster Interconnect:

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

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

    cf:eventlog CF: Route recovery on /dev/fjgi7 to node arteco00. (#0000 1 2 2 2)

11. 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
```

5.2.4 Replacing the cluster interconnect (BLASTBAND)

The procedure for replacing the cluster interconnect varies depending on whether the host channel adaptor (hereafter HCA) is only connected to the cluster interconnect or connected to the cluster interconnect and GLS Public LAN.

5.2.4.1 HCA is only connected to the cluster interconnect

1. Copy down the instance number of the device driver for the HCA that will be removed. You can check the instance number using the "/usr/sbin/hcatoibdl" command described in the "GLSTBAND Documentation".
2. Stop the hardware monitoring facility of the machine management.

   ```
   # /usr/sbin/FJSVmadm/prephp
   ```

3. Check the location of the PCI slot on which the HCA 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/ibdl0  5    4294    1976   UP    YES       00.e0.00.c4.8c.ad
   2     /dev/ibdl2  5    4294    1976   UP    YES       00.e0.00.c4.8c.ad
   # /usr/sbin/FJSVmadm/inst2comp ibdl2
   pcipsy5:PCI#slot00
   # cfgadm -x led=fault,mode=blink pcipsy5:PCI#slot00
   # cfgadm -x led=fault,mode=off pcipsy5:PCI#slot00
   ```

4. Disconnect the HCA device from the cluster interconnect.

   ```
   # cfrecon -d /dev/ibdl2

   Delete Cluster Interconnect:

   device "/dev/ibdl2" successfully removed from node configuration
   # ibsadm refresh
   ```

5. Select the device information display screen from the subnet state display menu (/usr/sbin/ibdsp) on the node where the HCA is replaced. Then, copy down the port GUID of the HCA that is removed. For information on how to refer to the Port GUID, see "3.2.6 Device Information" of the "BLASTBAND GUIDE".
6. Stop the HCA subnet manager. X indicates the instance number of the device driver that is copied at step 1.

   ```
   # /usr/FJSVibmgt/bin/ibmgtstop fjhca X
   ```

7. CE removes the PCI card after removing the HCA at step 3 using PCI Hot Plug.

   ```
   # cfgadm -c disconnect pcipsy5:PCI#slot00
   # cfgadm pcipsy5:PCI#slot00
   Ap_Id                Type       Receptacle   Occupant     Condition
   pcipsy5:PCI#slot00   unknown    disconnected unconfigured unknown
   ```

8. CE installs a new HCA on the PCI slot.
9. Determine the value that indicates the PCI slot location on which the HCA is installed at step 8. See the "ap_id" format described in "1.3.1 Displaying the PCI card" of the "PCI Hot Plug Users' Guide".
10. Add the replaced HCA. Specify the value at step 9 after "cfgadm -c configure". Then, check if "Receptacle" is "connected" and "Occupant" is "configured" respectively.

```
# cfgadm -c configure pcipsy5:PCI#slot00
# cfgadm pcipsy5:PCI#slot00
Ap_Id                  Type         Receptacle   Occupant    Condition
pcipsy5:PCI#slot00     network/hp   connected    configured   ok
```

11. Check if the following LEDs work properly on the HCA port. If they don't, check the cable connection.
    - LinkUp LED : Check lighting and blinking.
    - x4LED : Check blinking.
12. Initialize the replaced HCA executing the following command. X indicates the instance number of the HCA-installed PCI slot at step 8.
    ```
    # /opt/FJSVibhd/bin/fjhcainit php fjhcaX
    ```
13. Start the subnet manager of the replaced HCA executing the following command. X indicates the instance number of the HCA-installed PCI slot at step 8.
    ```
    # /usr/FJSVibmgt/bin/ibmgtstart fjhca X
    ```
14. Delete the setting of the Port GUID at step 5 executing the "/usr/sbin/ibadmin" command to open the Subnet Setting Information Management Menu on the Super SM node. See "3.1.6 Delete port" of the "BLASTBAND GUIDE".
15. Start the Subnet Manager using the subnet setting information management menu (/usr/sbin/ibadmin) on the node where the HCA is replaced. For information on how to start the Subnet Manager, see "3.1.7 Start / stop SM" of the "BLASTBAND GUIDE".
16. Check that the Port GUID of the replaced HCA, not the previously used HCA, is displayed using the subnet state display menu (/usr/sbin/ibdsp) on the node where the HCA is replaced. See "3.2 Subnet state display menu" of the "BLASTBAND GUIDE".
17. Enable the replaced HCA for IBsocket using the following command:
    ```
    # /usr/sbin/ibsadm refresh
    ```
18. Restart the hardware monitoring facility of the machine management that is stopped at step 2.
    ```
    # /usr/sbin/FJSVmadm/postphp
    ```
19. Add the installed HCA in the cluster interconnect using the following command:
    ```
    # cfrecon -a /dev/ibdl2

    Add Cluster Interconnect:

    device "/dev/ibdl2" successfully added to node configuration
    ```
20. Check that the following message is output to the console of each node.
    cf:eventlog CF: Route recovery on /dev/ibdl2 to node tracy. (#0000 2 1 2 2)
21. Confirm that the HCA has successfully been added in the cluster interconnect.
    ```
    # cftool -d
    Number Device    Type Speed   Mtu    State   Configured  Address
    1      /dev/ibdl0  5   4294   1976    UP    YES      00.e0.00.c4.8c.ad
    2      /dev/ibdl2  5   4294   1976    UP    YES      00.e0.00.c4.8c.ad
    ```

5.2.4.2 When the cluster interconnect and GLS public LAN are connected to the HCA

Note:

In this section, the cluster interconnect and GLS public LAN are connected to the HCA. You cannot use a different device name before or after the HCA replacement with PCI Hot Plug. Be sure to install the same PCA on the same PCI slot as before its replacement.

1. Copy down the instance number of the HCA device driver that will be removed. You can check the instance number using the "/usr/sbin/hcatoibdl" command described in the "BLASTBAND GUIDE".

2. Stop the hardware monitoring facility of the machine management.

```
# /usr/sbin/FJSVmadm/prephp
```

3. 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 ibdl3
# /usr/sbin/ifconfig ibdl3 unplumb
# /usr/sbin/ibsadm refresh
```

4. Check the location of the PCI slots on which the HCA 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/ibdl0  5    4294    1976   UP    YES        00.e0.00.c4.8c.ad
2      /dev/ibdl2  5    4294    1976   UP    YES        00.e0.00.c4.8c.ad
# /usr/sbin/FJSVmadm/inst2comp ibdl2
pcipsy5:PCI#slot00
# cfgadm -x led=fault,mode=blink pcipsy5:PCI#slot00
# cfgadm -x led=fault,mode=off pcipsy5:PCI#slot00
```

5. Remove the HCA device from the cluster interconnect.

```
# cfrecon -d /dev/ibdl2

Delete Cluster Interconnect:

device "/dev/ibdl2" successfully removed from node configuration
# /usr/sbin/ibsadm refresh
```

6. Copy down the HCA Port GUID selecting the device information display screen from the state display menu (/usr/sbin/ibdsp) on the node where the HCA is replaced. For the Port GUID, see "3.2.6 Device Information" of the "BLASTBAND GUIDE".

7. Stop the subnet manager of the HCA. X indicates the instance number of the device driver at step 1.

```
# /usr/FJSVibmgt/bin/ibmgtstop fjhca X
```

8. CE removes the PCI card after removing the HCA using PCI Hot Plug at step 4.

```
# cfgadm -c disconnect pcipsy5:PCI#slot00
# cfgadm pcipsy5:PCI#slot00
Ap_Id                Type        Receptacle   Occupant     Condition
pcipsy5:PCI#slot00   unknown     disconnected unconfigured unknown
```

9. CE installs the HCA on the PCI slot where the PCI card is removed at step 8.

10. Add the HCA that is installed at step 9 using PCI Hot Plug. Then, check if "Receptacle" is "connected" and "Occupant" is "configured" respectively.

```
# cfgadm -c configure pcipsy5:PCI#slot00
# cfgadm pcipsy5:PCI#slot00
Ap_Id                Type         Receptacle   Occupant     Condition
pcipsy5:PCI#slot00   network/hp   connected    configured   ok
```

11. Check if the following LEDs work properly on the HCA port. If they don't, check the cable

connection.

- LinkUp LED : Check lighting and blinking.
- x4LED : Check blinking.

12. Format the replaced HCA executing the following command. X indicates the instance number of the HCA-installed PCI slot at step 1.

```
# /opt/FJSVibhd/bin/fjhcainit php fjhcaX
```

13. Start the subnet manager of the replaced HCA executing the following command. X indicates the instance number of the HCA-installed PCI slot at step 1.

```
# /usr/FJSVibmgt/bin/ibmgtstart fjhca X
```

14. Delete the setting of the Port GUID at step 6 executing the setting information management menu (/usr/sbin/ibadmin) on the Super SM node. See “3.1.6 Delete port” of the “BLASTBAND GUIDE”.
15. Start the Subnet Manager using the subnet setting information management menu (/usr/sbin/ibadmin) on the node where the HCA is replaced. For information on how to start the Subnet Manager, see “3.1.7 Start / stop SM” of the “BLASTBAND GUIDE”.
16. Check that the Port GUID of the replaced HCA, not the previously used HCA, is displayed using the subnet state display menu (/usr/sbin/ibdsp) on the node where the HCA is replaced. See “3.2 Subnet state display menu” of the “BLASTBAND GUIDE”
17. Enable the replaced HCA for Ibsocket using the following command:

```
# /usr/sbin/ibsadm refresh
```

18. Add the replaced HCA in the GLS setting using the following command. Change the interface name according to the GLS settings.
    - Fast switching mode

```
# /usr/sbin/ifconfig ibdl3 plumb
# /usr/sbin/ifconfig ibdl3 192.168.4.1 netmask + broadcast + -
trailers up
# /opt/FJSVhanet/usr/sbin/hanetnic add -n sha0 -i ibdl3
```

19. Restart the hardware monitoring facility of the machine management that is stopped at step2.

```
# /usr/sbin/FJSVmadm/postphp
```

20. Add the installed HCA to the cluster interconnect using the following command:

```
# cfrecon -a /dev/ibdl2

Add Cluster Interconnect:

device "/dev/ibdl2" successfully added to node configuration
```

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

    cf:eventlog CF: Route recovery on /dev/ibdl2 to node tracy. (#0000 2 1 2 2)

22. Confirm that the HCA has successfully been added in the cluster interconnect.

```
# cftool -d

Number Device    Type  Speed    Mtu    State Configured  Address
1      /dev/ibdl0  5    4294     1976   UP    YES      00.e0.00.c4.8c.ad
2      /dev/ibdl2  5    4294     1976   UP    YES      00.e0.00.c4.8c.ad
```