TruCluster Server

Cluster Release Notes

Part Number: AA-RHH1E-TE

September 2002

Product Version: TruCluster Server Version 5.1B

Operating System and Version: Tru64 UNIX Version 5.1B

This manual provides important information about TruCluster Server Version 5.1B.

For information that may become available after the publication of these release notes, go to the following URL:

http://www.tru64unix.compaq.com/docs/pub_page/tcr_update.html

Contents

About This Manual

1 Features

2 Recommendations, Issues, and Known Problems

Index

Tables

About This Manual

This manual provides important information about HP TruCluster Server
Version 5.1B, which runs on the HP Tru64 UNIX Version 5.1B operating
system.

Audience

This manual is intended for TruCluster Server users, administrators, and
programmers.

Organization

This manual is organized as follows:

Chapter 1	Lists new features in TruCluster Server Version 5.1B.
Chapter 2	Discusses recommendations, issues, and known problems with TruCluster Server Version 5.1B, including updates to manuals since their publication.

Related Documents

Consult the following manuals for assistance in cluster hardware and
software configuration, installation, administration, and programming
tasks:

- TruCluster Server *QuickSpecs* — Describes TruCluster Server Version
  5.1B, including information about its capabilities and the hardware
  it supports. You can find the latest version of the *QuickSpecs* at the
  following URL:

  http://www.tru64unix.compaq.com/docs/pub_page/spds.html

- *Cluster Technical Overview* — Provides an overview of the TruCluster
  Server technology.

- *Cluster Hardware Configuration* — Describes how to set up the systems
  that are to become cluster members, and how to configure cluster shared
  storage.

- *Cluster Installation* — Describes how to install the TruCluster Server
  software.

- *Cluster Highly Available Applications* — Describes how to deploy existing applications on a TruCluster Server cluster, and how to write cluster-aware applications.

- *Cluster Administration* — Describes cluster-specific administration tasks.

You can find the latest version of the TruCluster Server documentation at:

`http://www.tru64unix.compaq.com/docs/pub_page/doc_list.html`

For information that may become available after the publication of these release notes, go to the following URL:

`http://www.tru64unix.compaq.com/docs/pub_page/tcr_up-`
`date.html`

_____ **Note** _____

> Read the Tru64 UNIX *Release Notes* and the Tru64 UNIX *System Administration* manual to become familiar with restrictions and new features in the base operating system before you install, configure, and use your TruCluster Server cluster.

_____

Icons on Tru64 UNIX Printed Manuals

The printed version of the Tru64 UNIX documentation uses letter icons on the spines of the manuals to help specific audiences quickly find the manuals that meet their needs. (You can order the printed documentation from HP.) The following list describes this convention:

G   Manuals for general users

S   Manuals for system and network administrators

P   Manuals for programmers

R   Manuals for reference page users

Some manuals in the documentation help meet the needs of several audiences. For example, the information in some system manuals is also used by programmers. Keep this in mind when searching for information on specific topics.

The *Documentation Overview* provides information on all of the manuals in the Tru64 UNIX documentation set.

Reader's Comments

HP welcomes any comments and suggestions you have on this and other Tru64 UNIX manuals.

You can send your comments in the following ways:

- Fax: 603-884-0120 Attn: UBPG Publications, ZKO3-3/Y32

- Internet electronic mail: `readers_comment@zk3.dec.com`

  A Reader's Comment form is located on your system in the following location:

  `/usr/doc/readers_comment.txt`

Please include the following information along with your comments:

- The full title of the manual and the order number. (The order number appears on the title page of printed and PDF versions of a manual.)

- The section numbers and page numbers of the information on which you are commenting.

- The version of Tru64 UNIX that you are using.

- If known, the type of processor that is running the Tru64 UNIX software.

The Tru64 UNIX Publications group cannot respond to system problems or technical support inquiries. Please address technical questions to your local system vendor or to the appropriate HP technical support office. Information provided with the software media explains how to send problem reports to HP.

Conventions

The following typographical conventions are used in this manual:

`%` `$`	A percent sign represents the C shell system prompt. A dollar sign represents the system prompt for the Bourne, Korn, and POSIX shells.
`#`	A number sign represents the superuser prompt.
`% `**`cat`**	Boldface type in interactive examples indicates typed user input.
`file`	Italic (slanted) type indicates variable values, placeholders, and function argument names.

[|] { | }	In syntax definitions, brackets indicate items that are optional and braces indicate items that are required. Vertical bars separating items inside brackets or braces indicate that you choose one item from among those listed.
. . .	In syntax definitions, a horizontal ellipsis indicates that the preceding item can be repeated one or more times.
cat(1)	A cross-reference to a reference page includes the appropriate section number in parentheses. For example, cat(1) indicates that you can find information on the cat command in Section 1 of the reference pages.
Return	In an example, a key name enclosed in a box indicates that you press that key.
Ctrl/*x*	This symbol indicates that you hold down the first named key while pressing the key or mouse button that follows the slash. In examples, this key combination is enclosed in a box (for example, Ctrl/C).

1

Features

TruCluster Server Version 5.1B supports all of the features of TruCluster
Server Version 5.1A plus additional enhancements that are listed in
Table 1–1.

Table 1–1: TruCluster Server Version 5.1B Features

Feature	Description	For More Information
Cluster File System (CFS) load balancing	The CFS load balancing daemon (`cfsd`) can monitor and analyze file-system usage, make recommendations, and automatically relocate file systems.	*Cluster Administration*, cfsd.conf(4), and cfsd(8)
Parallel rolling upgrade	This feature reduces the amount of time required for a rolling upgrade. After the initial member has been updated, an administrator can upgrade the remaining cluster members in parallel. The number of members that can be upgraded simultaneously is limited by the quorum configuration of the cluster.	*Cluster Installation* and clu_upgrade(8)
`mount -o server=node`	The `mount -o server=node` command lets an administrator specify a cluster member as the initial server for a file system. (The `cfsd` daemon uses this option to implement load-directed mounting.)	*Cluster Administration*, mount(2), and mount(8)
Forced unmounting of AdvFS file systems `cfsmgr -U`	The `cfsmgr -U` command lets an administrator forcibly unmount an entire AdvFS domain that is being served by a cluster member.	*Cluster Administration* and cfsmgr(8)
CFS quiesce `freezefs` `thawfs`	To facilitate snapshots and cloning of multivolume domains, the `freezefs` command places an AdvFS domain into a metadata-consistent frozen state and guarantees that it stays that way until thawed. The file system thaws either by timing out or explicitly with the `thawfs` command.	*Cluster Administration*, freezefs(8), and thawfs(8)

Table 1–1: TruCluster Server Version 5.1B Features (cont.)

Feature	Description	For More Information
Relaxed `gated` requirement	Prior releases of TruCluster Server required that all cluster members run the `gated` routing daemon. Version 5.1B relaxes that restriction to allow a cluster member to run `gated`, run `routed`, or use static routing. When a cluster member runs `gated`, the administrator can control whether `gated` routing entries are manipulated by the cluster alias daemon, `aliasd`.	*Cluster Technical Overview* and *Cluster Administration*
Balancing CAA application resources `caa_balance`	The `caa_balance` command causes CAA to reevaluate application resource placement based on the current state of the resources on the cluster and the rules of placement for the resources. Balancing applications can be done on a cluster-wide basis, a member-wide basis, or with specified resources.	*Cluster Highly Available Applications* and `caa_balance`(8)
CAA availability measurement `caa_report`	CAA maintains a history of each application resource from the time it is first started. The `caa_report` command displays a report summarizing the percentage of time that all currently registered application resources have been in the ONLINE state. The data is obtained by analyzing the Event Manager events related to CAA.	*Cluster Administration* and `caa_report`(1)
CAA user-defined attributes	A user-defined attribute can be placed within a resource profile and accessed from a resource action script as an environment variable.	*Cluster Highly Available Applications*
CAA reason codes and environment locales	A reason code describes why an action script is being executed. It is available to an action script through an environment variable.\n\nThe environment locale where an action script is executed is now available to an action script through an environment variable.	*Cluster Highly Available Applications*
Directing output from CAA action scripts	You can redirect output from an action script so that it is displayed when CAA commands are executed.	*Cluster Highly Available Applications*

Table 1–1: TruCluster Server Version 5.1B Features (cont.)

Feature	Description	For More Information
cluamgr rpri=100	The value of 100 is a special case for the cluamgr rpri (router priority) option. For rpri values of 1-99, aliasd advertises a host route cost metric of 14. For an rpri value of 100, aliasd advertises a host route cost metric of 10.	cluamgr(8)
clua_active	The clua_active command determines whether the specified cluster alias is active for the current member, has a selection weight greater than 0 (zero), and is reachable through an external network interface.	clua_active(8)

2

Recommendations, Issues, and Known Problems

The following sections describe recommendations, issues, and problems with TruCluster Server Version 5.1B.

2.1 Hardware Configuration

The information in this section applies to configuring the hardware that will become a cluster.

2.1.1 Memory Channel Configuration Must Be Symmetrical

In a cluster that uses Memory Channel for its cluster interconnect, each cluster member must have the same number of Memory Channel cards. For example, you cannot configure one member with two Memory Channel cards and another member with only one Memory Channel card. The Memory Channel software depends on the cluster having a symmetrical Memory Channel configuration. See the *Cluster Technical Overview* and the *Cluster Hardware Configuration* manuals for more information about Memory Channel and supported configurations.

2.2 Installation

The information in this section applies to installation.

2.2.1 Update to Latest Firmware Before Installing Version 5.1B

Before installing Tru64 UNIX and TruCluster Server Version 5.1B, update all systems that will become cluster members with the latest firmware. A cluster member running old firmware may not be able to use some hardware connected to the cluster. For example, with old firmware, a member with a boot disk behind an HSZ80 or HSG80 controller may fail to boot, indicating "Reservation Conflict" errors.

To update a system's firmware, do the following:

1. Insert the firmware CD-ROM in the drive and boot from it:

    ```
    >>> boot cdrom_console_device_name
    ```

The firmware update utility automatically identifies your system type and model and determines the correct firmware revision required for your system.

2. Follow the instructions on the screen. The READ-ME-FIRST file, which describes the firmware changes included in the update, is displayed automatically.

3. When the firmware update is complete, power off the processor for at least 10 seconds to initialize the new firmware.

If you do not have access to a firmware CD-ROM, you can find the latest firmware at the following URL:

ftp.digital.com/pub/DEC/Alpha/firmware/readme.html

You can download the firmware and associated documentation with the anonymous File Transfer Protocol (FTP).

2.3 Cluster Creation and Member Addition

Information in this section applies to creating a cluster and adding cluster members.

2.3.1 Need Space in Base /usr to Build a Cluster Kernel

The `clu_create` command builds the kernel for the first cluster member on the base system in /usr. If there is not enough space in /usr, the `doconfig` command fails. This is not a fatal error. You can either free up disk space and rerun `clu_create`, or boot the clusterized `genvmunix` from the first member's cluster boot disk and then build a customized kernel for the member.

The preventive workaround is to make sure that you have enough space in /usr to build a kernel before running `clu_create`.

2.3.2 NIS Master and Cluster Creation

When a Tru64 UNIX system is configured as a NIS master before cluster creation, the `clu_create` command should configure the cluster as a NIS master. However, when the single-member cluster is booted, all `yp` maps except the `ypservers` map are still served from the base member's host name, not from the default cluster alias name. To fix this problem, run the `make` command to remake the maps so that the NIS service can use the failover mechanism available to the cluster alias. Use the following procedure to remake the maps:

1. Change directory to /var/yp/src:

```
            # cd /var/yp/src
```

2. Run the `touch` command to update the dates on the files:

    ```
    # touch *
    ```

3. Change directory to `/var/yp`:

    ```
    # cd ..
    ```

4. Run the `make` command to remake the databases:

    ```
    # make
    ```

2.3.3 Domain Issues When Re-Creating a Cluster with Different Disks

When you create a cluster, `clu_create` creates AdvFS domains in
`/etc/fdmns` for `cluster_root`, `cluster_usr`, and `cluster_var`, and for
the first member's boot domain (usually `root1_domain`). These domains
are in the `/etc/fdmns` directory on the base Tru64 UNIX operating system.

To re-create a cluster, you boot the Tru64 UNIX operating system to
multi-user mode and rerun `clu_create`. This works as long as you use the
same disks for the same cluster domains as you did when you first created
the cluster. However, if you specify different disks, `clu_create` indicates
it will remove the existing domains from `/etc/fdmns`, but does not do so.
For example:

```
 The following AdvFS boot domain is already configured:

 root1_domain dsk6a

 Do you want to reuse the disk associated with this AdvFS domain
 as the boot disk? [yes]: no

 The installation must remove this AdvFS domain in order to continue.
 Do you want to remove this domain? [yes]: yes

 Each member has its own boot disk, which has an associated
 device name; for example, 'dsk5'.

 Enter the device name of the member boot disk []:dsk17
 Checking the member boot disk: dsk17.
         .
         .
         .
Creating AdvFS domains:
   Creating AdvFS domain 'root1_domain#root' on partition '/dev/disk/dsk17a'.
 mkfdmn: domain '/etc/fdmns/root1_domain' already exists
 mkfdmn: can't create new domain 'root1_domain'

 *** Error ***
  Cannot create AdvFS domain '/dev/disk/dsk17a' on disk 'root1_domain'.

 *** Error ***
  clu_create: Failed creating AdvFS domains.

 *** Error ***
  clu_create: Failed to create a cluster.
```

The preventive workaround is to remove domains that will not use the same disks from `/etc/fdmns` before running `clu_create` to re-create the cluster. For example:

```
# rm -rf /etc/fdmns/root1_domain
```

2.4 Rolling Upgrade

This section discusses issues with rolling upgrade.

2.4.1 vfast Boot Error Message on Non-Rolled Members

When performing a rolling upgrade to Version 5.1B, after the lead member has been rolled, if you boot a member that has not been rolled, you will see a `vfast` error message containing the following string:

```
usr/sbin/vfast: /sbin/loader: Fatal Error: call to unresolved symbol \
from /usr/sbin/vfast
```

The problem occurs because the `vfast` startup script is installed in `/sbin/init.d` and, although `vfast` is not supported on Version 5.1A, any member that reboots will attempt to start `vfast`. Fortunately, the attempt fails and the error message is benign. This problem does not occur on a member that has rolled to Version 5.1B.

2.4.2 Rolling Upgrade, Secure Shell, Securing Remote Utilities, and Fully Qualified Domain Names

This note applies only when all of the following is true:

- Your cluster is running Version 1.0 of the Secure Shell (SSH).

- You have the `EnforceSecureRutils` variable set to `yes` in the `/etc/ssh2/ssh_config` configuration file

- You did not use a fully qualified domain name for the cluster. Note that both the Secure Shell documentation and `clu_create` recommend that you use fully qualified domain names. If you followed those guidelines, this note does not apply to your cluster.

When you update the cluster to Version 5.1B, the Secure Shell subset is a mandatory subset that, when `EnforceSecureRutils` is set to `yes`, requires the use of host-based authentication with a fully qualified domain name for the cluster. If the cluster does not use a fully qualified domain name, the traditional remote utilities (such as `rcp` and `rsh`) will fail authentication checks.

The workaround is to add the fully qualified cluster hostname to `/etc/hosts`, `/.rhosts` (if using), and `/.shosts`. Then create a symbolic link, using the fully qualified name, in `/etc/ssh2/knownhosts`. For

example, using a cluster name whose name is `deli` the following example shows the before and after view for each file, and the command used to create the link:

Before:

```
/etc/hosts:    16.140.160.124  deli
/.rhosts:      deli
/.shosts:      deli
```

After:

```
/etc/hosts:    16.140.160.124  deli.zk3.dec.com  deli
/.rhosts:      deli.zk3.dec.com
/.shosts:      deli.zk3.dec.com
```

```
# ln -sf /etc/ssh2/hostkey.pub \
  /etc/ssh2/knownhosts/deli.zk3.dec.com.ssh-dss.pub
```

2.4.3 If i18n Is in /usr, clu_upgrade Does Not Calculate Required Disk Space Correctly

The Worldwide Language Support (WLS) subsets are installed in `/usr/i18n`, which can either be part of `/usr` or a separate file system. When the WLS subsets are not installed in a separate file system, the `clu_upgrade` command does not calculate required disk space correctly during its `setup` stage. This miscalculation can lead to insufficient space in the file system that contains the WLS subsets when `clu_upgrade` creates tagged files during the `setup` stage. For example:

```
# clu_upgrade setup 1
 .
 .
 .
Checking inventory and available disk space.
Copying cluster kit 'xxx' to 'yyy'

Creating tagged files.
 .......................................................
 ......................NOTE: CFS: File system full: /usr

 NOTE: CFS: File system full: /usr

 NOTE: CFS: File system full: /usr
 NOTE: CFS: File system full: /usr
 .
 .
 .
[followed by a multitude of similar error messages]

*** Warning ***
 The above errors were detected during the cluster upgrade. If you believe that
 the errors are not critical to system operation, you can choose to continue.
 If you are unsure, you should check the cluster upgrade log and refer
 to clu_upgrade(8) before continuing with the upgrade.

 Do you want to continue the cluster upgrade? [no]:
```

If you find yourself in this situation:

1. Answer `no` to the prompt.

2. Run the `clu_upgrade undo setup` command to remove the tagged files and undo the `setup` stage.

3. Increase file system free space to the values specified in the *Cluster Installation* manual.

4. Run the `clu_upgrade setup` command again.

The preventive workaround is to make sure that file systems have the required amounts of available free space before running the `clu_upgrade setup` command.

2.4.4 clu_upgrade undo install: Error Message for /usr/.smdb./.wwinstall

When the Worldwide Language Support (WLS) subsets are installed on a cluster and a problem occurs with a rolling upgrade that requires undoing the `install` stage, the following message will be displayed:

```
Restoring tagged files.
       .........Cannot rename /usr/.smdb./.RollTemp...wwinstall \
       to  /usr/.smdb./.wwinstall

       /usr/.smdb./.RollTemp...wwinstall is not a tagged file
```

You can ignore this warning message because the `.wwinstall` directory is not an inventory item.

2.4.5 clu_upgrade undo install: Error Message for /etc/.Old..ifaccess.conf

When rolling to Version 5.1B, the following error message is encountered only if the rolling upgrade had progressed past the roll stage and is then undone back to the install stage:

```
clu_rollprop: /etc/.Old..ifaccess.conf does not exist
```

If you see this error message while undoing the install stage, you can safely ignore the message. Although the `/etc/.Old..ifaccess.conf` file does not exist, the real per-member `/etc/ifaccess.conf` files are still there.

2.4.6 clu_upgrade undo install Does Not Remove WLS Entries from /etc/motd

When undoing a rolling upgrade on a cluster with the Worldwide Language Support (WLS) subsets installed, WLS entries for both Tru64 UNIX Version 5.1A and Version 5.1B are left in the `/etc/motd` file. For example:

```
Tru64 UNIX Catalan Support V5.1B (rev. 231)
Tru64 UNIX Catalan Support V5.1A (rev. 168)
```

Edit the `/etc/motd` file manually to delete the WLS entries for Tru64
UNIX Version 5.1B.

2.4.7 Rolling Upgrade and Retired Hardware

Before performing a rolling upgrade to Version 5.1B, read the Hardware
Support Retirement Notices in the Tru64 UNIX Version 5.1B *Release Notes*
and verify that your cluster's hardware is supported in Version 5.1B.

2.5 Booting and Shutdown

This section discusses requirements and restrictions for booting members
into a cluster and for shutting down cluster members.

2.5.1 A Member May Hang on Boot in a Cluster Using a KZPBA-CB SCSI Bus Adapter

When you boot a member of a cluster with a KZPBA-CB SCSI bus adapter,
the member may hang during the boot. The console log for that member will
display messages similar to the following:

```
cam_logger: SCSI event packet
cam_logger: bus 10 target 15 lun 0
ss_perform_timeout
timeout on disconnected request
Active CCB at time of error
cam_logger: SCSI event packet
cam_logger: bus 10 target 15 lun 0
isp_process_abort_queue
IO abort failure (mailbox status 0x0), chip reinit scheduled
Active CCB at time of error
cam_logger: SCSI event packet
cam_logger: bus 10
isp_reinit
Beginning Adapter/Chip reinitialization (0x3)
cam_logger: SCSI event packet
cam_logger: bus 10
isp_reinit
Fatal reinit error 1: Unable to bring Qlogic chip back online
```

If you see messages like this, you must reset the system and then boot
it again.

On an AlphaServer GS80, GS160, or GS320, you can perform a system
control manager (SCM) `halt` and `reset`. On other systems, you will have to
do a hardware reset before booting again.

If the reset fails to solve the problem, you must cycle the power. On an AlphaServer GS80, GS160, or GS320, use the SCM `power off` and `power on` commands.

2.5.2 CNX Panic During Boot

When you boot a member in a cluster with a large storage configuration, the member may panic and display the following message:

```
CNX MGR: Invalid configuration for cluster seq disk
```

If this occurs, reboot the member.

2.5.3 Panic: clubase_cfg: no cluster name in /etc/sysconfigtab

This panic has been seen occasionally on clusters doing reboots under load. This is a console firmware problem. There is no workaround; boot again.

2.5.4 Panic: cb_open: failed SCSI

This panic ("cb_open: failed SCSI" followed by device information) has been seen occasionally on clusters doing reboots under load. This is a console firmware problem. There is no workaround; boot again.

2.5.5 Interactively Booting a Non-Voting Cluster Member

By design, a non-voting member cannot form a cluster. This note describes how to boot a non-voting member and form a cluster when that member is the only surviving member of the cluster. The following example sets both `cluster_expected_votes=1` and `cluster_node_votes=1` so that the member can boot and form a cluster regardless of current cluster membership or quorum vote configuration.

```
>>> boot -fl "ia"
  .
  .
  .
    Enter kernel_name [option_1 ... option_n]
     Press Return to boot default kernel
     'vmunix': vmunix clubase:cluster_node_votes=1 \
             clubase:cluster_expected_votes=1 [Return]
```

When you resume the boot, the member can form a cluster. When the member reaches multiuser level, log in and use the `clu_quorum` command to give the member a vote. (The vote used for the interactive boot is not written to the member's `sysconfigtab` file.) Then use `clu_quorum` to inspect all cluster quorum settings and adjust as needed. See the *Cluster Administration* manual for information on quorum vote configuration. In addition, the *Cluster Administration* manual has a section, Forming a Cluster When Members Do Not Have Enough Votes to Boot and Form a

Cluster, that provides additional information on how to interactively adjust quorum values when booting a cluster member.

2.6 File System

This section discusses issues with the CFS, AdvFS, and NFS file systems in a cluster.

2.6.1 The newfs Command Incorrectly Allows Creation of UFS File Systems on a Quorum Disk

Do not use the quorum disk for user data. The `mkfdmn` command prevents you from creating an AdvFS domain on a quorum disk. However, the `newfs` command incorrectly allows you to create a file system on a quorum disk. Do not use `newfs` to create a file system on a quorum disk.

2.6.2 CFS Relocation Failures Involving Applications That Wire Memory

Applications that use the `plock( )` or `mlock( )` system call to lock pages of physical memory can cause the `cfsmgr` command to fail when performing a manual relocation.

If the application uses `plock( )`, the domain or file system that contains the application executable cannot relocate. In the case of `mlock( )`, if the locked pages are associated with files, the file systems where those files reside cannot relocate.

In the event of failure, the `cfsmgr` command returns the following message:

```
Server Relocation Failed
Failure Reason: Invalid Relocation
```

To allow the relocation to complete for the domain or file system on which the executables reside, kill the processes that are running the executables using the `plock( )` and `mlock( )` system calls. Find out whether `collect` is running. If it is, kill `collect` and restart it with the `-l` (do not lock pages into memory) option.

2.6.3 The cfsstat directio Command Displays Incorrect Value for Fragment Writes on the Server

The `cfsstat` command returns different values for `fragment writes` depending on whether the command is run on the cluster member serving the file system or on a cluster member that is a client for the file system.

The value of `fragment writes` increments properly on the client but not on the server. For example:

On the server:

```
# /usr/bin/cfsstat directio

 Concurrent Directio Stats:
             0 direct i/o reads
           280 direct i/o writes
             0 aio raw reads
             0 aio raw writes
             0 unaligned block reads
             0 fragment reads
             0 zero-fill (hole) reads
           280 file-extending writes
             0 unaligned block  writes
             0 hole writes
             0 fragment writes
             0 truncates
```

On the client:

```
# /usr/bin/cfsstat directio

 Concurrent Directio Stats:
          1569 direct i/o reads
           240 direct i/o writes
             0 aio raw reads
             0 aio raw writes
             0 unaligned block reads
            37 fragment reads
             3 zero-fill (hole) reads
           230 file-extending writes
             0 unaligned block  writes
             0 hole writes
            10 fragment writes
             0 truncates
```

2.6.4 The freezefs -q Command Returns Incorrect Results for Non-AdvFS File Systems

AdvFS is the only file system type that is valid for `freezefs` command operations. An attempt to freeze any other type of file system, such as NFS, UFS, or `/proc`, will result in the following error:

```
ENOTSUP : Function not implemented
```

Users can learn to whether a file system is currently frozen by using the `freezefs` command's `-q` option or the `freezefs` system call's `FS_QUERY` flag. The `-q` option works correctly for AdvFS file systems. However, if you use `freezefs -q` to query whether a non-AdvFS file system is frozen, the

`freezefs` command will erroneously display a message indicating that the file system is frozen instead of displaying an error message.

2.6.5 Do Not Use the chfile -L on Command or the mount -o adl Command

In a cluster, do not use the `chfile -L on` *filename* command or the `mount -o adl` *file-system* command. Because the cluster environment does not correctly enforce the exclusion of certain types of data logging when a file is mapped with the `mmap()` system call, a system crash might result in inconsistent data being written to disk.

2.7 LSM

This section discusses problems with using the Logical Storage Manager (LSM) in a TruCluster Server cluster.

2.7.1 Problem Encapsulating Swap in Clusters with Long Host Names

LSM has a problem encapsulating swap in a cluster on members with base host names greater than 24 characters, for example, `reallyreallyreallyverylonghostname.foo.bar.com`. To work around this problem, reduce the base hostname to fewer than 25 characters.

2.7.2 Caution on Using volmigrate or volunmigrate Command on the cluster_usr Domain

Do not reboot any member of a cluster while an LSM `volmigrate` or `volunmigrate` operation on the AdvFS `cluster_usr` domain is in process.

Rebooting a node before the `volmigrate` or `volunmigrate` operation completes can result in the entire cluster hanging. This problem occurs only with the `cluster_usr` domain.

You do not need to reboot the cluster or any cluster member after migrating domains to or from LSM volumes.

2.8 CAA

This section discusses problems in the cluster application availability (CAA) subsystem.

2.8.1 Problems with caa_relocate When Multiple Interdependent Applications Are Specified

The `caa_relocate -s` command fails when it is used with resources with dependencies. To avoid this, do the following:

1. Identify applications with dependencies by using the `caa_stat -p` command. Those applications with entries for `REQUIRED_RESOURCES` have dependencies.

2. Use the `caa_relocate -f` command to relocate the applications with dependencies.

3. Use `caa_relocate -s member1 -c member2` to relocate the other applications.

2.8.2 SysMan Menu CAA Management on a Terminal Is Missing Navigation Buttons

When running the CAA Management branch of the SysMan Menu in a terminal screen, the OK, CANCEL, and HELP selections may not be visible in the details window if your screen is only 24 lines. To move from this screen, press Ctrl/c.

2.8.3 Updating the Registration of an Application Resource with No Balance Data Produces an Error

Updating the registration of an application resource (`caa_register -u` *resource*) with no data in the Balance field of the application profile will display an error message similar to the following one:

```
REBALANCE entry(ies) will be removed from clustercron
  Error when calling system (/var/cluster/caa/bin/caa_schedule \
  UNREGISTER <application>)
```

You can safely ignore these messages.

2.8.4 CAA Events Are Malformed When Viewed from the Event Manager Viewer

The Event Manager viewer may display malformed CAA event messages, or messages with missing information. For example, the message:

```
CAA named is transitioning from state ONLINE to state OFFLINE on skiing
```

is displayed as:

```
CAA named is transitioning from state to state skiing
```

To work around this problem, examine the messages in the `daemon.log` file for more complete information. The messages in the `daemon` log file are in a slightly different format from those that the Event Manager viewer displays.

2.8.5 SysMan Station Shows CAA Application Resources in UNKNOWN State as Having an Error

The SysMan Station shows any application resources that are in the `UNKNOWN` state as having an error and does not show what member the application is `UNKNOWN` on. For example, an application resource named `xyz` in the `UNKNOWN` state is shown underneath the cluster icon as `xyz (error)`.

2.9 Cluster Alias

This section discusses problems in the cluster alias subsystem.

2.9.1 Cluster Alias Does Not Appear to Evenly Distribute SSH Requests

Incoming connection requests addressed to a cluster alias are distributed among members of that alias according to the selection weight (`selw`) assigned each member of the alias. The default cluster alias has a `selw` of 3. However, because the Secure Shell (SSH) uses two connections to establish a single connection, connection requests to the default cluster alias that use SSH will be distributed such that some systems get two connections while some systems get one connection. There is no real distribution imbalance; connection requests are still distributed according to the selection weight assigned to cluster members.

The Cluster Alias chapter in the *Cluster Technical Overview* describes how incoming TCP connection requests and UDP packets are distributed among the members of a cluster alias.

2.10 Miscellaneous Administration

This section discusses issues with various administration tools that are used in a cluster.

2.10.1 RIS Boot Failures When Cluster Is RIS Server

If the system that became the initial cluster member was configured as a RIS server before the `clu_create` command was run, then the cluster creation process does not update the `sa` entry in `/etc/bootptab`. The `sa` entry remains the IP address of the standalone system. Because of this, attempts at RIS boots after clusterization fail to mount the root file system.

You must manually edit /etc/bootptab and update the sa entry to be the IP address for the default cluster alias.

2.10.2 The hwmgr -show comp Command May Report an Inconsistency Error When Creating a Clusterwide Name for a SCSI Device

After you have used the hwmgr -edit scsi command to create a clusterwide unique name for a SCSI device, a subsequent hwmgr -show comp command may report an inconsistency on the SCSI device. The inconsistency appears when the hwmgr -edit scsi command is invoked on the second and subsequent members for the same device. You can ignore the inconsistency error in this situation.

For example:

```
root> hwmgr -show comp -id 373 -full

HWID:  HOSTNAME   FLAGS  SERVICE  COMPONENT NAME
-----------------------------------------------------------------
373:   rovel-qa1  rcd-i  iomap    SCSI-WWID:ff10000b:"media_chngr"

DSF GROUP
INSTANCE GRPFLAGS GROUPID SUBSYSTEM   BASENAME  L1            L2
-----------------------------------------------------------------
0        40       81      cam_changer mc2       media_changer generic

DEVICE NODE
ID  LBdevT  LCdevT   CBdevT  CCdevT  BFlags CFlags Class Suffix L3B    L3
-------------------------------------------------------------------------
0   0       56008c0  0       13003b3 0x0    0x861  0x0   (null) (null) (null)

COMPONENT INCONSISTENCY
-----------------------
Component should not have an entry in the cluster database but it does.
```

2.10.3 Running Process Accounting on Large Clusters Can Exhaust Member Process Quotas

If process accounting is enabled on large clusters (six to eight members), cluster members may start swapping heavily and eventually exhaust their process quotas. A ps command on a member will show tens of thousands of icssvr_daemon_from_pool processes.

If you see this situation developing in a cluster that is running process accounting, use the accton command with no parameters to disable accounting.

2.10.4 When Manually Editing a File, End the File with a newline

Several cluster commands append information to existing system administration files whose format is one entry per line. If you manually add

an entry to the end of a file of this type, and do not also add a newline, the append by the command will concatenate to the current entry rather than place the new entry on its own line.

For example, assume you add the entry `somehostname root` to the end of the `/.rhosts` file but do not insert a newline. When you then run `clu_create`, the name of the new cluster (in this example, `deli.zk3.dec.com`) is appended to the file. The resulting entry will look like this:

```
somehostname rootdeli.zk3.dec.com
```

For this reason, when manually editing a file whose format is one entry per line, make sure that a newline is the last character in the file.

2.10.5 Cluster Hardware Operations That Require Backup

TruCluster Server systems maintain cluster-wide and member-specific hardware databases. The databases are synchronized between the cluster root file system and the member boot disks, with the databases maintained on the cluster root accepted as the most up-to-date source in the case of a discrepancy.

Certain hardware operations generate entries in the hardware databases. After you perform one or more of these operations, you must back up the cluster root file system and the member boot partitions so that they correctly reflect the hardware status of the cluster.

This is particularly important if the cluster root file system later becomes corrupted or unavailable due to a hardware problem and you must restore it from backup. The backup **must** reflect the disk storage environment known to the cluster root file system at the time it failed. If the backup of the cluster root file system does not reflect the current disk storage environment, the cluster root recovery procedures described in the *Cluster Administration* manual cause a panic and you must use the `clu_create` command to re-create the cluster.

If you perform any of the following actions, back up the cluster root file system and the member boot partitions:

- Delete a component via the `hwmgr -delete` command.
- Refresh the component database via the `hwmgr -refresh comp` command.
- Redirect a SCSI device via the `hwmgr -redirect scsi` command.

See the *Cluster Administration* manual for information on backing up and repairing a member's boot disk.

2.10.6 Configuring a Cluster Member as an IP Router

This note provides supplementary information for the Running IP Routers section in the *Cluster Administration* manual.

When configuring a cluster member as an IP router, use `gated`. However the `gated` daemon cannot be under `aliasd` control. Otherwise, `aliasd` will turn `gated` on and off, and point `gated` to the `/etc/gated.conf.member`*n* configuration file generated by `aliasd` rather than to your customized `/etc/gated` file. The *Cluster Administration* manual describes how to use the `cluamgr` command's `nogated` option, and how to set `CLUAMGR_ROUTE_ARGS=nogated` for that member's `/etc/rc.config` file.

We recommend that you do not configure a cluster member as a general purpose IP router unless you are an experienced network administrator.

We do not recommend using `routed` or static routing on a cluster member that will be an IP router.

2.11 Programming

This section discusses problems with cluster-specific application programming interfaces (APIs).

2.11.1 MC API Applications May Not Use Transfers Larger Than 8 KB with Loopback Mode Enabled on Clusters Utilizing Virtual Hubs

This note applies only when all of the following are true:

- The cluster has a Memory Channel cluster interconnect.
- The cluster has a virtual hub.
- An application has used the Memory Channel Application Programming Interface (MC API) to enable loopback mode (via a call to the MC API `imc_asattach()` function with the `IMC_LOOPBACK` flag).

In this situation, make sure that applications that use the MC API limit the size of Memory Channel transfers to a maximum of 8 KB blocks. Larger block sizes can result in a hard hang of a cluster member, which you must then power down and reboot to recover.

The following function, `imc_bcopy_safe()`, avoids this problem by copying data over Memory Channel in blocks no larger than 8 KB and by implementing a kind of flow control between the members. In place of calling the MC API `imc_bcopy()` function, applications can call the `imc_bcopy_safe()` function. You can use the following example as is, or as a coding guideline:

```
/*
 * This function is a workaround to a HW problem when MC VHUBs
 * are used on certain platforms.
 *
 * Problem: If too much data is sent over MC on a page that
 *          is set up for LOOPBACK, the remote node may
 *          get stuck on the PCI bus, because of an
 *          MC adapter buffer overflow problem combined
 *          with a PCI interface chip problem.
 *
 * Solution: Only copy 8KB of data at a time over MC. In
 *           between these calls, use the function imc_ckerrcnt_mr()
 *           to provide a sort of flow control to the other node.
 *
 * Usage: Use this function instead of direct calls to
 *        imc_bcopy().
 *
 * Note:  The use of this function has some
 *        performance implications for large amounts of data.
 */

#include <stdio.h>
#include <unistd.h>
#include <sys/imc.h>

long imc_bcopy_safe( void *, void *, long , long , long );

#define MCCOPY_SIZE (8*1024)    /* Max copy size to avoid system hangs! */
#define LOGICAL_RAIL  0

long imc_bcopy_safe(
          void *src,
          void *dest,
          long length,
          long dest_write_only,
          long first_dest_quad)
{
  char *source;
  char *tx_addr;
  long size, sz, last_quad;
  int prev_err, status;

  for(source = src, tx_addr = dest, size = length;
      size > 0;
      )
  {
      /* Set sized in one call copied to <= 8KB */
      sz = MIN(MCCOPY_SIZE,size);

      /*
       * Use the error check routine imc_ckerrcnt_mr() to provide
       * some flow control to the remote adapter through HW ACKs.
       * Handle MC errors.
       */
      do {
          prev_err  = imc_rderrcnt_mr(LOGICAL_RAIL);
          last_quad = imc_bcopy(source,tx_addr, sz,
                                dest_write_only, first_dest_quad);
      } while ((status = imc_ckerrcnt_mr(&prev_err,0)) != IMC_SUCCESS);

      source  += sz;
      tx_addr += sz;
      size    -= sz;
  } while (size);
```

```
      return(last_quad);
}
```

For more information about the Memory Channel API and loopback mode,
see the *Cluster Highly Available Applications* manual and imc_asattach(3).
For information about the imc_bcopy() function, see imc_bcopy(3).

2.11.2 Small Memory Leak When clu_get_info() Fails

The clu_get_info() function allocates memory for a clu_info_resp
structure. In most error cases, the function frees the structure before
returning. However, when called on a system that is not a cluster member,
clu_get_info() returns before freeing the clu_info_resp structure.
Unless you have a program that loops calling clu_get_info(), you will not
notice this problem.

The workaround is to avoid using clu_get_info() to test whether a
system is a cluster member. The following example illustrates one approach:

```
# include <sys/param.h>
# include <sys/sysconfig.h>
# include <stdio.h>

int get_cluInfo ()
{
    cfg_attr_t one_attribute[1];

    strcpy (one_attribute[0].name, "clu_configured");
    if (cfg_subsys_query (NULL, "generic", one_attribute, 1) ==
        CFG_SUCCESS)
        if ((one_attribute[0].status == CFG_ATTR_SUCCESS) &&
            (one_attribute[0].attr.num.val != 0))
            return (TRUE);
    return (FALSE);
}

main ()
{
    int i = 0;
    char host[MAXHOSTNAMELEN];

    if (!(gethostname (host, MAXHOSTNAMELEN)))
      {
        if (get_cluInfo ())
          {
            printf ("System %s is a cluster member\n", host);
          }
        else
          {
            printf ("System %s is a _not_ a cluster member\n", host);
          }
      }
    else
      {
        printf ("Cannot determine host name\n");
        exit (1);
      }
    exit (0);
```

```
}
```

2.12 SysMan Menu

This section discusses problems that you may encounter when you use
SysMan Menu in a cluster.

2.12.1 Non-root Users with No Home Directory Cannot Run System Management Applications

Most system management applications require `root` privileges to make
configuration changes. Non-root users are permitted to run system
management applications only to view the current configuration. They are
prevented from changing the configuration.

In a cluster, the system management applications use the remote shell
command (`rsh`) to execute commands at a remote host. Part of the
`rsh` command processing includes verifying access in the remote user's
`$HOME/.rhosts` file in their home directory. For this reason, a non-root
user without a home directory who runs a system management application
might encounter a core dump. Users can avoid these problems by ensuring
that they have home directories set up before attempting to use the system
management applications.

2.13 SysMan Station

This section discusses problems that you may encounter when using SysMan
Station in a cluster.

2.13.1 SysMan Station Might Display Cluster Status Incorrectly

The SysMan Station relies on events generated by the Event Manager
subsystem in order to monitor and display cluster status. In the following
situations, the SysMan Station may reflect the state of the system
incorrectly:

- After a cluster member has booted, the Network light in the Monitor
  window may indicate a warning state (yellow) when no network errors
  exist. This condition is caused by network events that are generated
  during the boot sequence. To clear this warning, follow these steps:

  1. Click on the Network light in the Monitor window to display the
     Network Event window.

  2. Click on the Clear Events button.

- If the cluster application availability daemon (`caad`) fails to start on
  a cluster member, the SysMan Station will not correctly display the

state of CAA objects. For example, this situation can happen when the TruCluster Server license is not loaded on all the cluster members. To obtain accurate information on CAA applications from the SysMan Station, follow these steps:

1. Start the caad daemon on the affected cluster members using the following command:

   ```
   # /usr/sbin/caad
   ```

2. Restart the smsd daemon using the following command:

   ```
   # /sbin/init.d/smsd restart
   ```

2.13.2 SysMan Station Might Display New Hardware Objects Incorrectly

If a new disk device is added or an existing disk device is replaced in a running cluster, the SysMan Station's Hardware View may display the new or modified disk object incorrectly. The disk object may be positioned incorrectly in the hardware hierarchy; for example, the disk may be drawn as a child of the host object instead of as a child of a SCSI bus.

To correct the view, restart the SysMan Station daemon (smsd) on each cluster member by performing the following steps on all affected members:

1. Close all open SysMan Station sessions.

2. Enter the following command:

   ```
   # /sbin/init.d/smsd restart
   ```

2.13.3 Properties Might Not Be Displayed for Selected Objects

Properties may not be displayed for selected objects. The Properties dialog box may appear briefly on the screen or may not be displayed at all.

To work around this problem, continue to try to display properties in the current SysMan Station client, or exit the SysMan Station client and start a new SysMan Station session.

2.13.4 Some SysMan Station Applications Display Wrong Target Member Name

When the following applications are launched from the SysMan Station, their title bars incorrectly display the name of the cluster member on which the SysMan Station client is running instead of the cluster member that is the target of the application's actions:

• Security Auditing Configuration

- Network Configuration Applications
- NFS Configuration Applications
- NTP Configuration Applications
- PPP Configuration Applications

The application is directed to the correct cluster member; only the name in the title bar is incorrect.

2.13.5 If a Cluster Member Panics, You Must Restart smsd on All Cluster Members

If a system panic occurs on any node in a cluster, you must restart all of the SysMan Station daemons, smsd, in that cluster in order to ensure the consistency and correctness of the SysMan Station Filesystem views.

See smsd(8) for information on how to stop and restart the SysMan Station daemons in a cluster.

2.14 Documentation

This section discusses TruCluster Server Version 5.1B documentation issues.

2.14.1 No Cluster LAN Interconnect Manual for Version 5.1B

The *Cluster LAN Interconnect* manual, which was shipped for Version 5.1A, is not included with Version 5.1B. The information that was in that manual has been updated and merged into the remaining manuals in the Version 5.1B cluster documentation set.

2.14.2 Typographical Error In the *Cluster Hardware Configuration* Manual

Section 2.2.2 "Memory Channel Restrictions" of the *Cluster Hardware Configuration* manual contains the following paragraph:

```
To prevent this situation in standard hub mode (two member
systems connected without a Memory Channel hub), install a
second Memory Channel rail. A hub failure on one rail will
cause failover to the other rail.
```

The "without" is incorrect; standard hub mode requires the use of a Memory Channel hub. The paragraph should read:

```
To prevent this situation in standard hub mode (two member
systems connected with a Memory Channel hub), install
a second Memory Channel rail. A hub failure on one rail will
cause failover to the other rail.
```

Index