TruCluster Software Products

Administration

Part Number: AA-R88JA-TE

January 1998

Product Version:	TruCluster Production Server Software Version 1.5 and TruCluster Available Server Software Version 1.5
Operating System and Version:	DIGITAL UNIX Version 4.0D

This manual describes how to manage a cluster that is running TruCluster Production Server Software or TruCluster Available Server Software.

Digital Equipment Corporation
Maynard, Massachusetts

Digital Equipment Corporation makes no representations that the use of its products in the manner described in this publication will not infringe on existing or future patent rights, nor do the descriptions contained in this publication imply the granting of licenses to make, use, or sell equipment or software in accordance with the description.

Digital conducts its business in a manner that conserves the environment and protects the safety and health of its employees, customers, and the community.

Contents

About This Manual

1 Understanding Available Server Environments and Clusters

2 Managing ASE Members and Cluster Members

3 Managing Networks in a TruCluster Software Configuration

6 Setting Up a Disk Service

7 Setting Up a User-Defined Service

8 Setting Up a DRD Service (PS)

9 Setting Up a Shared Tape Service

10 Managing ASE Services

11 Using the Cluster Monitor

12 Troubleshooting

A Error Messages

Examples

Figures

Tables

About This Manual

This manual describes how to manage a TruCluster Production Server configuration or TruCluster available server environment (ASE) configuration.

Audience

This manual is for the ASE or cluster administrator.

Organization

This manual consists of twelve chapters, three appendixes, and a glossary. A brief description of the contents follows:

Chapter 1 Introduces the concept of the available server environment (ASE); compares and contrasts the use of ASEs in Production Server clusters and Available Server configurations; explains the operation of the daemons, drivers, and subsystems that comprise a cluster or ASE; and provides a brief overview of the ASE manager utility (`asemgr`).

Chapter 2 Describes how to configure and manage ASE and cluster membership.

Chapter 3 Discusses how to configure and administer member and client networks in a Production Server cluster or Available Server configuration.

Chapter 4 Explains how to prepare and set up highly available services, including discussions of the concepts fundamental to all types of service, such as Automatic Service Placement (ASP) policies, file systems and the Logical Storage Manager (LSM), and service action scripts.

Chapter 5 Describes how to configure and start a Network File System (NFS) service.

Chapter 6 Describes how to configure and start a disk service.

Chapter 7 Describes how to configure and start a user-defined service.

Chapter 8 Describes how to configure and start a distributed raw disk (DRD) service.

Chapter 9 Describes how to configure and start a shared tape service.

Chapter 10 Describes how to manage highly available services, including how to display service information, and delete, relocate, and temporarily stop services. It also describes the various mechanisms for modifying service, focusing on how to perform complex changes to a service's Advanced File System (AdvFS) and Logical Storage Manager (LSM) configuration.

Chapter 11 Introduces the Cluster Monitor (`cmon`) and shows how to run the Cluster Monitor as a highly available service.

Chapter 12 Discusses how to investigate and resolve common TruCluster problems.

Appendix A Describes alert messages that may be logged during the operation of a cluster.

Appendix B Lists the kernel attributes provided by the cluster subsystems for Production Server clusters only.

Appendix C Lists the configuration variables provided with TruCluster software.

Glossary Describes the TruCluster software terms introduced in the documentation set.

Related Documents

Consult the following TruCluster Software Products documentation for assistance in cluster configuration, installation, and administration tasks:

- *Release Notes*—Documents known restrictions and other important information about the TruCluster software products.
- *Hardware Configuration*—Describes how to set up the processors that are to become cluster members, and how to configure cluster shared storage.
- *Software Installation*—Describes how to install the TruCluster software on the systems that are to participate in the cluster.
- *Application Programming Interfaces*—Describes the application programming interfaces (APIs) provided by the distributed lock manager (DLM) and cluster information services.
- *MEMORY CHANNEL Application Programming Interfaces*—Describes the APIs that allow programming to the features of the MEMORY CHANNEL™ hardware.

You may also find the following DIGITAL UNIX operating system manuals useful:

- *System Configuration and Tuning*
- *Network Administration*

- *System Administration*

For help with storage management, see the following manuals:

- POLYCENTER Advanced File System and Utilities for DIGITAL UNIX *Guide to File System Administration*
- POLYCENTER Advanced File System Utilities *Reference Manual*
- POLYCENTER Advanced File System Utilities *Release Notes*
- POLYCENTER Advanced File System Utilities *Installation Guide*
- DIGITAL UNIX *Logical Storage Manager* manual

Reader's Comments

DIGITAL welcomes any comments and suggestions you have on this and other DIGITAL 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
  ```

- Mail:

  Digital Equipment Corporation
  UBPG Publications Manager
  ZKO3-3/Y32
  110 Spit Brook Road
  Nashua, NH 03062-9987

  A Reader's Comment form is located in the back of each printed manual. The form is postage paid if you mail it in the United States.

Please include the following information along with your comments:

- The full title of the book and the order number. (The order number is printed on the title page of this book and on its back cover.)
- The section numbers and page numbers of the information on which you are commenting.
- The version of DIGITAL UNIX that you are using.
- If known, the type of processor that is running the DIGITAL UNIX software.

The DIGITAL 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 DIGITAL technical support office. Information provided with the software media explains how to send problem reports to DIGITAL.

Conventions

This manual uses the following typographical conventions:

#
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.

⋮
A vertical ellipsis indicates that a portion of an example that would normally be present is not shown.

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.

PS
Abbreviation for the TruCluster Production Server Software.

AS
Abbreviation for the TruCluster Available Server Software.

MC
Abbreviation for the TruCluster MEMORY CHANNEL Software.

1

Understanding Available Server Environments and Clusters

The TruCluster software products suite consists of three separately licensed products:

- TruCluster Available Server Software
- TruCluster Production Server Software
- TruCluster MEMORY CHANNEL Software

TruCluster MEMORY CHANNEL Software supplies an application programming interface (API) library that lets applications perform high-speed data transfers between systems connected to the MEMORY CHANNEL interconnect. (This API library is also included in the Production Server Software.) TruCluster MEMORY CHANNEL Software, unlike TruCluster Available Server Software and TruCluster Production Server Software provides neither shared storage nor application failover capabilities. Consequently, management of MEMORY CHANNEL Software configurations is largely a matter of setting up the appropriate hardware, installing the software, and understanding the MEMORY CHANNEL API library. These tasks are described in the *Hardware Configuration*, *Software Installation*, and *MEMORY CHANNEL Application Programming Interfaces* manuals. Therefore, the remainder of this manual focuses exclusively on managing Available Server and Production Server configurations.

This chapter provides an overview to understanding available server environments (ASEs), the additional components of a Production Server cluster, and how to use the `asemgr` utility.

1.1 Using Storage Availability Domains in an Available Server Configuration or a Production Server Cluster

TruCluster Available Server Software and TruCluster Production Server Software let you configure a highly integrated organization of member systems, services, and storage devices. From a client's perspective, this configuration appears to be a powerful single-server system, providing greater application **availability** than is possible with a single system, and scalability beyond the limits of a single symmetric multiprocessing system.

A key component of the TruCluster Available Server Software and TruCluster Production Server Software is the **storage availability domain**. A storage availability domain is a collection of nodes that can access commonly shared storage devices in an available server environment (ASE). These nodes are considered to be ASE members.

Because all members in a given ASE can access the same shared storage, an application that requires that storage can run on any member. Both Production Server Software and Available Server Software let you configure such an application so that it runs on a single ASE member and, upon a failure of that member, restarts on another. This application could be a service that exports Network File System (NFS) file systems to clients, a disk-based application like a database engine or mail service, a tape-based service, or a nondisk-based application, such as a remote login service.

The most significant difference between Production Server Software and Available Server Software is that Production Server Software lets you develop and deploy an application whose components run concurrently, with equal access to raw disk data, on any node in the Production Server configuration. A Production Server cluster provides an ideal environment for applications that require high availability and performance, such as highly parallelized databases and transaction processing systems. The means by which raw disk data is provided to the components of applications distributed throughout the cluster involves a special type of ASE service (provided only with Production Server Software) known as **distributed raw disk** (**DRD**). Use of a **distributed lock manager** (**DLM**) ensures synchronized access to the data provided clusterwide by DRD services.

Because a Production Server cluster and an Available Server configuration both employ ASE technology, the administrator of either fundamentally manages ASE membership, ASE services, and service storage. However, the distributed nature of services within a Production Server cluster makes the configuration and management of ASEs within the cluster somewhat different than managing the ASE in an Available Server configuration. This manual will make the necessary distinctions as appropriate. To begin, keep the following configuration rules in mind when dealing with a TruCluster configuration:

- An Available Server configuration consists of one ASE. A Production Server cluster must contain one ASE; it can include up to four ASEs.
- An Available Server configuration’s membership is equivalent to the membership of its sole ASE. All members are connected to all common, shared storage and the same primary network.

  A Production Server cluster’s membership is determined by the member systems’ common connection to the MEMORY CHANNEL interconnect.

- A Production Server cluster has two to eight members. A cluster member can also be a member of an ASE within the cluster, or the cluster member may not be a member of an ASE.
- An ASE contains from two to four members. As a result, a Production Server cluster can include at most four ASEs, each containing two systems.
- You establish ASE membership using the `asemgr` utility. Available Server Software uses a primary network interconnect (Ethernet, FDDI, or ATM) to maintain ASE membership.

  You establish cluster membership by installing the Production Server Software on each member system and, during installation, by specifying the addresses of all members' MEMORY CHANNEL interconnects in each member's `/etc/hosts` file. Within a Production Server cluster both cluster and ASE membership are maintained over the MEMORY CHANNEL interconnect.

1.2 Components of an Available Server Environment

An **available server environment** (ASE) is a multinode configuration in which member systems and highly available storage are connected to shared SCSI buses. Software running on each ASE member monitors the health of ASE member systems and shared storage. In case of a failure, the ASE software causes services to fail over to surviving systems in the ASE that share access to the associated storage. Scripts associated with each service control failover.

An Available Server configuration contains a single ASE. A Production Server cluster can contain one or more nonoverlapping ASEs. A given cluster member can be a member of at most one ASE. However, a cluster member does not have to be a member of an ASE.

ASE members run the ASE daemons and driver, which monitor the network interconnects and the status of the systems, disks, and shared SCSI buses in the ASE. The ASE daemons and driver are as follows:

- ASE director daemon—Runs on only one member of the ASE and controls the entire ASE.
- ASE agent daemon—Runs on each member of the ASE and controls ASE operations on that member.
- Host status monitor (HSM) daemon—Runs on each member of the ASE. Like the AM driver, it also monitors that ASE and reports any member system or network failure to the director and agent daemons. The HSM, with the help of the AM driver, detects SCSI bus partitions.

- Availability manager (AM) driver—Runs on each member of the ASE as part of its kernel. It monitors that ASE and reports any member system failure to the HSM daemon and device connectivity failures to the agent daemon.
- Logger daemon—Tracks all the ASE messages that are generated by the members of the ASE.

The following sections describe the ASE daemons and the AM driver.

1.2.1 The ASE Director Daemon

The ASE director daemon (asedirector) controls an entire ASE. It coordinates most of the activities that occur during ASE setup and operation and has a global view of the ASE. The ASE director daemon maintains information about ASE members and services, including which member system is running which service. It decides what actions to take when a change in the environment occurs and coordinates these actions in the ASE.

The ASE director daemon runs on only one member system in the ASE. If an ASE director daemon is not running on one of the members, the agent daemons on the members choose an ASE member to run the daemon.

The ASE director daemon ensures that all the services are always configured on all the member systems, using the ASE agent daemon running on each member to implement its decisions. It also maintains such information as the current state of services and member systems.

For example, I/O events, such as a device going off line or a disk reservation failure, are detected by the availability manager (AM) driver and reported to the director daemon by the agent daemon. Member and network events, such as a member system going down or a network partition, are detected by the HSM daemon and then reported to the director daemon.

In addition, the ASE director daemon handles all requests from the asemgr utility, such as configuring a service or displaying status.

1.2.2 The ASE Agent Daemon

An ASE agent daemon (aseagent) controls ASE operations on each member of an ASE and has a local view of the ASE. An ASE agent daemon synchronizes access to shared resources, using the AM driver interfaces to reserve disks and to receive notification of lost reservations and device connectivity losses.

Each ASE agent daemon reports local events (such as disk failures) to the ASE director daemon and also performs local ASE management tasks as requested by the director daemon. An ASE agent daemon invokes the commands to configure, start, and stop a service at the request of the director daemon.

An ASE agent daemon runs on each member of an ASE. On each member, the ASE agent daemon initializes the ASE, starts the HSM daemon, and starts the director daemon if necessary. For example, if the ASE director daemon terminates unexpectedly, the ASE agent daemons on the ASE members choose a member on which to run the ASE director daemon, and the ASE agent daemon on that member system starts the ASE director daemon.

1.2.3 The Host Status Monitor Daemon

A host status monitor (HSM) daemon (asehsm) runs on each member in an ASE and monitors member system status. It detects any breaks (partitions) in the network connections between member systems. The HSM daemon uses the availability manager (AM) driver to query systems over the SCSI bus. It uses network interfaces to query systems over the network.

In addition to providing the interface that can query hosts, the AM driver provides the HSM daemon running on a member system with the ability to transfer data when the network is not working.

The HSM daemon is started by the ASE agent daemon and reports to both the ASE director daemon (if it is running locally) and the ASE agent daemon. For example, if a member system goes down, the AM driver notifies the HSM daemon that the SCSI member system query has timed out or that it has noticed a break in the network connection.

1.2.4 The Availability Manager Driver

The availability manager (AM) driver is a kernel-level device driver that provides device reservations (locking), monitors remote hosts on the SCSI bus, and provides error and event notifications. Changes in the hardware run-time status are detected by the AM driver and reported to the host status monitor (HSM) daemon and the ASE agent daemon running on the member system.

The AM driver interfaces reserve disks and ensure that only one ASE member has access to a shared device at one time. They allow the agent daemon to query devices and the HSM daemon to query members.

If an I/O bus partition occurs (for example, the SCSI bus cable is disconnected from the member system), the AM driver notifies the HSM daemon that the system query failed. If a device is powered off, the AM driver notifies the ASE agent daemon that a device path failure has occurred, or that an I/O bus partition has occurred such that a system no longer has connectivity to a device.

1.2.5 The Logger Daemon

The logger daemon (`aselogger`) tracks all the ASE messages that are generated by all the members of an ASE. When you install the TruCluster software on a system, you are prompted to determine if you want a logger daemon running on the system. A logger daemon can be run on more than one member system in an ASE.

The logger daemon uses the DIGITAL UNIX event logging facility, `syslog`, which collects messages that are logged by the various kernel, command, utility, and application programs. Messages are logged to a local file or forwarded to a remote system, as specified in the local system's `/etc/syslog.conf` file.

The logger daemon collects messages generated by the `asemgr` utility, the ASE director daemon, the ASE agent daemon, and the logger daemon. Messages generated by the host status manager (HSM) daemon and the availability manager (AM) driver are logged only to the local system. If all the logger daemons in the ASE stop, daemon messages continue to be logged, but only locally.

See the DIGITAL UNIX *System Administration* manual, `syslog`(3), and `syslogd`(8) for information on system event logging. See Appendix A for a description of some ASE error messages.

1.3 Additional Components of a Production Server Cluster

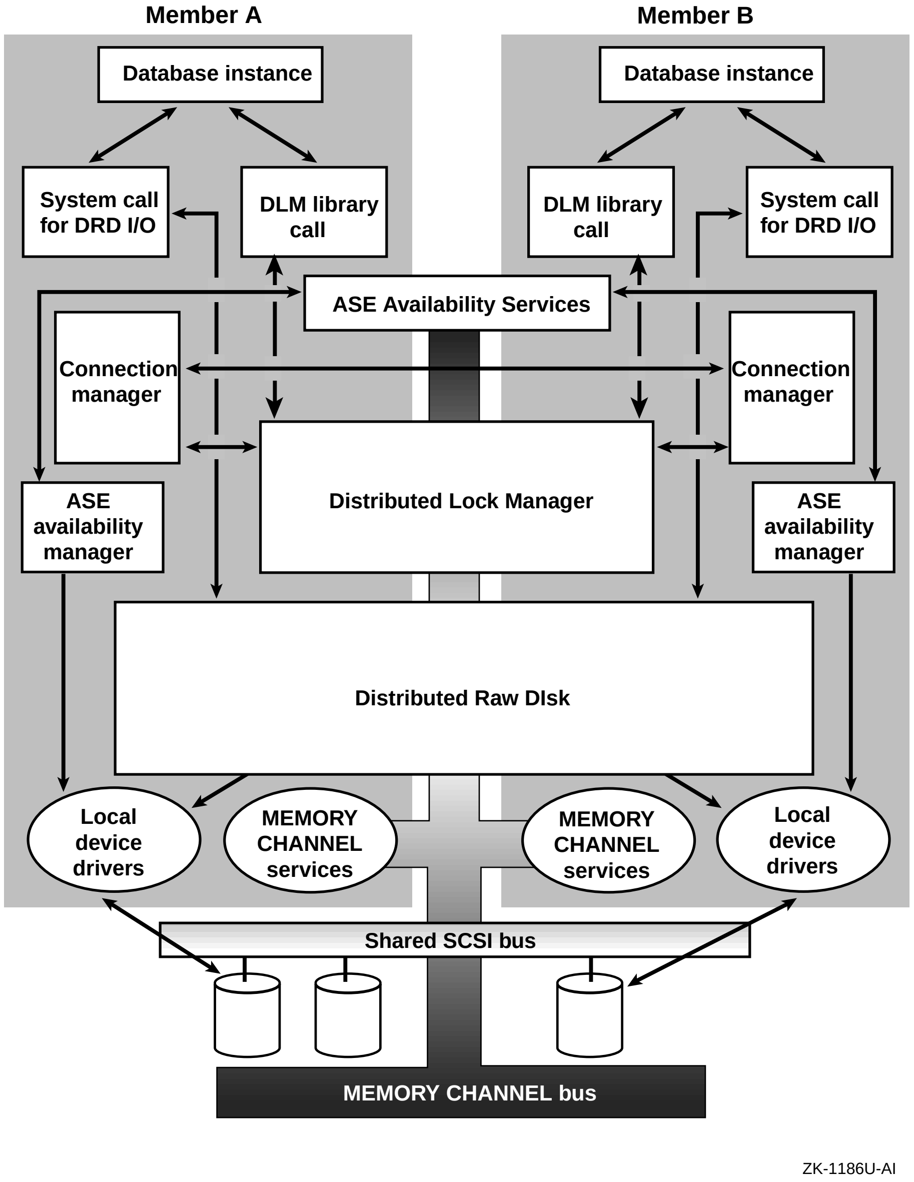
Although the ASE components discussed in Section 1.2 are fundamental to a Production Server cluster's ability to allow database system elements to fail over from member to member without disrupting access to data, there are several other technologies used in the cluster that are critical to the operation of highly available, large database systems:

- Distributed raw disk (DRD) services provide transparent remote access to cluster storage from any member system.
- Distributed lock manager (DLM) services allow the elements of a distributed database system to synchronize their activities from independent member systems.
- The connection manager supports the other subsystems by maintaining cluster membership and managing the addition and removal of members to and from the cluster.

- The MEMORY CHANNEL subsystem supports high-speed data sharing among member systems across the MEMORY CHANNEL interconnect.

Figure 1–1 shows the relationship of these components. The remainder of this chapter provides additional details on the operation of these components.

Figure 1–1: Overview of Production Server Software Subsystems



1.3.1 Distributed Raw Disk

Distributed raw disk (DRD) services allow a disk-based, user-level application to run within a cluster, regardless of where in the cluster the physical storage on which it depends is located. A DRD service allows an application, such as a distributed database system or transaction processing (TP) monitor, parallel access to storage media from multiple cluster members. Applications that perform I/O involving sets of large data files, random access to records within these files, and concurrent read/write data sharing can benefit from using the features of DRD. As deployed within an ASE, a DRD service can survive failures of both the server system and any mirrored disk participating in the service.

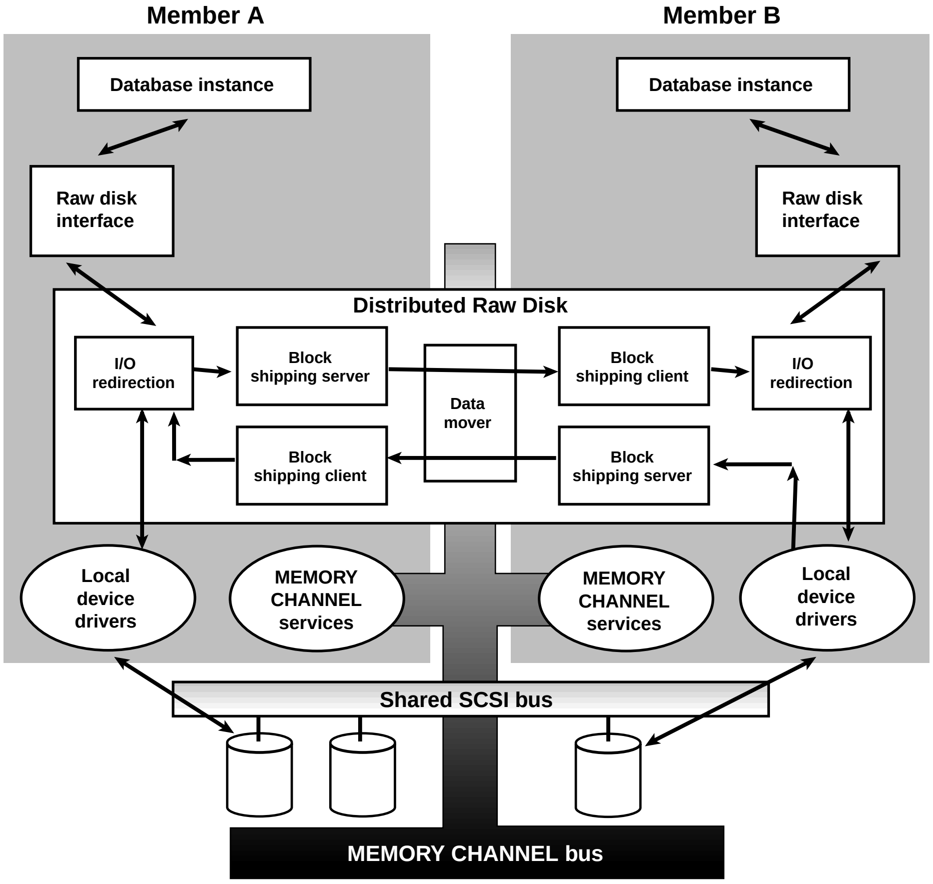
The DRD subsystem, shown in Figure 1–2, consists of four primary components:

- The raw disk interface (the DRD pseudodevice driver) on client and server nodes receives user requests through conventional system calls such as `open`, `close`, `read`, `write`, and `ioctl`. For this reason the driver is considered to be a raw (or character) device driver. Because it relies on an underlying physical device driver to control the disk device, the DRD driver is also considered a pseudodevice driver. When the DRD driver receives a user request, it first determines whether the node on which it is running is the server of the physical device that is the object of the request as follows:
  - If the node that receives the user request is serving the physical device that is the object of the request, the DRD driver considers the request to be a local request. The driver then passes the local request to the underlying physical device driver, such as the SCSI CAM driver or the **Logical Storage Manager (LSM)**.
  - If the node that receives the user request is not serving the physical device that is the object of the request, the DRD driver considers the request to be a remote request. The driver passes the remote request across the network transport to the other node that is the device's server node. See `drd`(7) for more information about the DRD pseudodevice driver.
- A block shipping client (`bsc`) that ships requests for access to remote DRD devices to the appropriate DRD services, and returns responses to the caller. See `drd` (7) for more information on the `bsc`.
- A block shipping server (`bss`) that accepts requests from `bsc` clients, passes them to a local device driver for service, and returns results to the clients. See `bssd`(8) for more information on the `bss`.
- A DRD management facility, not shown in Figure 1–2, that supports DRD device naming, device creation and deletion, device relocation, and

device status requests. See Chapter 8 for more information on DRD service administration.

The DRD subsystem, in conjunction with ASE services, is designed to provide applications with uninterrupted access to storage devices. Depending upon the hardware configuration of the cluster, DRD can withstand member failures, controller failures, and disk failures.

Figure 1–2: Distributed Raw Disk

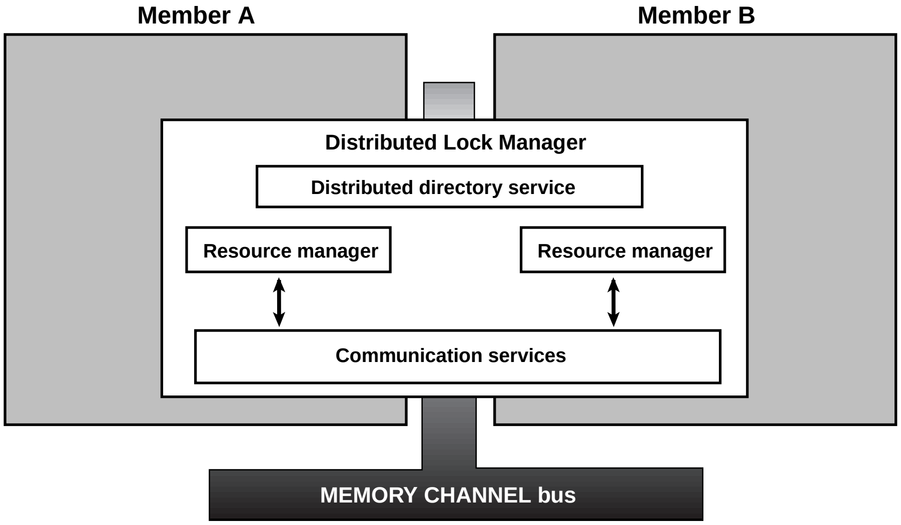


ZK-1188U-AI

1.3.2 Distributed Lock Manager

The distributed lock manager (DLM), shown in Figure 1–3, synchronizes access to the resources that are shared among cooperating processes throughout the cluster. For example, a distributed database application uses lock manager services to coordinate access to the shared disks participating in the database.

Figure 1–3: Distributed Lock Manager



ZK-1189U-AI

An application secures a lock on a named shared resource. Resource names can be single-dimensional or tree-structured. A resource tree allows you to create a hierarchy of locks and sublocks that reflect the structure of a shared resource. The DLM:

- Provides mutual exclusion, restricted sharing, and full sharing of data access
- Allows notification when a lock holder is blocking another process's access to a resource or when a queued lock request completes
- Allows conversion between less restrictive and more restrictive lock modes
- Provides services that return information about locks

The DLM employs a distributed, centralized tree design. It does not replicate lock information on each cluster member. Rather, the cluster member that manages a lock tree maintains all information about that tree. The member that holds a given lock is aware of only its contribution of that lock to the resource. Any member system can serve as the master for any lock tree, which distributes the overall lock management load.

The DLM uses a distributed directory service to quickly locate the directory node for a resource tree. A directory table associates a root resource name with the cluster member that is the manager of the resource. This directory table is identical on all cluster members.

The DLM is designed to handle member failures. If a lock holder fails, its locks are released. If a member system fails, a new lock master for locks previously mastered on that member is chosen and provided with all pertinent lock information.

The DLM also maintains a communications service that the connection manager uses to establish a communications channel between member systems.

1.3.3 Connection Manager

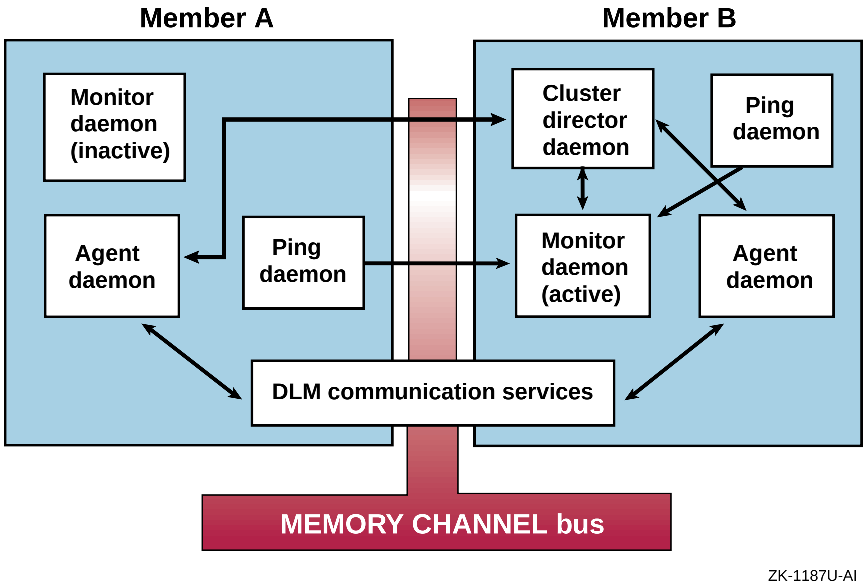
Systems in a Production Server cluster configuration share data and system resources, such as access to data and files. To achieve the coordination required to maintain data integrity, the systems must maintain a clear sense of cluster membership. The **connection manager** ensures that the clustered systems communicate with one another, and it enforces the rules of cluster membership.

The connection manager is a set of daemons that creates a cluster when the first member is booted, and reconfigures the cluster when other systems join or leave it. The overall responsibilities of the connection manager are to:

- Prevent partitioning.
- Track which nodes in the cluster are active and which are not.
- Add member systems to and remove systems from the cluster.
- Establish and maintain a high-performance, highly reliable communications path between each cluster member for use by the DLM. The DLM uses the configuration data and other services provided by the connection manager to maintain a distributed lock database.
- Maintain configuration information and make it available to the Cluster Monitor utility and other administrative tools.

Figure 1–4 shows the components of the connection manager.

Figure 1–4: Connection Manager



The connection manager consists of a kernel component that maintains the configuration information and, as shown in Figure 1–4, the following daemons that control and distribute configuration information:

- Monitor daemon (`cnxmond`)—The monitor daemon runs on all cluster members. It is in a standby state on all but one member. On the member on which it is active, the monitor daemon acquires a MEMORY CHANNEL spinlock, registers an IP alias named `cluster_cnx`, and starts the cluster director daemon (`cnxmgrd`). The acquisition of the spinlock ensures that only one cluster director daemon is running at any given time in the cluster and prevents multiple registrations of the `cluster_cnx` service. When active, the monitor daemon receives membership requests and periodic keep-alive pings from member systems, and interacts with the cluster director daemon to maintain and distribute cluster configuration information. The monitor daemon also receives event information (such as cluster interconnect failure) from agent daemons.

  The monitor daemon passes information related to membership requests, pings, and events to the cluster director daemon, which maintains the cluster membership list and other configuration information. See `cnxmond`(8) for a description of the monitor daemon.

- Cluster director daemon (`cnxmgrd`)—The cluster director daemon runs on a single cluster member and forms a new cluster by adding systems as they request membership, or it recovers an existing cluster based on membership information from the latest configuration. If the system running the cluster director daemon fails, the monitor daemon on another system becomes active, acquires the MEMORY CHANNEL spinlock, and starts the cluster director daemon. See cnxmgrd(8) for a description of the cluster director daemon.

- Agent daemon (cnxagentd)—The agent daemon runs on all cluster members and acts as an remote procedure call (RPC) server to receive configuration data and instructions from the cluster director daemon. See cnxagentd(8) for a description of the agent daemon.
- Ping daemon (cnxpingd)—The ping daemon runs on all cluster members and acts as an RPC client to periodically interact with the monitor daemon. See cnxpingd(8) for a description of the ping daemon.

The TruCluster software installation procedure adds or modifies system startup scripts to automatically start these daemons each time the system boots.

1.3.4 MEMORY CHANNEL

In a Production Server configuration, all cluster members must have a direct connection to all other members to facilitate communications among members and provide a fast and reliable transport for passing messages throughout the cluster. This version of the TruCluster software product supports the MEMORY CHANNEL interconnect, a specialized interconnect designed specifically for the needs of clusters.

The MEMORY CHANNEL interconnect is based on a **peripheral component interconnect** (PCI), which cluster members use to communicate among themselves on a private subnet. (See the TruCluster Software Products *Hardware Configuration* manual and TruCluster Software Products *Software Installation* manuals for instructions on how to set up the MEMORY CHANNEL subnet.) Each cluster system has a MEMORY CHANNEL interface card that connects to a MEMORY CHANNEL hub. The MEMORY CHANNEL hub provides both broadcast and point-to-point connections between cluster members. In most two-member cluster configurations, a physical MEMORY CHANNEL hub is not used. Instead, the members utilize the **virtual hub mode** of the MEMORY CHANNEL interface card.

The Production Server configuration fails over from one MEMORY CHANNEL interconnect to another if a configured and available secondary MEMORY CHANNEL interconnect exists on all member systems, and one of the following situations occurs in the primary interconnect:

- More than ten errors are logged within one minute
- A link cable is disconnected

- The hub is turned off

After the failover completes, the secondary MEMORY CHANNEL interconnect becomes the primary interconnect. Another interconnect failover cannot occur until you fix the problem with the interconnect that was originally the primary.

If more than ten MEMORY CHANNEL errors occur on any member system within a one-minute interval, the MEMORY CHANNEL error recovery code attempts to determine if a secondary MEMORY CHANNEL interconnect has been configured on the member as follows:

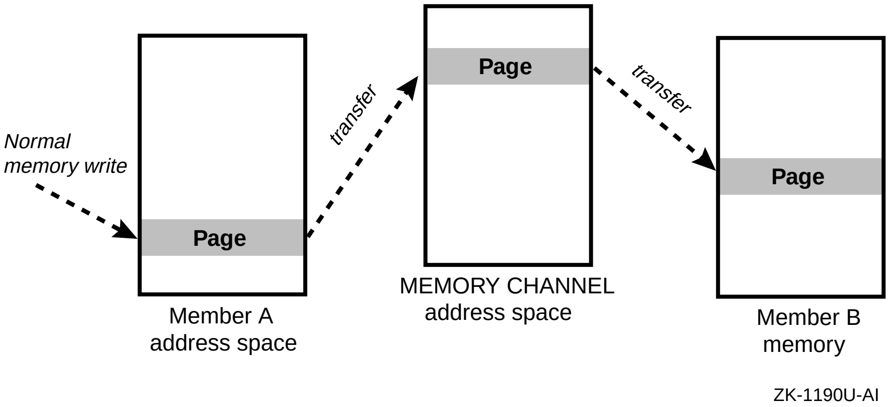
- If a secondary MEMORY CHANNEL interconnect exists on all member systems, the member system that encountered the error marks the primary MEMORY CHANNEL interconnect as bad and instructs all member systems (including itself) to fail over to their secondary MEMORY CHANNEL interconnect.
- If any member system does not have a secondary MEMORY CHANNEL interconnect configured and available, the member system that encountered the error displays a message indicating that it has exceeded the MEMORY CHANNEL hardware error limit and panics.

The MEMORY CHANNEL interconnect:

- Allows a cluster member to set up a high-performance, memory-mapped connection to other cluster members. These other cluster members can, in turn, map transfers from the MEMORY CHANNEL interconnect directly into their memory. A cluster member can thus obtain a write-only window into the memory of other cluster systems. Normal memory transfers across this connection can be accomplished at extremely low latency (3 to 5 microseconds).
- Has built-in error checking, virtually guaranteeing no undetected errors and allowing software error detection mechanisms, such as checksums, to be eliminated. The detected error rate is very low (on the order of one error per year per connection).
- Supports high-performance mutual exclusion locking (by means of spinlocks) for synchronized resource control among cooperating applications.

Figure 1–5 shows the general flow of a MEMORY CHANNEL transfer.

Figure 1–5: MEMORY CHANNEL Transfer



You need at least one MEMORY CHANNEL **adapter** installed in a PCI slot in each member system and a link cable to connect the adapters. If you have more than two members in your cluster, link cables are used to connect the MEMORY CHANNEL adapters to a MEMORY CHANNEL hub.

A redundant MEMORY CHANNEL configuration can further improve reliability and availability. In this case, you need a second MEMORY CHANNEL hub, a second MEMORY CHANNEL adapter in each cluster member, and link cables to connect the second MEMORY CHANNEL adapters to the MEMORY CHANNEL hub.

See the TruCluster Software Products *Hardware Configuration* manual for information on how to configure the MEMORY CHANNEL interconnect in a cluster.

1.4 Using the asemgr Utility

The `asemgr` utility allows you to administer the available server environment (ASE) and configure and manage services. The `asemgr` utility has an interactive mode and a command-line interface. If you enter the `asemgr` command with no options, the utility displays menus and task items and prompts you for information about the task you want to perform.

You can use the command-line interface for the `asemgr` utility if you want to include the `asemgr` command in shell scripts. The syntax for the command is as follows:

/usr/sbin/asemgr [*options*]

The *options* are as follows:

–d [-h *member*]|[-v *service*]|[-l]

Displays the status of all the member systems (-h) and services (-v) or specific member systems and services. Also displays the member systems that are running the logger daemon (-l).

–d [-C [*database*]]|[-c *service*]

Displays the contents of the current or specified ASE database (-C [*database*]) or the contents of the specified service (-c *service*).

–m *service member*

Relocates the specified service to the specified member system. When you **relocate a service**, you stop the service on the member system currently running the service and start the service on another member system.

–r *service*

Restarts a service.

–s *service* [*member*]

Starts the specified service and places it on line, making it available to clients. When the *member* parameter is specified, the service is started on that member, regardless of the service's current ASP policy.

–x *service*

Stops the specified service and places it off line, making it unavailable to clients.

Some ASE administrative tasks can lock the ASE. If you try to run the asemgr utility and the ASE is locked, the following message is displayed:

```
ASE is locked by 'hostname'
```

This message indicates that the task cannot be performed because another member system is running the asemgr utility.

2

Managing ASE Members and Cluster Members

An available server environment (ASE) manages a collection of systems and the shared SCSI buses to which they are connected, and provides an environment in which services can be started, stopped, and automatically relocated in response to a software or hardware failure. ASEs provide high availability and increase storage capacity, reduce performance bottlenecks, and permit a wider range of configurations.

An Available Server configuration contains but a single ASE. A Production Server configuration contains one or more ASEs. As discussed in the TruCluster Software Products *Software Installation* manual, when you install the Production Server Software, you identify each ASE that is to exist within the cluster by assigning an ASE identifier (ASE_ID) to those members that are to participate in an ASE.

The **ASE_ID** is a value from 0 to 63 that cluster software uses to uniquely identify the ASE in which the system resides within the cluster. Each ASE has a unique ASE_ID; all systems in the same ASE share the same ASE_ID. (In an Available Server configuration, all systems have an ASE_ID of 0. You are not prompted to supply an ASE_ID during software installation.)

The following sections describe how to manage the membership of ASEs and clusters. It discusses the following tasks:

- Adding members to a cluster (Production Server cluster only) (Section 2.1)
- Adding members to an ASE (Section 2.2)
- Enabling ASE on an existing cluster member (Production Server cluster only) (Section 2.3)
- Deleting members from an ASE (Section 2.4)
- Displaying the status of ASE members (Section 2.5)
- Initializing ASE member systems (Section 2.6)
- Stopping and restarting ASE activity (Section 2.7)
- Shutting down a member (Section 2.8)

2.1 Adding a New System to a Cluster (PS)

To add a new system to an existing Production Server configuration, follow these steps:

1. Follow the instructions in the TruCluster Software Products *Hardware Configuration* manual to physically connect the new system to existing shared storage, the MEMORY CHANNEL interconnect, and external networks.
2. Connect the new system to the MEMORY CHANNEL subnet, turn on the power, and boot the system.
3. Install the same versions of the DIGITAL UNIX operating system and TruCluster software on the new system as you installed on existing cluster member systems. Decide which available server environment (ASE), if any, the new system will belong to and specify the correct ASE_ID during cluster base **subset** configuration.
4. Ensure that the settings of the following `/etc/sysconfigtab` attributes are the same on the new system as on all current member systems:
   - `dochecksum`—Enables or disables Transmission Control Program/Internet Protocol (TCP/IP) checksums. (See Appendix B for more information.)
   - `rx_mapping_enabled`—Enables or disables copy avoidance. (See Appendix B for more information.)
   - `rm_rail_style`—Configures the reliability style of the MEMORY CHANNEL interconects on a cluster member. (See Appendix B and the TruCluster Production Server Software *MEMORY CHANNEL Application Programming Interfaces* guide for more information.)
   - `enable_extended_uids`—Enables or disables extended UIDs in the base operating system. (See the DIGITAL UNIX *Release Notes* for more information.)

Warning

Failure to match the current configuration can cause one or more systems to panic when you attempt to add the new system to the cluster.

5. Modify the new `/etc/hosts` file on the new system to add entries for the IP address and hostname associated with the MEMORY CHANNEL subnet for each existing member system (including the new system).
6. Modify the `/etc/hosts` file on existing cluster member systems to include an entry for the new system's MEMORY CHANNEL IP address and hostname.
7. If the new cluster member is to belong to an ASE, run the `asemgr` utility on an existing member of the ASE to which the new system will belong and add the new member. The updated ASE database will be propagated to all ASE members when the new system is rebooted.

See Section 2.2 for instructions on adding a new member to an ASE.

2.2 Adding Member Systems to an ASE

The following requirements pertain to adding member systems to an available server environment (ASE):

- The host name and IP address for each member system must be included in all the member systems' local `/etc/hosts` files. For Available Server configurations, this host name can correspond to the network interface you specify in the `HOSTNAME` configuration variable in the local `/etc/rc.config` file. For a Production Server configuration, it must correspond with the cluster interconnect name that the cluster installation script automatically adds to the `/etc/rc.config` file in the `CLUSTER_NET` configuration variable.
- You must include network interface names in each member system's local `/etc/hosts` file, and they must be configured on the member system. See your software installation manual and the DIGITAL UNIX *Network Administration* manual for more information.
- After you set up the cluster hardware configuration and install the TruCluster software, immediately use the `asemgr` utility to configure the cluster's ASEs, and add the member systems to each ASE. Add all the member systems from the same system. See Section 2.3 for more information.
- If you want to change the name of a member system, you must use the `asemgr` utility to delete the member system from the ASE, change the name of the system, and then add the renamed member system to the ASE. Make sure you also make changes as appropriate to members'`/etc/host` files. See Section 2.4 for more information.
- You must ensure that the ASE daemons do not time out because other system processes have a higher scheduling priority. The ASE daemons should have a scheduling priority that is higher than normal system

processes, because the daemons must be able to respond to administrative commands and other events in the ASE. The daemons' high priority enables the ASE to operate even when the member systems are busy. See Section 12.3 for more information.

- Each member system that participates in an ASE must reside in the same Berkeley Internet Name Domain (BIND) domain.

Use the `asemgr` utility to add one member system at a time to the ASE. The system on which you run the `asemgr` utility for the first time is your first member system. You must add at least one other member system to your ASE. Add all the member systems from the same system. Do not run the `asemgr` utility on one system and add one member system, and then run the `asemgr` utility on another system and add a different member system.

In an Available Server configuration, after you enter the names of all the member systems, you are prompted for additional network interfaces for each member system. Before you add a member system, all network interfaces must be configured on the system. See Section 3.3.3 for information about using multiple networks in an Available Server configuration.

If you want to add member systems to an existing ASE, choose the "Add a member" item from the Managing the ASE menu. A list of the current member systems is displayed and you can add additional member systems, one at a time. You are then prompted for additional network interfaces for the new member systems. Example 2–1 shows how to add member systems to the ASE.

Example 2–1: Adding Member Systems to the ASE

```
                Managing the ASE

    a)  Add a member
    d)  Delete a member
    n)  Modify the network configuration
    m)  Display the status of the members
    C)  Display the configuration of the ASE database
    l)  Set the logging level
    e)  Edit the error alert script
    t)  Test the error alert script
     )  Enable ASE V1.5 functionality

    q)  Quit (back to the Main Menu)
    x)  Exit to the Main Menu             ?)  Help


Enter your choice [x]: a

Member List: tototc, gideontc

Enter a new member: daffytc

Member List: tototc, gideontc, daffytc
```

Example 2–1: Adding Member Systems to the ASE (cont.)

```
Is this correct (y/n) [y]: y

Would you like to define any other network interfaces to daffytc
 for ASE use (y/n)? [n]: n

                ASE Network Configuration

    Member Name          Interface Name       Monitor
    ___________          ______________       _______
    tototc                 tototc             Yes
    gideontc               gideontc           Yes
    daffytc                daffytc            Yes

Is this configuration correct (y|n)? [y]: y
```

Note

After an ASE member system has been deleted from an ASE, attempts to add it back into the ASE may fail. To resolve this problem, perform one of the following actions before trying to add the affected system back into the ASE:

- Reboot the member system.
- Enter the following command:

```
% /sbin/init.d/asemember restart
```

2.3 Enabling ASE on an Existing Cluster Member (PS)

When you install the TruCluster Production Server software on a member system, the installation procedure asks if you intend to run the available server environment (ASE) on that system. If you answer "no" at that time, and later decide to run ASE on that system, you must perform the following steps to enable ASE:

1. Select a value for the member system’s ASE_ID. The member’s ASE_ID must be the same as the ASE_IDs of the other member systems in the ASE it is joining.
2. Determine whether this system should run the ASE logger.
3. Enter the following commands, specifying the selected ASE_ID for *<var>* where appropriate:

```
# rcmgr set ASE on

# rcmgr set ASE_ID <var>
# rcmgr set ASELOGGER 1 if you wish to run the logging daemon
#
```

4. Rebuild the kernel using the `doconfig` program. See the DIGITAL UNIX *System Administration* manual for instructions on running the `doconfig` program.
5. Reboot the system.

2.4 Deleting Member Systems from an ASE

To delete member systems from the available server environment (ASE), choose the "Delete a member" item from the Managing the ASE menu. You then specify the number associated with the member system you want to delete. If a member system is running a service, ASE relocates the service to another member system.

You cannot delete the member system on which you are running the `asemgr` utility. To delete the last member system in an ASE, you must delete the TruCluster software subsets from that system.

2.5 Displaying the Status of the Member Systems

To display the status of the member systems in an available server environment (ASE), choose the "Display the status of the members" item from the Managing the ASE menu. The status of each member system and the agent daemon running on the system are displayed. See Section 1.2 for a description of the ASE daemons.

Example 2–2 shows an example of member system status in the ASE.

Example 2–2: Displaying Member System Status

```
                 Member Status

Member:             Host Status:     Agent Status:
tototc              UP               RUNNING
daffytc             UP               RUNNING
```

The director daemon obtains system status from the host status monitor (HSM) daemons running on all the member systems. The following table describes the information for the `Host Status` field:

Host Status	Description
UP	The member system is up and can be accessed by the member system that is running the ASE director daemon using the cluster interconnect. The member system can be queried over the cluster interconnect, and can add, delete, start, and stop services.
DOWN	The member system cannot be accessed by the member system that is running the director daemon using any network or shared SCSI bus. The member system does not answer queries over the cluster interconnect or SCSI bus, and cannot start or stop services.
DISCONNECTED	The member system is disconnected from all monitored networks. Any services running on the member system are stopped, and no services can be added, deleted, or started on the member system.
NETPAR	There is a network partition between the member system and the member system running the director daemon, although the member systems can communicate using SCSI bus queries. Services that are currently running on the member system remain running, but the member system cannot start or stop any service until it leaves this state.

The director daemon determines the status of the agent daemons running on the member systems. The following table describes the information in the `Agent Status` field:

Agent Status	Description
RUNNING	The ASE agent daemon is running on the member system.
DOWN	The ASE agent daemon is not running on the member system.
INITIALIZING	The ASE agent daemon that is running on the member system is in its initialization phase and will be running soon.

Agent Status	Description
UNKNOWN	The ASE director daemon cannot determine the state of the agent daemon on the member system.
INVALID	The ASE director daemon reports an invalid state for the agent daemon on the member system.

2.6 Initializing ASE Member Systems

If an available server environment (ASE) does not work correctly, make sure that you have adhered to the requirements in this manual and in the TruCluster Software Products *Hardware Configuration* manual. You should also read the TruCluster Software Products *Release Notes*. Running the `clu_ivp` utility may reveal the cause of errors. If you cannot fix the problem, you can initialize one or all of the member systems in an ASE.

Initializing a system stops any running ASE daemons and removes any member system and service information from the ASE database on the system. After you initialize a system, it can be added to an existing ASE or used in a new ASE.

You may want to initialize a system if you cannot add it to an ASE, or if the ASE database is corrupted on the member system. However, you may need to initialize all the ASE member systems to solve the problem.

The following sections describe how to initialize one or all of the ASE systems.

2.6.1 Initializing One System

To initialize one system, follow these steps:

1. If the system is already a member system, use the `asemgr` utility to delete the member system from the ASE. If you cannot delete the member system, you cannot initialize only this member.
2. If the system is not an ASE member system, delete the `/usr/var/ase/config/asecdb` ASE database file, if it exists, from the system.
3. Invoke the `/usr/sbin/asesetup` command on the system.
4. Run the `asemgr` utility on an existing member system and add the initialized system to the ASE.

2.6.2 Initializing All the Member Systems

Initializing all the member systems returns the ASE to a state that includes no member systems or services. After you do this, you must add the member systems and set up your services again.

To initialize all the member systems in an ASE, follow these steps:

1. If possible, use the `asemgr` utility to display the status of the member systems, network, and services in the ASE. This information will help you to re-create your ASE.
2. If possible, use the `asemgr` utility to delete all the services from the ASE. This allows you to save any Logical Storage Manager (LSM) or Advanced File System (AdvFS) disk configurations on a specific system.
3. Delete the `/usr/var/ase/config/asecdb` ASE database file from all the systems.
4. Invoke the `/usr/sbin/asesetup` command on each system.
5. Run the `asemgr` utility on a system, add the other initialized systems to the ASE, one at a time, and set up your services.

2.7 Stopping and Restarting ASE Activity

To change your available server environment (ASE) hardware configuration or perform maintenance, you may have to stop all activity in the ASE.

To stop all ASE activity, follow these steps:

1. Use the `asemgr` utility to place each ASE service off line, stopping the services.
2. Invoke the `/sbin/init.d/asemember stop` command on all the member systems.

After you stop ASE activity, you can perform the desired maintenance.

To restart ASE activity, follow these steps:

1. Invoke the `/sbin/init.d/asemember start` command on all the member systems.
2. Use the `asemgr` utility to place the ASE services on line.

2.8 Shutting Down a Cluster Member (PS)

Because each cluster member must maintain a kernel state regarding the clusterwide activities of the connection manager and distributed lock manager (DLM), you cannot shut a cluster member down to single-user mode and then bring it back up to multiuser mode. A full halt or complete reboot is required.

All normal methods of shutting down a single system and rebooting work for a cluster member. That is, the `shutdown -h` and `shutdown -r` commands (and halt and reboot console operations) work normally for systems running TruCluster Production Server Software with one exception.

On multiprocessing systems that are cluster members, pressing the halt button (or typing Ctrl/P at an AlphaServer 8200/8400 system console) does not cause a full halt of the member. To bring a multiprocessing system in a cluster to a full halt, enter one of the following console commands immediately after pressing the halt button:

- Initialize the console by using the console's `init` command. This stops all CPUs and resets all buses and is the quickest and surest way to bring the system to a full halt.
- Halt each CPU by using the console's `halt` command. If you wish to halt all the CPUs in order to examine hardware registers or memory locations, type `halt 1`, `halt 2`, ... This prevents corruption of system data and guarantees that the MEMORY CHANNEL hardware will time out so that other cluster members will realize the member is down. To force a crash dump, use the appropriate console command. (This will safely halt all CPUs and generate a crash dump at the next boot.)
- Reboot the system by using the console's boot command.

3

Managing Networks in a TruCluster Software Configuration

Member systems in a Production Server cluster employ networks in the following two ways:

- The available server environment (ASE) daemons on cluster members communicate with each other using a single network associated with the MEMORY CHANNEL interconnect. This is how the ASE infrastructure uses the network.
- Clients access ASE services over networks. This is how ASE services use networks.

The network used by the cluster infrastructure can be made more reliable by configuring redundant MEMORY CHANNEL connections between member systems, as described in the TruCluster Software Products *Hardware Configuration* manual. If one MEMORY CHANNEL connection fails, the daemons will communicate over the other MEMORY CHANNEL connection, which maintains cluster operation. The MEMORY CHANNEL network still appears to the ASE infrastructure as a single network.

Only when a member system cannot access other member systems over either MEMORY CHANNEL connection can a full network partition occur. If a full network partition occurs, the services continue to run on the member system and can be automatically failed over if the system crashes, but you cannot use the `asemgr` utility to change the ASE or to manually relocate services until the full network partition has been resolved.

Member systems in an Available Server configuration also use networks in two ways:

- The ASE daemons on ASE members communicate with each other using a single network interconnect (Ethernet, FDDI, or ATM) to which all members are attached. An Available Server configuration does not use the MEMORY CHANNEL interconnect.
- As in a Production Server cluster, clients access ASE services over networks. This is how ASE services use networks.

Using multiple networks in an Available Server configuration has the following advantages:

- It allows you to increase the availability of applications and data. When you add a member system to an ASE, the `asemgr` utility prompts you for network interfaces for the member system. Configuring multiple network paths between member systems in an Available Server configuration reduces the chance that a member system will be erroneously considered unavailable.

  Networks and shared SCSI buses are used to query member systems and determine their viability. If you configure multiple network connections between member systems, instead of querying only over the network associated with the member system name, TruCluster Available Server Software queries over backup networks. Use the `asemgr` utility to specify primary and backup networks.

- If a network or network interface fails, member systems can still communicate over another network path, and ASE operation is not impaired. The ASE daemons on the member systems communicate using Transmission Control Program (TCP) connections over the network interface associated with the member system name. If you configure multiple network paths between member systems, and the path for the network interface associated with a member system name fails, the daemons will communicate over a backup network, which maintains ASE operation.

  Only when a member system cannot access other member systems over any of its configured network interfaces does a full network partition occur. For example, a full network partition occurs if only one network is used in an ASE and that path fails, or if more than one network is used and all the paths fail. If a full network partition occurs, the services continue to run on the member system and can be automatically failed over if the system crashes, but you cannot use the `asemgr` utility to change the ASE or to manually relocate services.

In either a Production Server cluster or an Available Server configuration, client access to ASE services over networks can be made more reliable by monitoring specific network interfaces and taking specific actions (such as relocating services) when a particular interface fails. Monitor an interface if you are concerned with client access to ASE services on a particular interface. Monitoring an interface allows you to customize ASE operation when a network interface fails. See Section 3.3.5 for a discussion of how to monitor network interfaces.

3.1 Network Requirements

The network requirements for Production Server clusters are as follows:

- Production Server clusters support only the MEMORY CHANNEL interconnect as a primary intracluster network.
- Available server environment (ASE) member systems must be on at least one Internet Protocol (IP) network subnet that is common to the ASE.
- Network interface names for common networks must be included in the local `/etc/hosts` file on each member system. See Section 3.2 for information.

The network requirements for an Available Server configuration are as follows:

- TruCluster Available Server Software supports only Ethernet, FDDI, and ATM network hardware.
- Member systems must be on at least one common IP network subnet.
- Your primary and backup networks should be set up before you set up the TruCluster Available Server Software hardware and software.
- Primary and backup networks in an Available Server configuration must be subnets that are common to all member systems. Network interface names for common networks must be included in the local `/etc/hosts` and `/etc/routes` files on each member system. See Section 3.2 for more information.
- If the primary network connected to the systems becomes saturated, TruCluster Available Server Software operation is impaired. If you receive messages indicating that you are out of `mbufs`, adjust the `ubcmaxpercent` and `ubcminpercent` parameters in the configuration file. See the DIGITAL UNIX *System Configuration and Tuning* manual for more information.

3.2 Defining Network Interfaces

When you add a member system to an available server environment (ASE) in either a Production Server cluster or an Available Server configuration, the `asemgr` utility prompts you for additional network interface names. In an Available Server configuration, before you add an interface, you must use the `netsetup` utility to define the network interface on the system. The Production Server installation script automatically defines the network interface for the MEMORY CHANNEL interconnect on the system, so there is no need to run the `netsetup` utility to define its interface.

The following example is part of an /etc/hosts file on a Production Server cluster and shows two member systems, gideonmc and totomc, and multiple network interfaces for the systems:

```
# Cluster member systems (MEMORY CHANNEL interconnect)
#
10.0.0.1 gideonmc.abc.def.com gideonmc
10.0.0.2 totomc.adc.def.com totomc
#
#
# FDDI ring #1 (Client network 1)
#
16.142.112.121 gideonfddi1.abc.def.com gideon1
16.142.112.122 totofddi1.abc.def.com toto1
#
# FDDI ring #2 (Client network 2)
#

16.142.96.121 gideonfddi2.abc.def.com gideon2
16.142.96.122 totofddi2.abc.def.com toto2
```

The following example is part of an /etc/hosts file on an Available Server configuration and shows two member systems, gideon and toto, and multiple network interfaces for the systems:

```
# ASE member systems
#
16.140.64.121 gideon.abc.def.com gideon
16.140.64.122 toto.adc.def.com toto
#
#
# FDDI ring #1
#
16.142.112.121 gideon1.abc.def.com gideon1
16.142.112.122 toto1.abc.def.com toto1
#
# FDDI ring #2
#
16.142.96.121 gideon2.abc.def.com gideon2
16.142.96.122 toto2.abc.def.com toto2
```

In an Available Server configuration, you must specify the interface names for the primary and backup networks in the local /etc/routes file on each member system. For each member system, you must define a host route to all other member systems. This definition is needed to fail over IP traffic between member systems when a network path fails.

For example, if your member systems are `gideon1` and `toto1`, where the number in the name refers to the subnet, and each member system also has interface names `gideon2` and `toto2`, then each member system's `/etc/routes` file must contain the following information:

```
-host gideon1 gideon1
-host gideon2 gideon2
-host toto1 toto1
-host toto2 toto2
```

See `routes`(4) for information on the file format.

3.3 Modifying the Network Configuration

To modify the network configuration, choose the "Modify the network configuration" item from the Managing the ASE menu.

The ASE Network Modify Menu allows you to do the following:

- Display the current network configuration (Section 3.3.1)
- Add and delete network interfaces (Section 3.3.2)
- Specify the primary network (Available Server configurations only) (Section 3.3.3)
- Specify backup networks (Available Server configurations only) (Section 3.3.3)
- Specify networks to ignore (Available Server configurations only) (Section 3.3.4)
- Specify network interfaces to monitor (Section 3.3.5)

The following sections describe how to display and modify the network configuration.

3.3.1 Displaying the Network Configuration

Choose the "Show the current configuration" item from the ASE Network Modify Menu to display the member systems, their interface names, whether monitoring is enabled, or, in an Available Server configuration, whether an interface is connected to a primary or a backup network.

The following example shows the network configuration of a Production Server cluster:

```
                ASE Network Configuration

    Member Name          Interface Name       Monitor
    ___________          ______________       _______
    totomc               totomc               No
    totomc               totofddi1            Yes
    totomc               totofddi2            No

    gideonmc             gideonmc             No
    gideonmc             gideonfddi1          Yes
    gideonmc             gideonfddi2          No
```

The following example shows the network configuration of an Available Server configuration:

```
                ASE Network Configuration

    Member Name          Interface Name       Member Net  Monitor
    ___________          ______________       __________  _______
    toto                 toto                 Primary     Yes
    toto                 toto1                Backup      No
    toto                 toto2                Backup      No

    gideon               gideon               Primary     Yes
    gideon               gideon1              Backup      No
    gideon               gideon2              Backup      No
```

3.3.2 Adding and Deleting Network Interfaces

Before you specify a network interface for a member system, the interface must be defined and configured on the system. (See Section 3.2 for more information.)

Choose the "Add network interfaces" item from the ASE Network Modify Menu to add a network interface. From the ASE Member Menu, choose the number of the member to which you want to add a network interface. For example, on a Production Server cluster:

```
                ASE Member Menu

Select a member to add an interface to:

    0)  gideonmc
    1)  totomc

    q)  Quit without making changes

Enter your choice: 1
Enter interface names for member 'totomc'
    Interface name (return to exit): totofddi1
```

To delete network interfaces, choose the "Delete network interfaces" item from the ASE Network Modify Menu. For example, on an Available Server configuration:

```
                ASE Member Menu

Choose a member to delete an interface from:

    0)  gideon
    1)  toto

    q)  Quit without making changes

Enter your choice: 1

         Network Interfaces for Member 'toto'

Choose one or more network interfaces to delete:

     )  toto     16.142.112.121       Not an option
    1)  toto1    16.142.112.122
    2)  toto2    16.142.96.122

    q)  Quit to previous menu

Enter your choices (comma or space separated):1
```

In an Available Server configuration, note that the member network interface cannot be deleted, regardless of whether it is defined as a primary or backup network. The member network interface is defined during software installation and establishes a system's membership in an ASE and its member name. For that reason, the member network interface name does not appear in the list of interfaces eligible for deletion. Similarly, in a Production Server cluster, you cannot delete a member's MEMORY CHANNEL interface.

3.3.3 Specifying Primary and Backup Networks (AS)

In an Available Server configuration, the primary network is the network that is used most frequently to query other member systems. Backup networks are also used for queries, but at a slower rate. Interfaces for primary and backup networks must be common to all the member systems and included in each member system's local `/etc/hosts` and `/etc/routes` files. See Section 3.2 for more information.

Choose the "Specify the primary ASE member network" item from the ASE Network Modify Menu to select an interface for the primary network. For example:

```
             ASE Member Primary Network Menu

Choose one of the networks to be the ASE member primary network:

    0)  16.142.112.0     (toto1, gideon1)
    1)  16.142.96.0      (toto2, gideon2)

    q)  Quit to previous menu

Enter your choice: 0
```

Choose the "Specify a backup ASE member network" item from the ASE Network Modify Menu to select backup network interfaces for the ASE. For example:

```
                ASE Member Backup Network Menu

Choose one of the networks to be the ASE member backup network:

    0)  16.142.112.0     (toto1, gideon1)
    1)  16.142.96.0      (toto2, gideon2)

    q)  Quit to previous menu

Enter your choices (comma or space separated): 1

        16.142.96.0      (toto2, gideon2)

Are the above choices correct (y|n)? [y]: y
```

3.3.4 Specifying a Network to Ignore (AS)

In an Available Server configuration, choose the "Specify an ASE member network to be ignored" item from the ASE Network Modify Menu to specify a network that you want to configure but you do not currently want the member system to use. For example:

```
                Ignore ASE Member Network Menu

Choose a network not to be used as an ASE member network:


    0)  16.142.112.0     (toto1, gideon1)
    1)  16.142.96.0      (toto2, gideon2)

    q)  Quit to previous menu

Enter your choices (comma or space separated): 0

        16.142.112.0     (toto1, gideon1)

Are the above choices correct (y|n)? [y]:n
```

3.3.5 Monitoring Network Interfaces

You can monitor any network interface or any member system and take specific actions (such as relocating services, or sending mail or a page to an administrator) when a particular interface fails. Monitor those interfaces that are critical to clients accessing services. TruCluster software allows you to monitor up to four interfaces per member system at the same time.

In a Production Server cluster, you can monitor a member’s Memory Channel interfaces or any network interface that allows client access to the cluster’s services. Similarly, in an Available Server configuration, you can monitor a member’s primary and backup network interfaces. In either a cluster or an Available Server configuration, you can monitor a network interface that is not on a subnet common to all member systems.

If a monitored network interface fails, the TruCluster software runs the error Alert script (see Section 12.1.5), which invokes the member’s `/var/ase/lib/ni_status_awk` script. By default, if all monitored network interfaces on the member are down, the `/var/ase/lib/ni_status_awk` script stops all the services running on that member and starts them on another member.

However, you can customize the `/var/ase/lib/ni_status_awk` script on each member system to specify a different action to take. For example, you can edit the script so that services relocate to another member system if any network interface fails or if a particular interface fails. In addition, because the error Alert script is propagated on all the member systems, you can edit the error Alert script itself, so that the actions will be the same on all systems. Use the `asemgr` utility to edit the error Alert script.

Choose the "Specify network interfaces to be monitored" item from the ASE Network Modify Menu to monitor specific interfaces. For example, on an Available Server configuration:

```
                ASE Member Menu

Choose a member to modify:

    0)  gideon
    1)  toto
    q)  Quit without making changes

Enter your choice: 0

           Network Interfaces for Member 'toto'

Choose one or more network interfaces:

    0)  toto           16.140.64.122         (monitored)
    1)  toto1          16.140.112.122        (monitored)
    2)  toto2          16.140.96.122         (monitored)

    q)  Quit to previous menu
    n)  Do not monitor any interfaces

Enter your choices (comma or space separated): 1

        toto1  16.140.112.122

Are the above choices correct (y|n)? [y]: y
```

3.4 Using Multiple Client Networks and ASE Services

In either a Production Server cluster or an Available Server configuration, member systems, Network File System (NFS) services, tape services, and disk services that have IP addresses use the networking subsystem. The following sections apply only if your member systems are connected to more than one client network. If subnets are used, the term network is used in the following sections to refer to a subnet.

3.4.1 How ASE Services Use Multiple Networks

You can connect the member systems in either a Production Server cluster or an Available Server configuration to several client networks. All the member systems must be able to access each network, so that clients react correctly when the TruCluster software relocates an NFS or tape service (or a disk service that has an IP address).

Between the networks, there should be a separate router system that is not a member system. Do not use a member system as a general-purpose IP router, because system performance will be unpredictable.

To enable clients to access an NFS or tape service (or disk service that has an IP address), the service name is assigned its own Internet address. The service name that you choose must be native to one of the networks. On that network, the **Address Resolution Protocol** (**ARP**) translates the Internet address associated with the service name to the hardware address of the member running the service. If the service is relocated, the ARP translates that Internet address to the Ethernet address of the new server. Therefore, the ARP broadcasts enable clients to recognize when a service has relocated to a different member system.

After they receive the new ARP address translation, clients on the network that is native to the service name will start to send data to the new member system that is running the service. Clients on a network that is not native to the service name forward their packets through the router system to the network that is native to the service name.

The router system processes the ARP broadcasts sent from the member systems. Clients that are not on the native network should know how to send data through a router to the service name address. Clients that are on the native network only need to know how to react to ARP broadcasts.

3.4.2 Getting Faster Access to Services

If a client is not on the network that is native to an NFS or tape service name (or a disk service name that has an IP address), the client must send packets through the router system to reach the service address. Network traffic must go through an extra hop to access the service because the packets are forced to pass through an extra system.

A TruCluster Available Server Software feature that enables you to bypass this step requires that clients on a network that is not native to the service name use the Routing Information Protocol (RIP) routing protocol and respond to host routes. This feature broadcasts host routes on networks that are not native to the service name. The technique that ARP uses to handle service relocations is still used on the network that is native to the service name.

Host routes direct the clients to the member system that is running the service, without requiring clients to send data through the router system. Using this method, member systems do a restricted form of routing. Only host routes associated with NFS, tape, or disk service names are advertised.

You must manually enable this feature. If this feature is not enabled, and you have multiple networks and a separate router system, clients on all networks will react correctly to service relocations. However, some network traffic will require an extra step to reach the service.

To enable this feature, perform the following tasks:

1. Run the `netsetup` script on all systems.
2. Choose the "Enable/Disable Network Daemons and Add Static Routes" menu items.
3. Enter `yes` at the prompt that asks if you want to be an IP router.
4. Choose the `gated` option and do not specify any flags.
5. Choose the "Exit" menu item.
6. Do not restart the network services.
7. Kill the `routed` daemon if it is running.
8. Enter the following command:

   ```
   # rcmgr set ASEROUTING yes
   ```

The `ASEROUTING` configuration variable allows for host-based routes from a server that has multiple network interfaces. This can make for faster connections to a service by avoiding routers and making use of the multiple interfaces.

However, only clients that can can listen to dynamic routing updates sent with the RIP routing protocol will benefit from setting `ASEROUTING=yes`. Clients that simply specify a default router will not benefit.

Note

> You can use the `ASEROUTING` configuration variable only with the old `gated` daemon (`ogated`). (`ogated` is the default selection in the `netsetup` script.) If you use the `ASEROUTING` configuration variable when the new `gated` daemon is running on ASE members, all service operations will fail and error messages are entered in the `daemon.log` file.

Setting `ASEROUTING` to `yes` results in modifications to the `/etc/ogated.conf` files on all ASE members. If you have modified the `/etc/ogated.conf` files on ASE members, these changes might interfere with `ASEROUTING` behavior. Therefore, if you customize the `/etc/ogated.conf` files on ASE members, do not use the `ASEROUTING` option.

If you created services before you enabled this feature, you must modify all the services; this will delete the services, add the services, and start the advertising of the host-based routes.

4 Preparing to Set Up Highly Available Services

This chapter describes the preparation required to set up the services in an available server environment (ASE) within your TruCluster configuration. Later chapters describe how to set up and manage specific types of ASE services.

Specifically, this chapter discusses the following topics:

- How to add a service (Section 4.1)
- How to use the Automatic Service Placement (ASP) policies (Section 4.2)
- How to use disk quotas in an ASE (Section 4.3.3)
- How to use the UNIX File System (UFS) in your services (Section 4.3.5)
- How to use the Advanced File System (AdvFS) in your services (Section 4.3.6)
- How to use the Logical Storage Manager (LSM) in your services (Section 4.3.7)
- How to install the applications you want to fail over in an ASE (Section 4.4)
- How to use your own action scripts (Section 4.5)

This chapter also lists the steps you follow to ensure that you have performed all the necessary service preparation tasks.

After you perform the preparatory tasks, you can use the `asemgr` utility to add the service to the ASE. After you specify the necessary service information, the TruCluster software updates the ASE database, propagates the database changes to all members of the ASE, and adds the service to all the members. The TruCluster software then chooses a member to run the service and starts the service on that member.

4.1 Preparing to Add a Service

Before you add a service, you must plan on how you want to set up the service and perform some preparatory tasks. Not all services require that you perform all the tasks. For example, some services consist of only a disk configuration or only an application.

To ensure that you are ready to use the asemgr utility to set up a service, follow these steps:

1. Prepare the disk configuration for the service. Depending on the service, you may have to install and set up specific software, such as UNIX File Service (UFS), Network File System (NFS), Advanced File System (AdvFS), or Logical Storage Manager (LSM) on all the member systems. If you want to use disk quotas, you must set up the quota files.
2. Install the application that you want the service to fail over to. Any application that you use in the service must be installed on all the member systems.
3. Create the action scripts that ASE will use to fail over the application. At a minimum, you must create start and stop action scripts for the service.
4. Determine if you want to restrict the service so that it runs only on specific member systems. A service's Automatic Service Placement (ASP) policy determines which members can run the service.

After you perform the previous tasks, use the asemgr utility to add the service to the ASE. The following sections describe ASP policies, how to prepare a disk configuration, and how to use action scripts. The following chapters describe how to add a specific type of service.

4.2 Understanding ASP

The Automatic Service Placement (ASP) policy for a service determines which member systems are allowed to run the service. You must choose an ASP policy for each service. The asemgr utility prompts you for the ASP policy when you add a service.

Whenever the TruCluster software automatically starts a service (for example, when a service relocates because of a member system failure), it chooses a member system to run the service. The member system it chooses depends on the service's ASP policy and which member systems are available.

Note

You can manually relocate a service by using the asemgr utility to override the ASP policy.

You can use a service's ASP policy to specify a master/standby configuration, where one member runs all the applications and the other member is used only if the master system fails. You can also specify that any member system can run the service. This ASP policy ensures that the services are distributed equally among the members.

When you add a service, the `asemgr` utility prompts you for the service's ASP policy. There are three ASP policies:

- Balanced Service Distribution—If you choose this policy, the TruCluster software considers the member that is running the least number of services the member most favored to run the service. Using this policy, services are distributed equally among the members.
- Favor Members—If you choose this policy, the TruCluster software prompts you for a list of available server environment (ASE) members. The first member on the list is the member most favored to run the service. If that member is unavailable, the second member on the list is the most favored member. If all the members on the list are unavailable, the member that is running the least number of services is the most favored member. This accommodates the previously mentioned master/standby configuration.
- Restrict to Favored Members—This policy is similar to the Favor Members policy. If you choose this policy, you are prompted for a list of ASE members. However, unlike the Favor Members policy, if all the members on the list are unavailable, the TruCluster software will not start the service. This policy ensures that the service will never run on a member that is not on the list, unless you manually relocate the service to that member.

After you choose one of the previous ASP policies, you must specify if you want the TruCluster software to relocate a service to a more highly favored member if it becomes available. For example, if the most highly favored member fails, its services relocate to another member. If the most highly favored member system becomes available again, depending on the option you choose, the TruCluster software can either relocate the service back to the original member system or keep the service running on the current member.

Also, if you specify only one favored member, you must indicate if you want the TruCluster software to change the one favored member to the member that is specified when you manually relocate the service. If you choose this option and manually relocate the service, the member to which you relocated the service is now the only favored member for the service. For example, if you specify that a service can run only on `member1` and you manually relocate the service to `member2`, the TruCluster software can make `member2` the only member on which the service can run.

Example 4–1 shows how to select the Favor Members ASP policy and specify two favored members.

Example 4–1: Selecting the Favor Members ASP Policy

```
        Selecting an Automatic Service Placement (ASP) Policy

Select the policy you want ASE to use when choosing a member
to run this service:

    b)  Balanced Service Distribution
    f)  Favor Members
    r)  Restrict to Favored Members

    x)  Exit to Service Configuration      ?)  Help

Enter your choice [b]: f

        Selecting an Automatic Service Placement (ASP) Policy

Select the favored member(s) IN ORDER for service ’disk1’:

    1)  gideon
    2)  toto
    3)  nobrain

    x)  No favored members              ?)  Help

Enter a comma-separated list [x]: 2,1

        Selecting an Automatic Service Placement (ASP) Policy

Do you want the ASE to relocate this service to a more highly
favored member if one becomes available while the service
is running (y/n/?): y
```

4.3 Using Disks in ASE Services

If you are setting up a distributed raw disk (DRD), Network File System (NFS), disk, or tape service, you must install and set up the disk configuration that will be used in the service. Note that user-defined services do not utilize disks. In general, you set up the disk configuration in the same way as in an environment other than an ASE.

However, once the disk is used in an ASE service, it must be managed within the ASE. The ASE must control the disk if a service is running. For example, do not manually unmount a file system that is being used by an online ASE service.

The following sections contain general information about using disks in the ASE, in addition to information about using the UNIX File System (UFS), NFS, and the Logical Storage Manager (LSM).

4.3.1 General Disk Requirements

The following requirements apply to all disk configurations:

- A disk cannot be used in more than one service, because a service must have exclusive access to a disk. When you use a disk in a service, use the entire disk. This requirement also applies to disks that are divided into partitions and used in AdvFS domains and **LSM disk groups**.
- Use the least possible number of disks per service. Using a small number of disks in a service is recommended because, if a disk fails, the service stops. However, if you use LSM or **RAID** to mirror the disks, only a complete mirrored volume failure (both disks in the mirrored volume fail) will cause the service to stop.
- Make sure that only the ASE service processes are accessing the disk. A service's **mount point** can be used only for the ASE service and should not have any other use in the system. See Section 4.3.2 for more information.
- Do not locally mount the disks that you will use in a service; the TruCluster software mounts them for you.
- Do not unmount a disk used in an online service.
- If a network disconnection occurs or if I/O to a nonmirrored disk fails, a service cannot be stopped. The TruCluster software will reboot the member running the service.

The following sections describe how to control access to disks and how to use UFS, AdvFS, and LSM in your services.

4.3.2 Controlling Access to Disks

You must be able to control the processes that access the disks used in your services. This is because the TruCluster software must be able to stop a service in order to relocate a service, place a service off line, or modify a service. If the TruCluster software is unable to stop a service, ASE operation may be impaired.

Notes

If a service fails to stop, the `asemgr` utility prompts you with a number of choices, so you can fix the problem. See Chapter 10 for more information.

If, while attempting to fail over a service, an ASE member cannot stop the service, the member will reboot itself. The rebooting sequence allows another member to provide the service. In addition, it often corrects the situation that prevented the member that originally owned the service from stopping the service.

If the TruCluster software cannot stop a service that uses mounted disks, you may see the following error message in the `daemon.log` file:

```
device busy
```

This message indicates that the TruCluster software could not unmount a disk, possibly because it could not stop all the processes accessing the disk.

The TruCluster software cannot stop a service that uses mounted file systems, filesets, or volumes unless it can unmount them. The TruCluster software may not be able to unmount a disk in the following situations:

- A process that accesses the disk was invoked by the service's start action script, but the service's stop action script does not invoke a command to stop the process.
- A process that the start action script did not start and that is unrelated to the ASE is accessing the disk. This could occur if a user logs in to the system on which the file system is locally mounted and changes directory to the mount point.

You must ensure that only processes started and stopped by the service's action scripts can access the disks used in a service. Make sure that all the processes invoked by the start action script are stopped by the stop action script. The actions in the start script must be reversed by the actions in the stop script.

If you want to allow users to access the mounted disks, use an NFS service and allow users to access only the directory that is exported, not the directory that is mounted locally.

To enable access through NFS, create an entry in each member system's `/etc/fstab` file and specify the service name as the remote host from which to NFS-mount the file system. You must also specify the exported file system path.

If you must allow access to the local mount point, ensure that the service's stop action script is able to stop these processes.

4.3.3 Using Disk Quotas

You can enable disk quotas on file systems or filesets that are used in an ASE service. Quotas allow you to limit the number of blocks and inodes (or files) that a user or a group of users can allocate. You can set a separate quota for each user or group of users on each file system. See the DIGITAL UNIX *System Administration* manual for detailed information about UFS and AdvFS disk quotas.

To use disk quotas in an ASE, you must mount the /proc file system on each member system. Add the /proc file system to the /etc/fstab file as follows:

```
/proc      /proc   procfs rw
```

When you add a service that uses disk quotas, the TruCluster software creates a /var/ase/config/fstabs/*service_name* file on each member system. The file contains records that describe each mount point used by the service. The records use the format of the /etc/fstab file.

When a service is started on a member, the TruCluster software creates a link to the /var/ase/config/fstabs/*service_name* file in the /var/ase/config/fstabs.running directory. The TruCluster software removes the link when it stops the service. Therefore, only one member system has a pointer to the file at any time.

There are two methods you can use to set up disk quotas on a UNIX file system in the ASE:

- You can:

    1. Use the quotacheck command to set up the quota.user or quota.group file. You can place these files in the root of the file system or in some other location.
    2. Use the asemgr utility to specify the file system in a service. When prompted for quota files, specify the pathname of the quota.user or quota.group file.

- Alternatively, you can:

    1. Use the asemgr utility to specify the file system.
    2. When prompted for quota files, specify the pathname of the quota.user or quota.group file. You can place these files in the root of the file system or in some other location.
    3. After you start the service, use the edquota command to specify the quota limits in the files. You must execute this command on the system that is running the service.

If you use the mkfset command to create an AdvFS fileset, the command automatically sets up quota files in the root of the fileset. You cannot specify another location. To set up quotas on an AdvFS fileset in the ASE, use the asemgr utility to specify the fileset and, when prompted for quota files, specify the quota.user or quota.group file. After you add the service, you must use the vedquota command to specify the quota limits in the files.

Place the quota files on a file system used in the service, so the files are relocated with the service. For AdvFS, you must locate the quota files on the fileset. If you are using UFS and want to specify another location, you must manage the quota files separately on each system. Do not locate quota files on a file system that is used by another service.

To display the mount options for a file system or fileset, including whether quotas are enabled, use the asemgr utility to display service status.

To change the quota files or disable quotas for a file system or fileset, modify the service and select the file system or fileset. Then, choose to modify the quota options. To disable quotas, specify none when prompted for the quota files.

4.3.4 Backing Up Disks

There are three methods to back up a disk that is used in an ASE service:

- Use the asemgr utility to relocate the service to a specific member so the service will not move. You must perform the backup or restore from this member.
- Use the asemgr utility to place the service that uses the disk off line, which stops the service. Perform the backup from any member system. Note that if you are using AdvFS or LSM disks, they must be configured on the system from which you are performing the backup.
- Use POLYCENTER NetWorker Save and Restore for DIGITAL UNIX to back up a disk that is used in an NFS service or in a disk service that uses an IP address. NetWorker treats an ASE service as an independent client and stores the storage indexes under the name of the service. This enables you to back up and recover the service's storage independent of the member system running the service. See the NetWorker documentation for information about using NetWorker to back up an ASE service's storage.

4.3.5 Using UFS

To use the UNIX File System (UFS) in the ASE, you first set up the disks as you would for any UFS before you add the service. That is, you must use the disklabel command to partition a disk and the newfs command to create a file system on a partition. See the DIGITAL UNIX *System Administration* manual for information about setting up UFS.

When you use the asemgr utility to add an NFS, disk, or tape service that uses UFS, you are prompted for the following information:

- Device special file name (for example, /dev/rz2c)
- Mount point or export directory
- Netgroup or system name to which the service is restricted (optional and only for NFS services)
- Access mode (either read/write or read-only)
- Additional mount options other than the default options specified in mount(8).

Example 4–2 shows how to specify a UFS when setting up a disk service.

Example 4–2: Specifying a UFS in a Disk Service

```
                     Specifying Disk Information

Enter one or more device special files, AdvFS filesets, or LSM
volumes to define the disk storage for this service.

    For example:    Device special file:     /dev/rz3c
                    AdvFS fileset:           domain1#set1
                    LSM volume:              /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.

Enter a device special file, an AdvFS fileset, or an LSM volume as
storage for this service (press ’Return’ to end): :/dev/rz20c

                  Mount Point

The mount point is the directory on which to mount ‘/dev/rz20c‘.
If you do not want it mounted, enter "NONE."

Enter the mount point or NONE: /usr/dbase


              UFS File System Read-Write Access

Mount ‘/dev/rz20c‘ file system with read-write or read-only access?

    1)  Read-write
    2)  Read-only

Enter your choice [1]: 1

You may enable user and group quotas on this file system by specifying
names for the quota files.  If you accept the default quota file names,
the quota assignments you make with edquota will relocate with the file
system.  Enter "none" to disable quotas.

User quota file [/dbase/quota.user]:  Return

Group quota file [/dbase/quota.group]:  Return
              UFS Mount Options Modification

Enter a comma-separated list of any mount options you want to use
```

Example 4–2: Specifying a UFS in a Disk Service (cont.)

```
for '/dev/rz20c' (in addition to the UFS-specific defaults listed
in the mount.8 reference page).  If none are specified, only the
default mount options are used.

Enter options (Return for none):  Return
```

4.3.6 Using AdvFS

You can use the Advanced File System (AdvFS) in an NFS service or a disk service. AdvFS provides you with fast file system recovery. If you want to use AdvFS, it is important that you read the following documentation:

- The POLYCENTER Advanced File System Utilities *Installation Guide* describes how to install the POLYCENTER Advanced File System Utilities software. The POLYCENTER Advanced File System Utilities software is a separately licensed product.
- The DIGITAL UNIX *System Administration* manual and `advfs`(4) provide an introduction to AdvFS. Also see the AdvFS online help.

To use AdvFS in an ASE, you first set up AdvFS in the same way that you would in an environment other than ASE before you add a service. You must set up the volumes, domains, and filesets that you will use in the service. Do this on the same member system on which you will run the `asemgr` utility to add the service to the ASE.

The following example shows how to create a domain named `dom1` on the volume `/dev/rz10c`, and then create a fileset named `set1` on `dom1`:

```
# mkfdmn /dev/rz10c dom1

# mkfset dom1 set1
```

The following sections describe the AdvFS requirements and how to specify AdvFS information with the `asemgr` utility.

4.3.6.1 AdvFS Requirements

AdvFS has the following requirements:

- You must set up the AdvFS volumes, domains, and filesets on the same member on which you will run the `asemgr` utility to add the service to the ASE.
- A service can use more than one AdvFS domain, but a domain cannot be used by more than one service.
- The AdvFS domain names must be unique in the ASE.

- To modify an AdvFS configuration that a service uses, the disks must be configured on the system on which you make the modifications. See Chapter 10 for information.
- When you delete a service that uses AdvFS, the `asemgr` utility prompts you for a member on which to leave the AdvFS domain configured. This enables you to use the storage configuration again.
- If you create an NFS, disk, or tape service that uses AdvFS and choose not to have the TruCluster software automatically mount the filesets, a member system may panic unless the following conditions are met:
  - Before you add the service, make sure that the fileset is not already mounted.
  - If you mount a fileset in your own user-defined action scripts, make sure that the user-defined stop action script unmounts the fileset and returns an error code if the unmount fails.

4.3.6.2 Specifying AdvFS Filesets in a Service

If you use AdvFS in a service, you are prompted for the following information when you add the service:

- Fileset name (for example, `dom1#fset2`)
- Mount point or export directory (optional for disk services)
- Netgroup or machine name to which the service is restricted (optional and only for NFS services)
- Access mode (either read/write or read-only)
- Mount options other than the default options specified in the `mount`(8) reference page

When you specify a fileset, the `asemgr` utility displays the disk partitions and any **LSM volumes** that comprise the fileset.

Example 4–3 shows how to specify an AdvFS fileset that uses an LSM volume.

Example 4–3: Specifying AdvFS Filesets in an NFS Service

```
                 Specifying Disk Information

Enter one or more device special files, AdvFS filesets, or LSM
volumes to define the disk storage for this service.

For example:        Device special file:      /dev/rz3c
                    AdvFS fileset:            domain1#set1
                    LSM volume:               /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.
```

Example 4–3: Specifying AdvFS Filesets in an NFS Service (cont.)

```
Enter a device special file, an AdvFS fileset, or an LSM volume as
storage for this service (press ’Return’ to end): dom1#set1

ADVFS domain ‘dom1‘ has the following volume(s):

     /dev/vol/dg3/vol01

Is this correct (y/n) [y]: y

Following is a list of device(s) and pubpath(s) for disk group dg3:

        DEVICE  PUBPATH

        rz32c   /dev/rz32c

Is this correct (y/n) [y]: y

Enter the directory pathname(s) to be NFS exported from the storage
area "dom1#set1".  Press ’Return’ when done.

Enter a directory pathname: /usr/staff

  Enter a host name, NIS netgroup, or IP address for the NFS
  exports list (press ’Return’ for all hosts): staff_group

Enter a directory pathname: Return


              AdvFS Fileset Read-Write Access

Mount ‘dom1#set1‘ fileset with read-write or read-only access?

    1)  Read-write
    2)  Read-only

Enter your choice [1]: 1

You may enable user and group quotas on this file system by specifying
names for the quota files.  If you accept the default quota file names,
the quota assignments you make with edquota will relocate with the file
system.  Enter "none" to disable quotas.

User quota file [/var/ase/mnt/ase4/usr/staff/quota.user]:
Return

Group quota file [/var/ase/mnt/ase4/usr/staff/quota.group]:
Return
              AdvFS Mount Options Modification

Enter a comma-separated list of any mount options you want to use for
the ‘dom1#set1‘ fileset (in addition to the defaults listed in the
mount.8 reference page).  If none are given, only the default mount
options are used.

Enter options (Return for none): bg
```

See Section 10.6 for information about modifying services that use AdvFS.

4.3.7 Using LSM

You can use Logical Storage Manager (LSM) volumes in an NFS, disk, or tape service. LSM provides high data availability for disk storage devices. It protects against data loss, improves disk I/O performance, and allows you to perform disk management functions without disrupting access to disks. If you want to use LSM volumes in an ASE service it is important that you read the DIGITAL UNIX *Logical Storage Manager* manual.

To use LSM, make sure that the LSM software is installed and initialized before you add the service. You must set up a `rootdg` disk group on each member system, set up the disk groups, and configure the LSM volumes that you will use in the ASE services. You can specify an LSM volume in a service, or you can create a UNIX file system or an AdvFS domain and fileset on top of an LSM volume, and use the file system or fileset in a service.

Note

> You must set up a service's LSM disk groups and volumes on the same member on which you will run the `asemgr` utility to set up the service.

If you delete a service that uses LSM, the `asemgr` utility prompts you for a member on which to keep the service's storage configuration. The disk groups will remain imported only on that system.

To set up LSM disk groups and volumes for a service, follow these steps:

1. Ensure that each member system has a `rootdg` disk group set up on local (nonshared) disks. Configure the `rootdg` disk group on more than one disk. This ensures that if one disk fails, the disk group is still available.
2. On one member, initialize the disks to be used in each disk group.
3. Create the disk groups for the service and add the initialized disks to it. Note that a service can use more than one disk group, but a disk group can be used only in one service.
4. Create the volumes for the disk groups.
5. Optionally, create UNIX file systems or AdvFS domains and filesets on the volumes.
6. Run the `asemgr` utility and add the service.

The following sections describe the LSM requirements and how to set up a new LSM disk configuration for an ASE service. However, you may also want to use a configuration that was used in a previous service or for some purpose other than ASE. See Section 4.3.7.5 for information about using an existing LSM configuration.

4.3.7.1 LSM Requirements

LSM has the following requirements:

- All ASE members must have the LSM software installed.
- All ASE members must have a `rootdg` disk group set up on local (nonshared) disks. Configure the `rootdg` disk group on more than one disk. This ensures that if one disk fails, the disk group is still available.
- Do not use the `rootdg` disk group in an ASE service because it cannot be relocated.
- All LSM disk group names in the ASE must be unique.
- LSM nopriv disks that are created as part of the LSM encapsulation process cannot be used in an ASE service. LSM encapsulation converts a disk partition that contains data into an LSM disk. However, if a nopriv disk is created in order to reduce the number of configuration copies in a disk group, use the following command syntax to create an LSM disk without a configuration copy or a kernel log copy:

  voldisksetup -i [*disk*] [nconfig=0 nlog=0]

  This command initializes the specified *disk* with a disk header that LSM recognizes when it is restarted, but without a configuration or kernel log copy. The disk then can be added to an LSM disk group to create volumes, and the volumes can then be used in an ASE service.
- Each disk group must have four to eight copies of the configuration and kernel log files.
- You must set up a service's disk groups and volumes on the same member on which you will run the `asemgr` utility to set up the service.
- When modifying a service's LSM configuration, the disk groups used in the service must be imported to the machine on which you will run the `asemgr` utility. LSM configuration changes can be made only on an imported disk group.
- You can modify the name of a service that uses LSM only if the service is on line. In addition, you must run the `asemgr` utility on the member that is running the service.
- A disk or disk group can be used in only one service, but a service can use more than one disk or disk group.
- When a failed or previously unavailable part of a mirrored volume becomes available, you can reincorporate the device into the service

without interrupting the service. To do this, resynchronize the mirrored volume outside of the ASE on the member to which the disk groups are imported. Then, rereserve the devices by using the `asemgr` utility's Advanced Utilities menu.

- If a service uses LSM mirrored volumes, do not modify the service while a mirrored volume is synchronizing (using `volplex att` or `volrecover`), because the synchronization will abort and then restart. The abort will not corrupt the volume, but it will delay the volume synchronization.
- If a disk that is included in a plex that is part of an LSM mirrored volume goes off line, the data in the volume is still available, as long as one complete plex of the volume remains on line. Thus, the service that uses the mirrored volume remains available if a disk fails. When the failed disk is replaced, you can reincorporate the device into the ASE service by resynchronizing the mirrored volume outside of the ASE and then rereserving the devices using the `asemgr` utility's Advanced Utilities menu.

  In rare cases, if any disk that is part of an LSM disk group is not accessible when the service is started, you may access stale data or a stale LSM configuration. This can occur if two member systems access different subsets of the disks in the disk group. By default, the TruCluster software prevents accessing stale data by disabling the feature that allows you to start an unsynchronized mirrored volume.

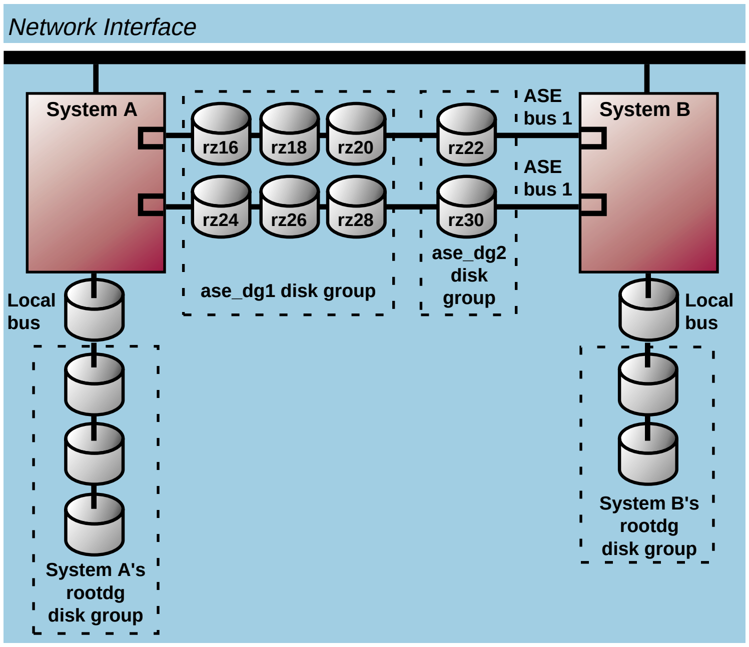
4.3.7.2 Initializing LSM

To use LSM in your ASE services, you must set up LSM and create a separate `rootdg` disk group on each member systems before adding services. See the DIGITAL UNIX *Logical Storage Manager* manual for information about using individual commands to customize the `rootdg` disk group initialization.

Use the `volsetup` command to create the `rootdg` disk group. When you create the `rootdg` disk group on each member system, you must specify disks only on the system's local SCSI bus. Do not specify disks that are on the ASE shared bus.

See Figure 4–1 for an example of an LSM disk configuration that uses two disk groups, `ase_dg1` and `ase_dg2`, on two shared buses.

Figure 4–1: Logical Storage Manager Disk Configuration



ZK-1061U-AI

4.3.7.3 Configuring LSM Disks and Disk Groups for the ASE

You must initialize the LSM disks and include them in disk groups that are set up specifically for ASE. The DIGITAL UNIX *Logical Storage Manager* manual describes how to determine the private region size and parameters for a disk group, how to initialize disks for LSM, and how to set up and add disks to disk groups. Figure 4–1 shows that disks on the two ASE shared buses are included in two shared disk groups, `ase_dg1` and `ase_dg2`.

The `voldisksetup` utility sets up a disk for LSM use by modifying the disk label and initializing the private region. The `voldg` command adds an initialized disk to a disk group. You can also use the `voldiskadd` utility to initialize disks and add them to a disk group.

For each disk group, the number of copies and the size of the LSM configuration database must be configured to provide the best failover performance. However, you also must ensure that enough copies of the configuration database are available.

By default, the `voldisksetup` utility sets up LSM disks with a private region size of 512 sectors, two copies of the configuration database, and two copies of the log regions. It is recommended that you set up four to eight copies of the LSM configuration database for each disk group used in the

ASE. For higher database availability, place the copies on different SCSI buses, disk storage boxes, or RAID controllers.

Example 4–4 and Example 4–5 create the LSM configuration shown in Figure 4–1. Example 4–4 shows how to use the `voldisksetup –i` command to initialize six entire disks with four copies of the configuration database and four copies of the log regions distributed among the disks. Example 4–4 also shows how to use the `voldg` command to create the `ase_dg1` disk group and add the initialized disks to the disk group.

Example 4–4: Initializing LSM Disks and Disk Groups

```
# voldisksetup -i rz16 nconfig=1 nlog=1

# voldisksetup -i rz18 nconfig=1 nlog=1

# voldisksetup -i rz20 nconfig=0 nlog=0

# voldisksetup -i rz24 nconfig=0 nlog=0

# voldisksetup -i rz26 nconfig=1 nlog=1

# voldisksetup -i rz28 nconfig=1 nlog=1

# voldg init ase_dg1 ase_16=rz16

# voldg -g ase_dg1 adddisk ase_18=rz18

# voldg -g ase_dg1 adddisk ase_20=rz20

# voldg -g ase_dg1 adddisk ase_24=rz24

# voldg -g ase_dg1 adddisk ase_26=rz26

# voldg -g ase_dg1 adddisk ase_28=rz28
```

Example 4–5 shows how to initialize two entire disks, with default private region parameters, create the `ase_dg2` disk group, and add the initialized disks to the disk group.

Example 4–5: Initializing LSM Disks with Private Region Parameters

```
# voldisksetup -i rz22 nconfig=2 nlog=2 privlen=512

# voldisksetup -i rz30 nconfig=2 nlog=2 privlen=512

# voldg init ase_dg2 ase_22=rz22

# voldg -g ase_dg2 adddisk ase_30=rz30
```

4.3.7.4 Creating and Configuring LSM Volumes for the ASE

If you want to use LSM volumes in your ASE services, you must configure them into the disk groups that you set up for the ASE. See the DIGITAL UNIX *Logical Storage Manager* manual for information about creating LSM volumes.

You can use the `volassist` command to create LSM volumes, as follows:

```
# volassist -g ase_dg2 make vl2 100m nmirror=2
```

The following command creates a striped LSM volume, `v1_ase`:

```
# volassist -g ase_dg1 make v1_ase 64m usetype=fsgen layout=stripe \
      nstripe=3 stwidth=8k ase_16 ase_18 ase_20
```

The following command mirrors the striped LSM volume created by the previous command:

```
# volassist -g ase_dg1 mirror v1_ase layout=stripe nstripe=3 \
     stwidth=8k ase_24 ase_26 ase_28
```

After a volume is created, you can create a UNIX file system on the volume, using the `newfs` command. For example:

```
# newfs /dev/rvol/ase_dg2/vl2 rz26
```

You can also use LSM volumes in AdvFS domains. The following example uses the `mkfdmn` command to create a new AdvFS domain on an existing LSM volume, and then uses the `mkfset` command to create a fileset on the domain:

```
# mkfdmn /dev/vol/ase_dg1/v1_ase dom_ase1
# mkfset dom_ase1 set_ase1
```

4.3.7.5 Using an Existing LSM Disk Configuration

The previous sections describe how to set up a new LSM disk configuration for an ASE service. However, you may also want to use a configuration that was used in a previous ASE service or for some other purpose. To do this, the disk groups used in the service must be imported to the machine on which you are running the `asemgr` utility. The disk groups cannot be imported to any other system.

To determine the system to which a disk group is imported, use the following command:

```
# voldg list
```

If the disk group is imported on the member system on which you will add the service, you can run the `asemgr` utility and add the service.

If the disk group is not imported on the member system on which you will add the service, you must import the disk group only to the member on which you will add the service.

Use the following command syntax to deport a disk group:

voldg deport [*disk_group*]

To import the disk group, perform the following tasks on the member system on which you will run the `asemgr` utility to add the service:

1. Define the disks that are used in the disk group by using the following command syntax:

   voldisk define [*disk*]

2. Place on line the disks that are used in the disk group by using the following command syntax:

   voldisk online [*disk. . .*]

3. Import the disk group by using the following command syntax:

   voldg import [*disk_group*]

4. Restart the volumes by using the following command syntax:

   volrecover -sb [*disk_group*]

After the disk group is imported, you can run the `asemgr` utility and add the service.

4.3.7.6 Specifying LSM Volumes in a Service

After you set up the LSM disk configuration, you can use the `asemgr` utility to set up an NFS, disk, or tape service that uses an LSM volume. You can specify the following information about an LSM volume:

- LSM volume character device name if you want to use a raw disk
- Device special file if the LSM volume is being used by a UNIX file system

- AdvFS fileset name if the LSM volume is being used by an AdvFS domain
- Mount point or export directory (optional for disk services)
- Netgroup or machine name to which you want to restrict access (optional and only for NFS services)
- Access mode (either read/write or read-only)
- Mount options other than the default options specified in `mount`(8)

When you specify an LSM, the `asemgr` utility displays the disk partitions that comprise the volume.

Example 4–6 shows how to specify an LSM volume in a disk service.

Example 4–6: Specifying LSM Volumes in a Disk Service

```
                 Specifying Disk Information

Enter one or more device special files, AdvFS filesets, or LSM
volumes to define the disk storage for this service.

For example:        Device special file:      /dev/rz3c
                    AdvFS fileset:            domain1#set1
                    LSM volume:               /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.

Enter a device special file, an AdvFS fileset, or an LSM volume as
storage for this service (press ’Return’ to end): /dev/vol/ase_dg2/vl2

                  Mount Point

The mount point is the directory on which to mount /dev/vol/ase_dg2/vl2.
If you do not want it mounted, enter "NONE."

Enter the mount point or NONE: NONE

Following is a list of device(s) and pubpath(s) for disk group ase_dg2:

        DEVICE  PUBPATH

        rz22c   /dev/rz22c
        rz30c   /dev/rz30c

Is this correct (y/n) [y]: y
```

See Example 4–3 for an example of specifying an LSM volume that uses an AdvFS fileset.

Note

When an ASE service using LSM volumes is relocated while a volume is open, LSM displays the following message on the console:

```
volklog_dgfree:
  Can’t free kernel logging area for vol_reset_kernel of group dg4
```

This message indicates that the LSM kernel log for the diskgroup was not cleared before the diskgroup was deported. When the diskgroup is reimported, the log area will be cleared. For this reason, you can safely ignore the console message.

4.3.7.7 Using LSM Mirroring in the ASE

You can use mirrored LSM volumes in your ASE services. If a disk that is included in a plex that is part of an LSM mirrored volume goes off line, the data in the volume is still available, as long as one complete plex of the volume remains on line. Thus, the service that uses the mirrored volume remains available if a disk fails.

When the failed disk is replaced, you can reincorporate the device into the ASE service by resynchronizing the mirrored volume outside of the ASE, and then rereserving the devices using the asemgr utility’s Advanced Utilities menu.

In rare cases, if any disk that is part of an LSM disk group is not accessible when the service is started, you might access stale data or a stale LSM configuration. This can occur if two member systems access different subsets of the disks in the disk group.

By default, the TruCluster software prevents accessing stale data by disabling the ASE_PARTIAL_MIRRORING run-time configuration variable that allows you to start an unsynchronized mirrored volume.

If you use LSM mirroring in a service and the ASE_PARTIAL_MIRRORING run-time variable is disabled (the default behavior), the service can start only if the member system can access all of the disks in the disk groups used in the mirroring. This ensures that the service cannot be started with obsolete LSM configuration information, which guarantees data integrity but limits service availability.

If you use the default behavior, when you use the asemgr utility to set up a service that uses an LSM mirrored volume, choose an ASP policy that will not automatically relocate the service to a more highly favored member if the system becomes available. If a plex fails, the service remains available on the member. Otherwise, if the plex fails and the service tries to relocate, the service will not start.

For maximum service availability, you can set the ASE_PARTIAL_MIRRORING variable to on by invoking the following command on all the members in the ASE:

```
# rcmgr set ASE_PARTIAL_MIRRORING on
```

If you do this, a service that uses LSM mirroring can start when only one plex of the data is available, but there is a remote possibility that the service will use obsolete LSM configuration information.

If the ASE_PARTIAL_MIRRORING variable is disabled, you can force a service to start by enabling the ASE_PARTIAL_MIRRORING variable, manually restarting the service, and then disabling the variable. Do this if you know that a disk has failed and you want to relocate the service.

4.3.7.8 Understanding the LSM Pseudodevice

The TruCluster software uses a pseudodevice if a service includes both UFS and LSM. When the TruCluster software mounts a UNIX file system on an LSM volume, it maps the volume to an ASE pseudodevice (/dev/ase_*nnn*) and then mounts the pseudodevice. For example, if you have an NFS service using the /dev/vol/dg1/vol01 volume, the following information is displayed when you invoke the mountcommand:

```
# mount
⋮
/dev/ase_001 on /usr/var/ase/mnt/ase18/usr/ase18 type ufs (rw)
```

A pseudodevice is used because when LSM volumes move from one member to another, the major and minor numbers for the volume may change. Because NFS uses the major and minor numbers for its file handle, a change in these numbers means that the file handle may correspond to different file systems on each member. ASE guarantees consistent file handles by using a pseudodevice.

4.4 Installing an Application to Fail Over

If your service uses an application that you want to fail over, you must install the application before you set up the service. For example, when you set up an NFS mail service, you must set up the mail hubs on the member systems before you use the asemgr utility to add the service to the ASE. In addition, before you set up a disk service that makes a database program highly available, you must install the database program on all the members.

Only certain types of applications can be made highly available with an ASE service. The application must have the following characteristics:

- The application must run on only one system at a time.
- The application must be able to be started and stopped using a set of commands that are performed in a specific order. When you set up a service, these commands are included in a set of programs called **action scripts**.

When you install the application that you want to fail over, you must ensure that the application is located in the correct directory and has the correct access mode.

Section 4.5 describes how to set up the action scripts that the TruCluster software uses to fail over your application.

4.5 Using Action Scripts

The TruCluster software uses action scripts to fail over the services in the ASE. Action scripts break down a procedure (for example, starting a service on a member system) into a series of steps, which are executed in order. The TruCluster software makes certain that each step is successfully completed. The order of the steps ensures that any dependencies are met before the next step is performed.

There are five types of action scripts: add, delete, start, stop, and check action scripts. In addition, there are two versions of each type of action script: internal and user-defined action scripts. These action scripts are executed at specific times and perform specific tasks.

The types of action scripts are as follows:

- Add action script—After you use the `asemgr` utility to set up an ASE service, the TruCluster software executes each add action script on all the member systems to configure the service on the members. The TruCluster software executes add action scripts on all the members because each member must be able to run every service. An add action script contains all the commands you need to set up the system environment to enable the service to run. For example, an add action script could edit system files.
- Delete action script—If you use the `asemgr` utility to delete a service from the ASE, the TruCluster software executes each delete action script on all the members to remove the service from the members. Delete action scripts reverse any service setup tasks that the add action scripts perform. For example, if an add action script made changes to a file, a delete action script will edit that same file and remove the changes.

- Start action script—When the TruCluster software starts a service on a member system, it executes each start action script only on the member that the TruCluster software chooses to run the service. This is because only one member runs a service at any time. Start action scripts contain all the commands that are necessary to start a service on a member. For example, a start action script could invoke the application that you want to make highly available in the ASE.
- Stop action script—When the TruCluster software stops a service on a member system, it executes each stop action script to stop the service on that member. Stop action scripts reverse the tasks that the start action scripts perform. For example, if a start action script invokes an application, the stop action script will include a command to stop that application. If the stop action scripts do not stop all the processes accessing the disks used in a service, the disks cannot be unmounted and the service cannot be stopped. See Section 4.3.2 for more information.
- Check action script—Check action scripts determine if a service is running. When the director daemon starts, it interrogates each member's agent daemon to determine which services are running. The agent daemons invoke the services' check action scripts and collect their exit status to determine if the services are running. The agent daemons also invoke the check action scripts when you stop or delete a service to verify the state of the service. Usually, check action scripts check for a specific running process by using the `ps` command. For example, the following command checks for the `mountd` daemon and starts the daemon if necessary:

  ```
  /bin/ps -e | grep "mountd" | grep -v grep || /usr/sbin/mountd
  ```

  In addition, some applications can create a file that contains the process identification number (PID) of an active daemon. Check action scripts could check for the existence of that PID to determine if a service is running.

For the five types of action scripts, there are two versions of each type of script, as follows:

- Internal action scripts—Internal action scripts are used only in NFS, disk, and tape services. These scripts contain the programs that make the UNIX file systems, AdvFS filesets, or LSM volumes highly available. These scripts cannot be manually edited. You specify information in the internal scripts only by responding to the `asemgr` utility prompts when you create a service. Because user-defined services do not use disks, there are no internal action scripts for user-defined services. Usually, for NFS services, the internal action scripts are the only scripts you need.

- User-defined action scripts—User-defined action scripts contain the commands that the TruCluster software uses to fail over your application. For example, if you want to fail over a database application, set up user-defined action scripts that include the commands to start and stop the application.

  You must create user-defined action scripts. The TruCluster software provides skeleton scripts that you can edit using the `asemgr` utility. You can also create your own scripts outside of the ASE and then use the `asemgr` utility to specify them in the service. If you create your own scripts, they must be installed locally on each system. User-defined services require you to create user-defined action scripts, because they do not use any internal action scripts. A user-defined check action script is supported only in a user-defined service. Although NFS and disk services use internal scripts to fail over disks, if you also want to fail over an application, you must create user-defined scripts to start and stop the application.

4.5.1 Execution Sequence for Action Scripts

Internal and user-defined action scripts are executed in a specific order, depending on the task the TruCluster software is performing. The following information shows the sequence of execution for an NFS service that uses both user-defined action scripts and internal action scripts.

To add an NFS service that is configured for Advanced File System (AdvFS) to the ASE, the TruCluster software does the following:

1. Runs the internal add action script, which performs the following tasks:
   a. Invokes the `MAKEDEV` command to create the device files for the devices in the service.
   b. Creates the `/etc/exports.ase.`*`service`* file.
2. Runs any user-defined add action script.

To delete an NFS service from the ASE, the TruCluster software does the following:

1. Runs any user-defined delete action script.
2. Runs the internal delete action script, which removes the `/etc/exports.ase.`*`service`* file.

To start an NFS service on a member system, the TruCluster software does the following:

1. Runs the internal start action script, which performs the following tasks:

    a. Sets up AdvFS domains, as necessary.

    b. Invokes the `fsck` command on any UNIX file systems used in the service.

    c. Mounts all the file systems.

    d. Adds the following line to the `/etc/exports.ase` file:

       `.INCLUDE /etc/exports.ase.` *`service`*

    e. Aliases the virtual host using the following command:

       `ifconfig ln0 alias` *`hostname`*

    f. Starts NFS locking by invoking the `statd` and `lockd` daemons.

2. Runs any user-defined start action script.

To stop an NFS service on a member system, the TruCluster software does the following:

1. Runs any user-defined stop action script.

2. Runs the internal stop action script, which performs the following tasks:

    a. Stops NFS locking by stopping the `statd` and `lockd` daemons.

    b. Removes the virtual host alias by using the following command:

       `ifconfig ln0 -alias` *`hostname`* command.

    c. Removes the `.INCLUDE /etc/exports.ase.` *`service`* line from the `/etc/exports.ase` file.

    d. Unmounts all the file systems.

    e. Unconfigures the domain.

4.5.2 Exit Codes for Action Scripts

When the TruCluster software invokes an action script, it usually considers a 0 (zero) exit code as a success. An exit code of 1 indicates that the script failed. The exception is the check action script, which exits with an exit code that is between 100 and 200 to indicate that the service is running, and an exit code that is less than 100 to indicate that the service is not running. A stop script can produce the exit code 99, which indicates that the service could not be stopped because the service was busy.

All standard output and standard error output from your script goes to the ASE logger daemon, if running, or to the `syslog` daemon. If an internal action script exits with a 0 (zero), it is logged as an informational message; if a user-defined action script exits with a zero, it is a notice. If either type of script exits with an exit code other than zero, the messages are logged as errors.

4.5.3 Specifying User-Defined Action Scripts for a Service

You use the `asemgr` utility to specify any user-defined action scripts when you add or modify a service. If you want to fail over an application, at a minimum, you must create start and stop action scripts to start and stop the application. If you need to set up the system environment in order for the service to run, then you also must create add and delete action scripts. For user-defined services, it is recommended that you create a check action script, so the TruCluster software can determine if the service is running.

There are three ways to specify a user-defined action script for a service:

- You can create a script that will perform the desired task outside of the ASE and specify the pathname location of the script when the `asemgr` utility prompts you for the script name. Your script will be copied into the ASE database. After you specify the script, you can edit the script only by using the `asemgr` utility. This is because the TruCluster software uses the copy of the script that is in the ASE database and not the one that is located on the system.
- When the `asemgr` utility prompts you for a script name, you can specify `default`. You then edit the skeleton action script that the TruCluster software provides, and include the commands that will perform the desired task.
- You can create a script that will perform the desired task outside of the ASE and copy it to all the members. Specify `default` when prompted for a script name, and edit the default script to include a one-line pointer to the pathname of the existing script. If you use this method, you can modify the script by manually editing it on each member. Because you do not have to use the `asemgr` utility, you can modify the script without interrupting the service.

In addition to specifying the user-defined action scripts, the `asemgr` utility allows you to specify the following script information:

- Arguments that are passed to the script—Arguments can be useful if you have a generic script to which you need to pass the service name or an action in order to make it work.

- Timeout value—The timeout value for a script is the specified length of time the TruCluster software waits for your script to finish running. This value should be the maximum amount of time that the script needs. If your script runs longer than the timeout value (for example, because it hangs), the TruCluster software considers the script failed and reports the failure as a timeout of the script.

Example 4–7 shows how to specify the pathname of an action script at the `asemgr` prompt. The script must already be installed on your system.

Example 4–7: Specifying Your Own Action Scripts

```
Enter the full pathname of your start action script or "default"
for the default script (x to exit): /usr/sbin/dbase_account

Enter the argument list for the start action script
(x to exit, NONE for none): start

Enter the timeout in seconds for the start action script [60]:  Return
```

Example 4–8 shows how to specify and edit a default skeleton script that the TruCluster software provides.

Example 4–8: Editing Default Action Scripts

```
Enter the full pathname of the start action script or "default"
(for the default script (x to exit): default

Enter the argument list for the start action script
(x to exit NONE for none) [prophecy_1]: NONE

Enter the timeout in seconds for the start action script [60]:80

Modifying the start action script for ‘prophecy_1‘:

    f)  Replace the start action script
    e)  Edit the start action script
    g)  Modify the start action script arguments [none]
    t)  Modify the start action script timeout [80]
    r)  Remove the start action script
    x)  Exit - done with changes

Enter your choice [x]: e

PATH=/sbin:/usr/sbin:/usr/bin
export PATH
ASETMPDIR=/var/ase/tmp

if [ $# -gt 0 ]; then
```

Example 4–8: Editing Default Action Scripts (cont.)

```
        svcName=$1
else
        svcName=
fi

# Start prophecy_1 database program:
su - prophecy -c dbstart
exit 0
:wq
```

Action scripts can test the ASE_STATE variable to determine the state of a service. If the ASE_STATE variable is STARTING, this indicates that the service's stop action script was invoked because of a system reboot. If the ASE_STATE variable is RUNNING, this indicates that the service's stop action script was invoked because the service was modified, deleted, placed off line, or relocated.

A stop action script can test the value of the ASE_STATE variable and, if the value is STARTING, perform some tasks to ensure that erroneous processes are not running.

You may want to save the user-defined action scripts in a file that you can access from outside of the ASE. For example, you may want to archive the files, or you may want to test the scripts outside of the ASE. To do this, use the asemgr utility to edit the scripts. Then, while in the editor, use a command similar to the following vi editor command to copy the script to a file:

```
:w /tmp/foo
```

4.5.4 Debugging User-Defined Scripts

Always test your script outside of the ASE to ensure that it works. When you are ready to set up a service that uses the script, first set the ASE logging level to informational, so that all messages are logged. After you set up the service to use the scripts, examine the syslog daemon logs for any problems. If you have added any debug echoes to the script, they will show up in the log file.

If your stop script fails when you use the asemgr utility to modify a service or place it off line, the TruCluster software places the service off line and prompts you to ensure that the service has stopped. If this occurs because of an error in your stop script, use the asemgr utility to edit the script and correct the problem, then place the service back on line.

If a stop service action fails to stop a service, the asemgr utility provides you with a number of opportunities to fix the problem. See Section 10.3 for information.

5

Setting Up an NFS Service

A Network File System (NFS) service includes one or more file systems, Advanced File System (AdvFS) filesets, or Logical Storage Manager (LSM) volumes that a member system exports to clients, making the data highly available. NFS services can also include highly available applications.

An NFS service name is assigned its own Internet address. The member system that runs the service responds to this address. This makes the service autonomous and not dependent on the availability of any particular member system. Clients access the service by including the service name and the exported directory path in their `/etc/fstab` file. If the service stops on a member system, it fails over to a viable system, and clients only experience a short timeout.

The NFS service name also allows you to use the POLYCENTER NetWorker Save and Restore (NetWorker) to back up the service's storage. NetWorker treats the NFS service as an independent client and stores the storage indexes under the name of the service. This enables you to back up and recover the service's storage independent of the member system running the service. See the NetWorker documentation for information about using NetWorker to back up an NFS service's storage.

To set up an NFS service, you should be familiar with setting up NFS in general, the `/etc/exports` file, and the `/etc/fstab` file.

Before you set up your NFS service, both the client and member systems must be running NFS Version 2.0 or Version 3.0 and use the Address Resolution Protocol (ARP). You also must prepare the shared disks that will be used in the service and install any application used in the service.

To fail over an application, in addition to disks, at a minimum, you must create a user-defined start action script that includes the commands to start the application, and create a user-defined stop action script that includes the commands to stop the application. See Chapter 4 for more information about preparing disks, applications, and action scripts for a service.

5.1 NFS Service Requirements

Network Files System (NFS) services have the following requirements:

- NFS service names must be included in the local `/etc/hosts` file on each member system before you set up the NFS service.
- You cannot use an NFS service name that is the same as the name of a member system. Service names and member system names must be unique.
- When exporting directories from a service, the directory names must be unique within the entire ASE, and unique from any nonshared exports made from the `/etc/exports` file from any ASE member nodes.
- An NFS service name must adhere to the conventions for naming a system, as described in the DIGITAL UNIX *Installation Guide*.
- If you are using a distributed database lookup service such as the Network Information Service (NIS), be sure that the service name information is local to all the member systems. To do this, make all the member systems either master or slave servers, or specify the service name information in the local `/etc/hosts` file. Be sure your `/etc/svc.conf` file specifies `local` as the first entry.
- Do not manually mount or dismount disks that are used in an NFS service.
- If you set up an NFS service in your ASE, do not use the `automount` command option `/net –hosts` on any client system that accesses the service's NFS file systems.

  If a client uses the `/net/`*`hostname`* or `/net/`*`service_name`* mount points to NFS-mount a file system from a member system, and TruCluster software relocated the NFS sevice to a different member system, the client may receive "stale file handle" error messages.
- Use the `automount` command `-p` option to limit the local loopback mounts to primary Internet addresses. When passed this option, `automount` queries each of the system's configured network interfaces for its primary Internet address. `automount` uses local loopback mounts for all Internet addresses produced from this query. All Internet alias addresses for the system will be treated as remote addresses and will use NFS mounts.

  By default, `automount` bypasses NFS for all local Internet addresses, including Internet alias addresses. The `automount -p` option enables Internet alias addresses to use NFS for all automounted file systems, and is necessary for ASE servers to avoid difficulty modifying or stopping services.

5.2 NFS Service Components

When you add a Network File System (NFS) service to an available server environment (ASE), the `asemgr` utility prompts you for service-specific information, in addition to information that is similar to what you specify with the `nfssetup` script. See `nfssetup`(8) for more information.

You can specify the following NFS service information:

- Service name—To enable clients to access an NFS service, the service name is assigned its own Internet address. The service name and Internet address must be included in each member system's local `/etc/hosts` file before you set up the NFS service. See `hosts`(4) for more information. In addition, an NFS service name must adhere to the conventions for naming a system, as described in the DIGITAL UNIX *Installation Guide*.
- Automatic Service Placement (ASP) policy—See Chapter 4 for information about the ASP policies.
- UNIX file systems, Advanced File System (AdvFS) filesets, or Logical Storage Manager (LSM) volumes—See Chapter 4 for information about setting up disks.
- Mount point to be exported for each file system, fileset, or volume—Client systems will specify this mount point in their `/etc/fstab` files to access the service's file systems, filesets, or volumes.
- Names of the remote hosts, network groups, or Internet Protocol (IP) addresses to which you want the file systems, filesets, or volumes restricted—See `exports`(4) for information about remote host access to NFS mount requests.
- NFS locking area—If you have more than one writable disk area in a service, you must specify which area to use for the NFS locking area. The NFS locking software maintains some data that must be failed over, so the data must be placed on a writable, shared disk.
- Mode (either read/write or read-only) and any mount options other than the default options for each file system, fileset, or volume.

If you also want to fail over an application, you must modify the NFS service and specify the action scripts. See Chapter 4 for information about action scripts. See Chapter 10 for information about modifying services.

5.3 Understanding the Service Exports File

When you add a Network File System (NFS) service to an available server environment (ASE), the TruCluster software edits the `/etc/exports.ase` file on each member system and includes an entry that specifies the service's exports file. For example:

```
 # more exports.ase

 .INCLUDE /etc/exports.ase.aseba1
 .INCLUDE /etc/exports.ase.aseba2
 #
```

Service exports file names have the following syntax:

/etc/exports.ase.*service*

The *service* variable specifies the service name.

A service exports file contains a list of all the file systems and filesets in the service and their mount points, using a format that is similar to the `/etc/exports` file. It can include the remote hosts, network groups, or Internet Protocol (IP) addresses to which the service's file systems or filesets are restricted. If none are specified in the file, then all remote hosts can mount the directory. See `exports`(4) for information about the file format.

Entries in service exports files include a `–m` option, which specifies the actual mount point for a file system or fileset.

Note

> Do not manually edit the `/etc/exports.ase.`*service* file to modify services; instead, use the `asemgr` utility to make modifications.
>
> To delete a file system or fileset from an NFS service, use the `asemgr` utility to remove its entry from the `/etc/exports.ase.`*service* file. When deleting a file system or fileset, the `asemgr` utility prompts you to invoke an editor, providing the opportunity to delete the entry at this time. If you choose to not run an editor at this time, then you must remember to do so later.

The following example shows an exports file for an NFS service with two file systems:

```
#
#  ASE exports file for service aseba2 (edit only with asemgr)
#

#/dev/rz25c exports (after this line) - DO NOT DELETE THIS LINE
/ase/aseba2 -m=/var/ase/mnt/aseba2/ase/aseba2 -ro=0
```

```
#/dev/rz26c exports (after this line) - DO NOT DELETE THIS LINE
/ase/aseusr -m=/var/ase/mnt/aseba2/ase/aseusr  testit milan tabby
#
```

5.4 Adding a Basic NFS Service

To add a Network File System (NFS) service to an available server environment (ASE), choose the "Adding a new service" item from the Service Configuration menu and provide the appropriate information for the service at the prompts. Example 5–1 shows an example of adding a basic NFS service that includes a UNIX file system and a Logical Storage Manager (LSM) volume.

Example 5–1: Adding a Basic NFS Service

```
# asemgr

.
.
.
      Adding a service

Select the type of service:

    1)  NFS service
    2)  Disk service
    3)  User-defined service
    4)  DRD service
    5)  Tape service

    q)  Quit without adding a service
    x)  Exit                ?)  Help

Enter your choice [1]: 1

You are now adding a new NFS service to the ASE.

An NFS service consists of an IP host name and disk configuration
that are failed over together.  The disk configuration can include
UFS file systems, AdvFS filesets, and LSM disk groups.

                     NFS Service Name

The name of an NFS service is a unique IP host name that has been
set up for this service.  This host name must exist in the local
hosts database on all ASE members.

Enter the NFS service name: ase3


Checking to see if ase3 is a valid host...

                     Specifying Disk Information

Enter one or more UFS device special files, AdvFS filesets, or LSM
volumes to define the disk storage for this service.

    For example:        Device special file:      /dev/rz3c
                        AdvFS fileset:            domain1#set1
                        LSM volume:               /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.

Enter a device special file, an AdvFS fileset, or an LSM volume as
```

Example 5–1: Adding a Basic NFS Service (cont.)

```
storage for this service (press ’Return’ to end): /dev/rz25c

Enter the directory pathname(s) to be NFS exported from the storage
area "/dev/rz25c".  Press ’Return’ when done.

Enter a directory pathname: /ase_dir
Enter a host name, NIS netgroup, or IP address for the
NFS exports list (press ’Return’ for all hosts):  Return

Enter a directory pathname:  Return

              UFS File System Read-Write Access

Mount /dev/rz25c file system with read-write or read-only access?

    1)  Read-write
    2)  Read-only

Enter your choice [1]: Return

You may enable user and group quotas on this file system by specifying
full path names for the quota files.  If you place the files within
the service’s file systems, the quota assignments you make with
edquota will relocate with the service.  Enter "none" to disable
quotas.

User quota file [/var/ase/mnt/ase3/ase_dir/quota.user]:  Return

Group quota file [/var/ase/mnt/ase3/ase_dir/quota.group]:  Return

              UFS Mount Options Modification

Enter a comma-separated list of any mount options you want to use
for "/dev/rz25c" (in addition to the UFS-specific defaults listed
in the mount.8 reference page).  If none are given, only the
default mount options are used.

Enter options (Return for none): noexec

Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press ’Return’ to end): /dev/vol/dg3/vol04

Enter the directory pathname(s) to be NFS exported from the storage
area "/dev/vol/dg3/vol04".  Press ’Return’ when done.

Enter a directory pathname: /ase_data

Enter a host name, NIS netgroup, or IP address for the

NFS exports list (press ’Return’ for all hosts): net_staff
Enter a directory pathname:  Return

The following is a list of device(s) and pubpath(s) for disk group dg3:

        DEVICE  PUBPATH

        rz20c   /dev/rz20c

Is this correct (y/n) [y]: y

              UFS File System Read-Write Access

Mount /dev/vol/dg3/vol04 file system with read-write or read-only access?

    1)  Read-write
    2)  Read-only

Enter your choice [1]: 2

              UFS Mount Options Modification
```

Example 5–1: Adding a Basic NFS Service (cont.)

```
Enter a comma-separated list of any mount options you want to use
for "/dev/vol/dg3/vol03" (in addition to the UFS-specific defaults listed
in the mount.8 reference page).  If none are given, only the default
mount options are used.

Enter options (Return for none): nosuid

Enter a device special file, an AdvFS fileset, or an LSM volume as
storage for this service (press 'Return' to end):  Return

NFS needs a disk area that is writable to keep some state information
for NFS locking during ASE operation.  Choose a disk area that is
writable and will not fill up.

Select the disk area to use for the NFS locking information:

    1)  /dev/rz25c  (UFS)
    2)  /dev/vol/dg3/vol04 (UFS)
    x)  Exit

Enter your choice [1]: 1


        Selecting an Automatic Service Placement (ASP) Policy

Select the policy you want ASE to use when choosing a member
to run this service:

    b)  Balanced Service Distribution
    f)  Favor Members
    r)  Restrict to Favored Members

    x)  Exit to Service Configuration     ?)  Help

Enter your choice [b]: b


       Selecting an Automatic Service Placement (ASP) Policy

Do you want ASE to relocate this service to a more highly favored
member if one becomes available while the service is running (y/n/?):y

Enter 'y' to add Service 'ase3' (y/n):y

Adding service...
Starting service...
Saving the updated database...
Service successfully added...
```

5.5 Adding an NFS Mail Service

You can use the TruCluster software to set up a mail system and make it highly available. The `sendmail` program uses the Simple Mail Transfer Protocol (SMTP) to deliver mail messages between users, systems, and networks. You can set up member systems as mail hubs (servers) so that other systems in your mail environment send mail to and through the mail hubs. If a problem occurs in a mail hub, the TruCluster software can fail over the mail to another hub and reroute incoming mail to the new hub.

Before setting up a mail service, you should understand how the sendmail program works. The sendmail program can receive mail from an SMTP connection or directly from a process; that is, from the mail system or some user interaction. For example:

1. The sendmail program writes the message to a mail queue area, which is the /var/spool/mqueue directory by default.
2. After the entire mail message is written to the mail queue area, sendmail tells the sending process that the mail was received, so the sending process is assured that the mail was delivered. If the machine crashes at this point, a copy of mail remains in the mail queue area, so the mail is not lost.
3. After a secure copy of the message is in the mail queue area, sendmail parses the address and delivers the message according to the instructions in the sendmail configuration file, /var/adm/sendmail/sendmail.cf.
4. The sendmail program passes the mail to another delivery agent, such as DECnet, UNIX-to-UNIX Copy Program (UUCP) or another SMTP. In addition, local mail is passed to the local mailer (/usr/bin/mail) which, by default, delivers it to the system mailbox, /var/spool/mail/*username*. If the address is not local, sendmail passes the mail to the mail delivery agent on the remote machine. If sendmail cannot pass the mail (for example, the remote machine is down), the mail remains in the queue area to be processed at a later time.

To set up a highly available mail service with the TruCluster software, the file systems or filesets containing the /var/spool/mail mailbox directory and the /var/spool/mqueue mail queue area must be shared between two or more systems. The /var/spool/mail mailbox directory must be shared so that the mail drop is available for mail delivery and processing on all the member systems that are set up as mail hubs. The /var/spool/mqueue queue area must be shared to ensure that any mail that remains in the queue area can be processed even if the hub that queued the mail is not available. In addition, both areas must have a common floating network connection to which mail can be sent.

You can share the mailbox directory and the queue area by using the TruCluster software to set up an NFS service to share the file systems or filesets and by modifying all the mail hubs' sendmail.cf configuration files.

The following steps describe how to set up two member systems as mail hubs (server1 and server2) using an NFS service named mail_hub:

1. Set up the disk or disks that will contain the /var/spool/mail and /var/spool/mqueue file systems.
2. Use the asemgr utility to set up the NFS service mail_hub that will export the /var/spool/mail and /var/spool/mqueue file systems.

   Note

   NFS locking (the lockd daemon) must be set up and running on the member systems that are mail hubs because locking will be done on both exported spool areas.

3. Modify the mail_hub service's exports file. Use the asemgr utility to make the /var/spool/mail and /var/spool/mqueue file systems accessible by root.
4. Set up the sendmail.cf configuration files on the member systems that will be mail hubs, server1 and server2. You must set up the sendmail.cf files to ensure that mail addressed to user@server1, user@server2, or user@mail_hub is delivered locally.
5. NFS-mount the file systems on the mail hub member systems. On both mail hub member systems, server1 and server2, use the mount command to NFS-mount /var/spool/mail and /var/spool/mqueue from the service (Internet host name) mail_hub and then add this mount point to the /etc/fstab file on each member system.
6. Optionally, define a Berkeley Internet Name Domain (BIND) mail exchanger (MX) record to point to all mail hubs. You can define an MX record so that if server1 is inaccessible, mail sent to server1 is forwarded to server2 and vice versa.

After you complete these steps, the mail service is ready to use. You can send mail to server1, server2, or mail_hub, and your mail will be delivered to the shared local /var/spool/mail area on the server1 and server2 mail hub member systems.

You can log in to server1, server2, or mail_hub and access your mail. You can also mount /var/spool/mail@mail_hub on another system and access your mail from that system. However, if one of the mail hub member systems goes down, mail sent directly to that mail hub member system will not be delivered until it reboots. You can fix this problem by defining the BIND MX records.

The following sections describe the steps in detail.

5.5.1 Preparing Disks for a Mail Service

To prepare the disks that will contain the shared /var/spool/mail and /var/spool/mqueue file systems, follow the guidelines specified in Chapter 4.

If you are using one disk partition for both the /var/spool/mqueue and /var/spool/mail directories, create an mqueue directory and a mail directory. You perform these tasks on only one mail hub member system.

The following example sets up a UNIX file system on an entire RZ10 disk and creates two directories:

```
# newfs /dev/rrz10c
# mount /dev/rz10c /mnt
# mkdir /mnt/mail
# chmod 1777 /mnt/mail
# mkdir /mnt/mqueue
# chmod 755 /mnt/mqueue
# umount /mnt
```

5.5.2 Using the asemgr Utility to Add a Mail Service

To add a mail service to your available server environment (ASE), run the asemgr utility on one mail hub member system, choose the "Add a new service" item from the Service Configuration menu and provide the information appropriate for your configuration at the prompts.

Example 5–2 shows how to add a service named mail_hub, which consists of the /dev/rz10c file system and the /var/spool/mqueue and /var/spool/mail directories. The example also shows how to restrict access to the service to the server1 and server2 mail hub member systems.

Example 5–2: Adding a Mail Service

```
# asemgr

.
.
.
          Adding a service

Select the type of service:

    1)  NFS service
    2)  Disk service
    3)  User configured service
    4)  DRD service
    5)  Tape service

    q)  Quit without adding a service
    x)  Exit                  ?)  Help

Enter your choice [1]: 1
```

Example 5–2: Adding a Mail Service (cont.)

```
You are now adding a new NFS service to the ASE.

An NFS service consists of an IP host name and disk configuration that
are failed over together.  The disk configuration can include UFS file
systems, AdvFS filesets, and LSM disk groups.

                     NFS Service Name

The name of an NFS service is a unique IP host name that has been set
up for this service.  This host name must exist in the local hosts
database on all ASE members.

Enter the NFS service name: mail_host

Checking to see if mail_host is a valid host...


                   Specifying Disk Information

Enter one or more UFS device special files, AdvFS filesets, or LSM
volumes to define the disk storage for this service.

    For example:        Device special file:      /dev/rz3c
                        AdvFS fileset:            domain1#set1
                        LSM volume:               /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.

Enter a device special file, an AdvFS fileset, or an LSM volume as
 storage for this service (press ’Return’ to end): /dev/rz10c

Enter the directory pathname(s) to be NFS exported from the storage
area /dev/rz10c.  Press ’Return’ when done.

Enter a directory pathname: /var/spool/mail
Enter a host name, NIS netgroup, or IP address for the
NFS exports list (press ’Return’ for all hosts):server1 server2

Enter a directory pathname: /var/spool/mqueue

Enter a host name, NIS netgroup, or IP address for the
NFS exports list (press ’Return’ for all hosts):server1 server2

Enter a directory pathname: Return

              UFS File System Read-Write Access

Mount /dev/rz10c file system with read-write or read-only access?

    1)  Read-write
    2)  Read-only

Enter your choice [1]: 1

You may enable user and group quotas on this file system by specifying
full path names for the quota files.  If you place the files within
the service’s file systems, the quota assignments you make with
edquota will relocate with the service.  Enter "none" to disable
quotas.

User quota file [/var/ase/mnt/mail_host/var/mail/quota.user]: none
Group quota file [/var/ase/mnt/mail_host/var/mail/quota.group]: none


              UFS Mount Options Modification

Enter a comma-separated list of any mount options you want to
use for /dev/rz10c (in addition to the UFS-specific defaults
listed in the mount.8 reference page).  If none are given, only
the default mount options are used.
```

Example 5–2: Adding a Mail Service (cont.)

```
Enter options (Return for none):  Return

Enter a device special file, an AdvFS fileset, or an LSM volume as
storage for this service (press 'Return' to end):  Return

        Selecting an Automatic Service Placement (ASP) Policy

Select the policy you want ASE to use when choosing a member
to run this service:

    b)  Balanced Service Distribution
    f)  Favor Members
    r)  Restrict to Favored Members

    x)  Exit to service config menu      ?)  Help

Enter your choice [b]: b

        Selecting an Automatic Service Placement (ASP) Policy

Do you want ASE to relocate this service if a more highly favored
member becomes available while the service is running (y/n/?):n

Enter 'y' to add Service 'mail_host' (y/n): y
Adding service...
Starting service...
Saving the updated database...
Service successfully added...
```

5.5.3 Modifying the Mail Service's Exports File

On one mail hub member system, use the `asemgr` utility to edit the `mail_hub` service's exports file and make the `/var/spool/mail` and `/var/spool/mqueue` directories accessible by root on all the mail hub member systems. You must add the `-root=0` option to the entries for the `/var/spool/mqueue` and `/var/spool/mail` directories in the `/etc/exports.ase.mail_hub` file.

To edit the `mail_hub` service's ASE exports file, follow these steps:

1. Invoke the `asemgr` utility and choose the "Modify a service" menu item from the Service Configuration menu.
2. Choose the name of the service you want to modify. In this example, choose `mail_hub`.
3. Choose the "General service information" menu item when prompted for what you want to modify.
4. Choose the disk area that contains the mail areas to modify. In this example, choose `/dev/rz10c`.

5. Choose the "Modify the NFS exports list" menu item. The asemgr utility invokes an editor (as defined by the EDITOR system variable) so you can edit the mail service's ASE exports file.
6. Edit the exports file to include the -root=0 option. For example, the file should look like the following:

   ```
   #
   #  ASE exports file for service mail_hub
   #

   /dev/rz10c exports (after this line) -
    DO NOT DELETE THIS LINE
   /var/spool/mqueue -root=0 server1 server2
   /var/spool/mail -root=0 server1 server2
   ```

7. Exit the asemgr utility. The mail_hub service's exports file, /etc/exports.ase.mail_hub, is updated on all the mail hub member systems.

5.5.4 Setting Up the sendmail.cf Configuration File

On each mail hub member system, you must set up the sendmail.cf configuration file to handle mail sent directly to the mail hub member systems and to the mail service name as local mail. For example, if server1 and server2 are mail hub member systems for the NFS mail service mail_hub, you must set up the sendmail.cf configuration file on both mail hub member systems to ensure that mail sent to server1, server2, and mail_hub is handled as local mail.

Because server1, server2, and mail_hub share the /var/spool/mail area, mail sent to any of the three addresses is delivered to the shared local /var/spool/mail area.

You can use several methods to configure sendmail to do this:

- Use the mailsetup command
- Modify the *server*.m4 file and then reconfigure sendmail
- Directly modify the sendmail.cf file

You must configure the sendmail.cf file on all the mail hub member systems. To do this, invoke the mailsetup command and choose the option to perform an advanced mail setup. Add server1, server2, and mail_host to the NICKNAMES FOR THIS MACHINE section. Example 5–3 shows how to use the mailsetup program.

Example 5–3: Using mailsetup to Configure the sendmail.cf File

```
# mailsetup

.
.
.

                   NICKNAMES FOR THIS MACHINE

Are there any other names that are used to send mail to this
machine?  For instance, if you have changed this host's name
(or plan to in the near future), a nickname allows sendmail to
recognize both names, "pearly" and the nickname, as synonyms
for this machine.

Another good use for nicknames occurs when a host receives mail
from multiple different networks.  A host's name may not be the
same on all of the different networks.  Again, nicknames allows
sendmail to recognize these different names as synonyms for this
host.

Do you wish to enter nicknames for this machine (y/[n])? y

The following have been defined for the nicknames for server1 class:

add to list, delete from list, or continue on (a/d/c)? a

Enter additions to class (space or <cr> separated) - end list with a <cr>


? server1 server2 mail_hub
?  Return

The following have been defined for the nicknames for server1
class:

    server1 server2 mail_hub

add to list, delete from list, or continue on (a/d/c)? c

.
.
.
```

If you have already set up mail using the `mailsetup` program, follow these steps to manually configure the `/var/adm/sendmail.cf` file:

1. Change your directory to `/var/adm/sendmail`.
2. Edit the `server1.m4` file and add `server1`, `server2`, and `mail_hub` to the definition of `_MyNicknames`:

   ```
   dnl -- Other names for me - aliases of my machine
   define(_MyNicknames,    {server1 server2 mail_hub})dnl
   ```
3. Use the `make` command to update the `server1.cf` file:

   ```
   # make -f Makefile.cf.server1

   # mv sendmail.cf sendmail.cf.sav

   # cp server1.cf sendmail.cf
   ```

4. Restart the sendmail program:

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

To directly edit the sendmail.cf file, follow these steps:

1. Change to the /var/adm/sendmail directory.
2. Edit the sendmail.cf file and add the following line:

   ```
   Cw server1 server2 mail_hub
   ```

3. Restart the sendmail program:

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

5.5.5 Mounting the Disks

After you complete the preliminary steps, you can use the mail service. The final step is to mount the /var/spool/mqueue and /var/spool/mail directories on the server1 and server2 mail hub member systems. Perform the following steps on both mail hub member systems:

1. Disable the sendmail program:

   ```
   # /sbin/init.d/sendmail stop
   ```

2. If your mail hub member systems are active servers, you must save the old /var/spool/mqueue and /var/spool/mail areas so you do not lose any mail or queue files:

   ```
   # cd /var/spool

   # mv mqueue mqueue.old

   # mv mail mail.old
   ```

   You can move the old mail files to the new /var/spool/mail area after it is set up. You can run any queue files later by using the following command:

   ```
   # sendmail -q -oQ/var/spool/mqueue.old
   ```

3. Re-create the directories:

   ```
   # mkdir mqueue

   # mkdir mail
   ```

4. Mount the mail service spool areas:

   ```
   # mount mail_hub:/var/spool/mqueue /var/spool/mqueue
   # mount mail_hub:/var/spool/mail /var/spool/mail
   ```

5. Start the sendmail program:

   ```
   # /sbin/init.d/sendmail start
   ```

6. Add the loopback mounts to the /etc/fstab file so they will mount on the next reboot. The lines in the /etc/fstab file should resemble the following:

   ```
   /var/spool/mqueue@mail_hub /var/spool/mqueue nfs rw,fg 0 0
   /var/spool/mail@mail_hub /var/spool/mail nfs rw,fg 0 0
   ```

   You must specify the fg option to ensure that /var/spool/mqueue is NFS-mounted before the sendmail program starts. Do not put the mount command into the background to retry the mount if the original mount fails, because the sendmail program could start before the mqueue area is mounted. This situation causes problems because sendmail tries to use the mount point for the mqueue area instead of the mounted file system.

5.5.6 Defining a BIND Mail Exchange Record

You can define a BIND mail exchanger (MX) record in a database file, such as the /etc/namedb/hosts.db file, on the primary BIND server to point to all your mail hubs. The sendmail program uses the BIND MX record to define a list of mail machines that can receive mail sent to a specific address. See the DIGITAL UNIX *Network Administration* manual for detailed information about BIND MX records.

The sendmail program delivers the mail to the machine with the lowest specified preference, if possible. If that machine is not available, it tries the machine with the next lowest preference, and so on.

You can specify both mail hub member systems as the mail exchange for each system. The following example shows the mail exchange resource records:

```
;name  ttl class type preference deliver-to

server1.foo.com IN MX 1 server1.foo.com

   IN MX 100 server2.foo.com

server2.foo.com IN MX 1 server2.foo.com

   IN MX 100 server1.foo.com

mail_server.foo.com IN MX 100 server1.foo.com

   IN MX 100 server2.foo.com
```

In this example, all mail going to server1 goes to server1 if it is available, because it has a preference of 1. If server1 is unavailable, then the mail goes to server2, which delivers the mail to the shared /var/spool/mail area. Using this configuration, mail continues to be delivered to either mail hub member system internal address as long as at least one mail hub member system is available.

5.6 Accessing NFS Services from Client Systems

To access a Network File System (NFS) service from a client system, you must edit two system files:

- You must include the NFS service name and Internet address in a client system's local /etc/hosts file.
- You must edit the /etc/fstab file on a client system and include the NFS service name and its exported mount points. See Section 5.3 for information on ASE exports files for services.

6

Setting Up a Disk Service

A disk service includes one or more file systems, Advanced File System (AdvFS) filesets, or Logical Storage Manager (LSM) volumes that are highly available. Disk services can also include a disk-based application that is highly available. An example of an application that you can use in a disk service is a database application.

You can assign a disk service name its own Internet address. The member system that runs the service responds to this address. This makes the service autonomous and not dependent on the availability of any particular member system. Clients access the service by including the service name and the exported directory path in their `/etc/fstab` file. If the service stops on a member system, it fails over to a viable system, and clients only experience a short timeout.

If you assign an Internet address to a disk service name, you can use the POLYCENTER NetWorker Save and Restore (NetWorker) to back up the service's storage. NetWorker treats the disk service as an independent client and stores the storage indexes under the name of the service. This enables you to back up and recover the service's storage independent of the member system running the service. See the NetWorker documentation for information about using NetWorker to back up a disk service's storage.

Before you set up a disk service, you must prepare the disks to be used in the service and install any application used in the service. To fail over an application, in addition to disk, at a minimum, you must create a user-defined start action script that includes the commands to start the application, and create a user-defined stop action script that includes the commands to stop the application. See Chapter 4 for more information about preparing disks, applications, and action scripts for a service.

Note that a disk service is not the same as a distributed raw disk (DRD) service. Disk services typically involve file system usage, while DRD services provide access to raw physical disks throughout a Production Server cluster.

6.1 Disk Service Requirements

Disk services have the following requirements:

- Disk service names that are also Internet Protocol (IP) names must be included in the local `/etc/hosts` file on each member system before you set up the disk service.
- You cannot use a disk service IP name that is the same as the name of any system. Service names and member system names must be unique.
- A disk service IP name must adhere to the conventions for naming a system, as described in the DIGITAL UNIX *Installation Guide*.
- A disk service IP name can be associated with any IP subnet that is directly connected to all the member systems. If you are using a distributed database lookup service such as the Network Information Service (NIS), be sure that the service name information is local to all the member systems. To do this, make all the member systems either master or slave servers, or specify the service name information in the local `/etc/hosts` file. Be sure your `/etc/svc.conf` file specifies `local` as the first entry.
- Do not manually mount or dismount disks that will be used in a disk service.
- To use NetWorker to back up disk service data, you must specify a mount point for each file system, fileset, or volume when you add the service.
- If you do not specify a mount point for an Advanced File System (AdvFS) fileset, a member system may panic unless the following conditions are met:
  - Before you add the disk service, make sure that the fileset is not already mounted.
  - If you mount a fileset in your own user-defined action scripts, make sure that the user-defined stop action script unmounts the fileset and returns an error code if the unmount fails.

6.2 Disk Service Components

When you add a disk service, you can specify the following information:

- Service name—A service name must be unique and cannot contain a slash (/). Optionally, you can specify a disk service name that is also an Internet Protocol (IP) host name. The IP host name must be specified in each member system’s local `/etc/hosts` file. See `hosts`(4) for more information. Service names that are IP host names must adhere to the conventions for naming a system, as described in the DIGITAL UNIX *Installation Guide*.

- Automatic Service Placement (ASP) policy—See Chapter 4 for information about the ASP policies.
- UNIX file systems, AdvFS fileset names, or Logical Storage Manager (LSM) volumes—See Chapter 4 for information about setting up disks.
- Mount point for each file system, fileset, or volume—Client systems will specify this mount point in their `/etc/fstab` files to access the service's file systems. You can specify NONE if you do not want the TruCluster software to automatically mount the file system, fileset, or volume.
- Mode (either read/write or read-only) and any mount options other than the default options for each file system, fileset, or volume.
- User-defined action scripts to fail over any application—See Chapter 4 for information about creating action scripts.

6.3 Adding a Basic Disk Service

Example 6–1 shows how to add a disk service that uses an Internet Protocol (IP) name, an Advanced File System (AdvFS) fileset on a Logical Storage Manager (LSM) volume, and pathnames for action scripts to start and stop the service.

Example 6–1: Adding a Disk Service Using Your Own Action Scripts

```
# asemgr
.
.
.
      Adding a service

Select the type of service:

    1)  NFS service
    2)  Disk service
    3)  User-defined service
    4)  DRD service
    5)  Tape service

    q)  Quit without adding a service
    x)  Exit                 ?)  Help

Enter your choice [1]: 2

You are now adding a new disk service to the ASE.

A disk service consists of a disk-based application and disk
configuration that are failed over together.  The disk
configuration can include UFS file systems, AdvFS filesets,
LSM disk groups, or raw disk information.

                    Disk Service Name

The name of a disk service must be a unique service name. Optionally,
an IP address may be assigned to a disk service.  In this case, the
name must be a unique IP host name set up for this service and
must be in the local hosts database on all the ASE member systems.
```

Example 6–1: Adding a Disk Service Using Your Own Action Scripts (cont.)

```
Enter the disk service name: disk1

Assign an IP address to this service? (y/n): y

Checking to see if disk1 is a valid host...

                  Specifying Disk Information

Enter one or more device special files, AdvFS filesets, or LSM
volumes to define the disk storage for this service.

    For example:     Device special file:   /dev/rz3c
                     AdvFS fileset:         domain1#set1
                     LSM volume:            /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.

Enter a device special file, an AdvFS fileset, or an LSM volumeas storage for this
service (press ’Return’ to end): domain2#fs1

AdvFS domain ‘domain2‘ has the following volumes:
     /dev/vol/dg3/vol01

Is this correct (y/n) [y]: y

Following is a list of device(s) and pubpath(s) for disk group dg3:

        DEVICE  PUBPATH
        rz20c   /dev/rz20c

Is this correct (y/n) [y]: y

                  Mount Point

The mount point is the directory on which to mount‘domain2#fs1‘.
If you do not want it mounted, enter "NONE".

Enter the mount point or NONE: /fs1_disk1

              AdvFS Fileset Read-Write Access

Mount ‘domain2#fs1‘ fileset with read-write or read-only access?

    1)  Read-write
    2)  Read-only

Enter your choice [1]: 2

              AdvFS Mount Options Modification

Enter a comma-separated list of any mount options you want to use
for ‘domain2#fs1‘ fileset (in addition to the defaults listed in
the mount.8 reference page).  If none are specified, only the
default mount options are used.

Enter options (Return for none): Return

Enter a device special file, an AdvFS fileset, or an LSM volume as
storage for this service (press ’Return’ to end):  Return

Modifying user-defined scripts for ‘disk1‘:

    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action

    x)  Exit - done with changes

Enter your choice [x]: 1
```

Example 6–1: Adding a Disk Service Using Your Own Action Scripts (cont.)

```
Modifying the start action script for ‘disk1‘:

    a)  Add a start action script
    e)  Edit the start action script
    g)  Modify the start action script arguments []
    t)  Modify the start action script timeout [60]
    r)  Remove the start action script
    x)  Exit - done with changes

Enter your choice [x]: a

Enter the full pathname of your start action script or "default"
for the default script (x to exit): /usr/sbin/dbase_account

Enter the argument list for the start action script (x to exit): start

Enter the timeout in seconds for the start action script [60]: Return

Modifying the start action script for ‘disk1‘:

    f)  Replace the start action script
    e)  Edit the start action script
    g)  Modify the start action script arguments [start]
    t)  Modify the start action script timeout [60]
    r)  Remove the start action script
    x)  Exit - done with changes

Enter your choice [x]: x

Modifying user-defined scripts for ‘disk1‘:

    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action

    x)  Exit - done with changes

Enter your choice [x]: 2

Modifying the stop action script for ‘disk1‘:

    a)  Add stop action script
    e)  Edit the stop action script
    g)  Modify the stop action script arguments []
    t)  Modify the stop action script timeout [60]
    r)  Remove the stop action script
    x)  Exit - done with changes

Enter your choice [x]: a

Enter the full pathname of your stop action script or "default"
for the default script (x to exit): /usr/sbin/dbase_account

Enter the argument list for the stop action script (x to exit): stop

Enter the timeout in seconds for the stop action script [60]: Return

Modifying the stop action script for ‘disk1‘:

    f)  Replace the stop action script
    e)  Edit the stop action script
    g)  Modify the stop action script arguments [stop]
    t)  Modify the stop action script timeout [60]
    r)  Remove the stop action script
    x)  Exit - done with changes

Enter your choice [x]:  Return

Modifying user-defined scripts for ‘disk1‘:
```

Example 6–1: Adding a Disk Service Using Your Own Action Scripts (cont.)

```
    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action
    x)  Exit - done with changes

Enter your choice [x]: x

    Selecting an Automatic Service Placement (ASP) Policy

Select the policy you want ASE to use when choosing a member
to run this service:

    b)  Balanced Service Distribution
    f)  Favor Members
    r)  Restrict to Favored Members
    x)  Exit to service config menu      ?)  Help

Enter your choice [b]: r

        Selecting an Automatic Service Placement (ASP) Policy

Select the favored member(s) IN ORDER for service ’disk1’:

    1)  gideon
    2)  toto
    x)  No favored members               ?)  Help

Enter a comma-separated list [x]: 2

        Selecting an Automatic Service Placement (ASP) Policy

Do you want the favored member to change to the one that is
specified when a manual relocation is performed (y/n/?): y

Enter ’y’ to add Service ’disk1’ (y/n): y

Adding service...
Starting service...
Saving the updated database...
Service successfully added...
```

6.4 Accessing Disk Services from Client Systems

To access a disk service that uses an Internet Protocol (IP) host name for a service name, client systems must include the disk service name and Internet address in their local `/etc/hosts` files.

7

Setting Up a User-Defined Service

A user-defined service consists only of an application that you want to fail over using your own action scripts. For example, you could set up a highly available Internet login service. The application in a user-defined service cannot use disks. If your application is disk-based, set up a disk service or Network File System (NFS) service instead.

Before you set up a user-defined service on each member system, you must install the application that you want to fail over. You also must set up the user-defined action scripts that the TruCluster software uses to fail over the application. See Chapter 4 for more information about installing applications and creating action scripts for a service.

7.1 User-Defined Service Components

When you add a user-defined service, you can specify the following information:

- Service name—Service names must be unique and cannot contain a slash (/).
- Automatic Service Placement (ASP) policy—See Chapter 4 for information about the ASP policies.
- User-defined action scripts to fail over your application—At a minimum, you must specify a start action script that includes the commands to start the application and a stop action script that includes the commands to stop the application. See Chapter 4 for more information about action scripts.

7.2 Adding a Basic User-Defined Service

Example 7–1 shows how to add a basic user-defined service. This example shows how to specify pathnames for the user-defined start and stop action scripts. In this example, the “Adding service” menu is accessed from the `asemgr` Service Configuration menu by choosing the “Adding a new service” item.

Example 7–1: Adding a Basic User-Defined Service

```
# asemgr

⋮

        Adding a service



Select the type of service:



    1)  NFS service
    2)  Disk service
    3)  User-defined service
    4)  DRD service
    5)  Tape service

    q)  Quit without adding a service
    x)  Exit                   ?)  Help


Enter your choice [1]: 3


You are now adding a new user-defined service to the ASE.


                User-Defined Service Name

The name of a user-defined service must be a unique service name
within the ASE.

Enter the user-defined service name: start1

Modifying user-defined scripts for ‘start1‘:

    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action
    5)  Check action
    x)  Exit - done with changes

Enter your choice [x]: 1

Modifying the start action script for ‘start1‘:

    f)  Replace the start action script
    e)  Edit the start action script
    g)  Modify the start action script arguments [start1]
    t)  Modify the start action script timeout [60]
    r)  Remove the start action script
    x)  Exit - done with changes

Enter your choice [x]: f

Enter the full pathname of your start action script or "default"
for the default script (x to exit): /usr/sbin/start1_start

Modifying the start action script for ‘start1‘:
```

Example 7–1: Adding a Basic User-Defined Service (cont.)

```
    f)  Replace the action script
    e)  Edit the start action script
    g)  Modify the start action script arguments [start1]
    t)  Modify the start action script timeout [60]
    r)  Remove the start action script
    x)  Exit - done with changes

Enter your choice [x]: x

Modifying user-defined scripts for ‘start1‘:

    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action
    5)  Check action
    x)  Exit - done with changes

Enter your choice [x]: 2

Modifying the stop action script for ‘start1‘:

    f)  Replace stop action script
    e)  Edit the stop action script
    g)  Modify the stop action script arguments [start1]
    t)  Modify the stop action script timeout [60]
    r)  Remove the stop action script
    x)  Exit - done with changes

Enter your choice [x]: f

Enter the full pathname of your stop action script or "default"
for the default script (x to exit): /usr/sbin/start1_stop

Modifying the stop action script for ‘start1‘:

    f)  Replace the stop action script
    e)  Edit the stop action script
    g)  Modify the stop action script arguments [start1]
    t)  Modify the stop action script timeout [60]
    r)  Remove the stop action script
    x)  Exit - done with changes

Enter your choice [x]: x

Modifying user-defined scripts for ‘start1‘:

    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action
    5)  Check action
    x)  Exit - done with changes

Enter your choice [x]: x


        Selecting an Automatic Service Placement (ASP) Policy
```

Example 7–1: Adding a Basic User-Defined Service (cont.)

```
Select the policy you want ASE to use when choosing a member
to run this service:

    b)  Balanced Service Distribution
    f)  Favor Members
    r)  Restrict to Favored Members

    x)  Exit to service config menu      ?)  Help

Enter your choice [n]: b

        Selecting an Automatic Service Placement (ASP) Policy

Do you want the ASE to relocate this service to a more highly
favored member if the member becomes available while the service
is running (y/n/?): n

Enter ’y’ to add Service ’start1’ (y/n): y

Service successfully added...
```

7.3 Adding a User-Defined Login Service

You can set up a user-defined network or login service that uses a pseudohost name as a service name. Users can log in by using the pseudohost name and perform network operations on the host. The pseudohost name has an Internet address and resembles other hosts. The TruCluster software aliases the pseudohost name to the member system that is running the login service.

To add a user-defined login service to your ASE, you must set up a user-defined start action script that invokes the `ifconfig alias` command and a user-defined stop action script that invokes the `ifconfig -alias` command. See `ifconfig`(8) for a description of this command. To make this easy, the TruCluster software provides you with a script, `/var/ase/sbin/nfs_ifconfig`, that can establish and remove a host name alias.

You can invoke the `/var/ase/sbin/nfs_ifconfig` script in your user-defined start and stop action scripts to start and stop the login service. To start the service, invoke the `/var/ase/sbin/nfs_ifconfig` script with the `start` argument. To stop the service, invoke the script with the `stop` argument.

To set up a user-defined login service, you must first add the pseudohost name to the `/etc/hosts` file on all the member systems. For example, to add a login service using the pseudohost name `ase10`, specify a line similar to the following in the `/etc/hosts` file on all the members:

```
6.140.64.52    ase10.ift.tec.com ase10
```

Run the `asemgr` utility to add the login service, using the pseudohost name as the service name. Use an Automatic Service Placement (ASP) policy that does not allow the service to relocate unless a member system fails, because users are logged out if the login service relocates.

Example 7–2 shows how to add a login service. This example shows how to edit the default user-defined start and stop action scripts and specify the `/var/ase/sbin/nfs_ifconfig` script with the appropriate argument.

Example 7–2: Adding a Login Service

```
# asemgr
.
.
.
        Adding a service

Select the type of service:

    1)  NFS service
    2)  Disk service
    3)  User-defined service
    4)  DRD service
    5)  Tape service

    q)  Quit without adding a service
    x)  Exit                    ?)  Help

Enter your choice [1]: 3

You are now adding a new user-defined service to the ASE.

                  User-Defined Service Name

The name of a user-defined service must be a unique service name
within the ASE environment.

Enter the user-defined service name:ase10
Modifying user-defined scripts for ‘ase10‘:


    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action
    5)  Check action
    x)  Exit - done with changes

Enter your choice [x]: 1

Modifying the start action script for ‘ase10‘:

    f)  Replace the start action script
    e)  Edit the start action script
    g)  Modify the start action script arguments [ase10]
    t)  Modify the start action script timeout [20]
    r)  Remove the start action script
    x)  Exit - done with changes

Enter your choice [x]: f

Enter the full pathname of your start action script or "default"
```

Example 7–2: Adding a Login Service (cont.)

```
for the default script (x to exit): default

Modifying the start action script for ‘ase10‘:

    f)  Replace the start action script
    e)  Edit the start action script
    g)  Modify the start action script arguments [ase10]
    t)  Modify the start action script timeout [60]
    r)  Remove the start action script
    x)  Exit - done with changes

Enter your choice [x]: e

#!/bin/sh
#
#
PATH=/sbin:/usr/sbin:/usr/bin

export PATH

ASETMPDIR=/var/ase/tmp

if [ $# -gt 0 ]; then
        svcName=$1
else
        svcName=
fi

#

# the nfs_ifconfig script will send any
# stdout/stderr to the following log file:
#
LOGGER=/var/ase/tmp/childLog.$$

rm -f ${LOGGER}

#
# Run the ase ifconfig script to start the alias.
# This will do an
# ifconfig <interface_id> alias ${svcName} to
# to get the login service going
#
/var/ase/sbin/nfs_ifconfig $$ start ${svcName} returnValue=$?
#
# If anything in the logger cat it to stdout which will then be
# sent to the syslog daemon.log
#
if [ -f ${LOGGER} ]; then
        cat ${LOGGER}
fi

rm -f ${LOGGER}

#
# exit with the return value of the nfs_ifconfig command.
#
exit ${returnValue}

:wq

Modifying the start action script for ‘ase10‘:

    f)  Replace the start action script
    e)  Edit the start action script
    g)  Modify the start action script arguments [ase10]
    t)  Modify the start action script timeout [60]
    r)  Remove the start action script
    x)  Exit - done with changes

Enter your choice [x]:  Return
```

Example 7–2: Adding a Login Service (cont.)

```
Modifying user-defined scripts for ‘ase10‘:

    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action
    5)  Check action
    x)  Exit - done with changes

Enter your choice [x]: 2

Modifying the stop action script for ‘ase10‘:

    f)  Replace the stop action script
    e)  Edit the stop action script
    g)  Modify the stop action script arguments [ase10]
    t)  Modify the stop action script timeout [60]
    r)  Remove the stop action script
    x)  Exit - done with changes

Enter your choice [x]: f

Enter the full pathname of your stop action script or "default"
for the default script (x to exit): default

Modifying the stop action script for ‘ase10‘:

    f)  Replace the stop action script
    e)  Edit the stop action script
    g)  Modify the stop action script arguments [ase10]
    t)  Modify the stop action script timeout [60]
    r)  Remove the stop action script
    x)  Exit - done with changes

Enter your choice [x]: e

#!/bin/sh
#
#
PATH=/sbin:/usr/sbin:/usr/bin
export PATH
ASETMPDIR=/var/ase/tmp

if [ $# -gt 0 ]; then
        svcName=$1
else
        svcName=
fi

#
# the nfs_ifconfig script will send any
# stdout/stderr to the following log file.
#
LOGGER=/var/ase/tmp/childLog.$$
rm -f ${LOGGER}
#
# Run the ase ifconfig script to stop the alias.
# This will do an
# ifconfig <interface_id> -alias ${svcName} to
# get the login service stopped
#
/var/ase/sbin/nfs_ifconfig $$ stop ${svcName} returnValue=$?
#
# If anything in the logger cat it to stdout which will then be
# sent to the syslog daemon.log
#
if [ -f ${LOGGER} ]; then
        cat ${LOGGER}
```

Example 7–2: Adding a Login Service (cont.)

```
fi

rm -f ${LOGGER}

#
# exit with the return value of the nfs_ifconfig command.
#
exit ${returnValue}
:wq

Modifying the stop action script for ‘ase10‘:

    f)  Replace stop action script
    e)  Edit the stop action script
    g)  Modify the stop action script arguments [ase10]
    t)  Modify the stop action script timeout [60]
    r)  Remove the stop action script
    x)  Exit - done with changes

Enter your choice [x]: x

Modifying user-defined scripts for ‘ase10‘:

    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action
    5)  Check action
    x)  Exit - done with changes

Enter your choice [x]: x

        Selecting an Automatic Service Placement (ASP) Policy

Select the policy you want ASE to use when choosing a member
to run this service:

    b)  Balanced Service Distribution
    f)  Favor Members
    r)  Restrict to Favored Members

    x)  Exit to service config menu      ?)  Help

Enter your choice [b]:  Return

        Selecting an Automatic Service Placement (ASP) Policy

Do you want ASE to relocate this service to a more highly favored
member if one becomes available while the service is running (y/n/?):n

Enter ’y’ to add Service ’ase10’ (y/n): y

Adding service...

Starting service...

Saving the updated database...

Service successfully added...
```

8

Setting Up a DRD Service (PS)

The distributed raw disk (DRD) subsystem of a TruCluster Production Server cluster allows a disk-based, user-level application to run within a cluster, regardless of where in the cluster the physical storage it depends upon is located. A DRD service enables you to provide to applications, such as database and transaction processing (TP) monitor systems, parallel access to storage media from multiple cluster members.

Applications that perform I/O involving sets of large data files, random access to records within these files, and concurrent read/write data sharing can benefit from using the features of DRD. The DRD subsystem driver is a pseudodevice driver that provides an abstraction of the physical storage throughout the cluster.

The available server environment (ASE) manager utility (`asemgr`) allows you to set up a DRD service within an ASE in a cluster, and make it highly available and eligible for failover among the cluster member systems within that ASE. When creating a DRD service, you specify the physical media that the service will provide clusterwide. The `asemgr` utility sets up the service name and the device special files by which the service is accessed. After a DRD service has been established within a given ASE, cluster members both within and outside that ASE can access the disk storage it provides.

This chapter discusses the following topics:

- The key concepts of the DRD subsystem implementation (Section 8.1)
- The semantics of DRD device volume names (Section 8.2)
- How to add a DRD service to a cluster (Section 8.3)
- How to modify an existing DRD service (Section 8.4)
- How to remove a DRD service from a cluster (Section 8.5)
- How to tune the DRD subsystem (Section 8.6)
- How to measure and test the DRD subsystem's performance (Section 8.7)
- Other considerations for managing DRD services (Section 8.8)
- How to locate and interpret DRD messages and troubleshoot the DRD subsystem (Section 8.9)

8.1 Overview of the DRD Subsystem Driver

The distributed raw disk (DRD) subsystem driver provides character device driver interfaces, receiving user requests through conventional system calls such as open, close, read, write, and ioctl. The DRD subsystem provides only raw disk capabilities: that is, file systems cannot be mounted on DRD devices.

Upon receipt of a user-level request, the DRD driver determines which member system is the server of the physical device, with the following results:

- If the physical device is being served by the member system that received the user request, it is the server and the request is considered a local request. The local request is passed on to the underlying physical device driver (for instance, the SCSI CAM driver or LSM driver).
- If the physical device is not being served by the member system that received the user request, another member system is the server and the request is considered a remote request. That member system could be in the same available server environment (ASE) as the server, or it could be in a different ASE within the cluster. A remote request is sent across a network transport to the server. The server then passes the request to the underlying physical device driver. When the local physical driver completes the request, the server returns the results and status to the client. Finally, the client returns results and status to the calling user-level program.

8.2 DRD Namespace

A distributed raw disk (DRD) device is a device special file that provides clusterwide access to a single underlying physical device (either a SCSI disk or Logical Storage Manager (LSM) volume). There is a one-to-one correspondence between a DRD device and an underlying physical device.

The asemgr, drd_mknod, and drd_ivp utilities create the DRD device special files that disk-based, clusterwide application programs open and to which they issue read and write system calls. These files reside in the /dev/rdrd/ directory and are assigned names such as drd1, drd2, and drd30. (The DRD subsystem reserves device number 0 for subsystem control purposes.)

The collection of DRD device special files forms the DRD namespace. As DRD services are added, the asemgr utility assigns DRD special file names sequentially. For example, if the file names /dev/rdrd/drd1, /dev/rdrd/drd2, and /dev/rdrd/drd3 are in use, the next new DRD service will be added as /dev/rdrd/drd4. To minimize holes in the DRD namespace, the asemgr utility reuses DRD device numbers for services that have been deleted.

The DRD namespace must be unique across the cluster. Therefore, in cluster configurations that include multiple available server environments (ASEs), the /dev/rdrd/drd1 special file must refer to the same disk regardless of which ASE is serving it. For this reason, the asemgr utility partitions the DRD namespace on a per-ASE basis. The first ASE is limited to DRD numbers drd1 through drd9999; the second ASE is limited to DRD numbers drd10000 through drd19999; and so on. The DRD namespace implementation accommodates a cluster consisting of 64 separate ASEs. Each ASE can have 9999 separate DRD services.

The asemgr utility creates the DRD device special files corresponding to DRD services within a single ASE. In cluster configurations consisting of a single ASE, the cluster administrator need not perform any explicit tasks to create these special files throughout the cluster. However, in cluster configurations consisting of multiple ASEs, the asemgr utility cannot create the special files for DRD services provided by one ASE on cluster members outside the serving ASE. In these configurations, the cluster administrator must perform the procedures detailed in Section 8.3, Section 8.4, and Section 8.5 to manage DRD services clusterwide.

Do not use the mknod command to create and assign names to DRD device special files. Unlike the SCSI device driver, the DRD device driver dynamically obtains a major number during system startup. This number can change each time the system is rebooted, and it can vary from member system to member system. For example, if you used the mknod to create a DRD device special file, its file handle would become stale at the next reboot because its major number will have changed. In some instances, you might obtain inconsistent results when using mknod for this purpose, including data corruption.

If you need to adjust DRD special filenames, use the drd_mknod command. If you need to create your own form of the DRD device special file namespace which better matches your usage model, set up symbolic links to the actual DRD device special files.

8.3 Adding a DRD Service

As described in Section 8.2, a distributed raw disk (DRD) device is a device special file that provides clusterwide access to a single, underlying physical device (either a SCSI disk or LSM volume).

A DRD service is a highly available service created and managed by means of the asemgr utility. A DRD service consists of one or more underlying DRD devices (each of which represents a single SCSI disk or LSM volume).

As a result, there is not necessarily a one-to-one correspondence between a DRD service and an underlying physical device.

A cluster administrator may group various physical devices into a single DRD service for functional or organizational reasons. For example, an administrator may create a DRD service consisting of all physical devices within a single storage enclosure to force them to be served by a single member system within the available server environment (ASE).

For systems with large numbers of physical devices, it is often useful to group devices into DRD services to limit the number of services that need to be maintained. Note that when a DRD service consists of multiple underlying physical devices, all devices are relocated and failed over as a group. As a result, any DRD service is available only if all of its underlying physical devices are operational; failure of any underlying physical device disables a DRD service.

Note

> When a DRD service consists of multiple underlying physical devices, the devices will not be considered for relocation by the `drd_balance` utility. See Section 8.6.1 for a discussion of DRD service placement and a description of the `drd_balance` utility.

Similarly, for DRD services that use Logical Storage Manager (LSM) volumes, the unit of failover and relocation is the LSM disk group. All disks within the disk group are relocated as one set. Because of this, you cannot assign separate LSM volumes within the same disk group to separate DRD services. However, you can have a single DRD service that consists of multiple LSM volumes within the same disk group, and you can employ multiple disk groups in a single DRD service.

When you add a DRD service, the `asemgr` utility prompts you for the following information:

- DRD service name—The name can be up to 64 characters long. The `asemgr` utility uses a single service name to identify and manage a DRD service, regardless of how many underlying devices participate in it.
- Physical disk(s) or LSM volume(s) to be used—You can associate a single physical disk with no more than one DRD service. Although you can specify multiple, nonoverlapping disk partitions in the device special file names you supply when creating a single DRD service, you cannot assign multiple partitions of the same physical disk to multiple DRD services. The `asemgr` utility assigns a discrete DRD device special file to each underlying physical device participating in the service.

The more disks that are configured into a single DRD service, the longer it takes for the service to relocate. For this reason, DIGITAL recommends that no more than 50 DRD disks participate in a single service.

An application accesses DRD devices by referring to the DRD device special files, and does not use the DRD service name. If you specify an LSM volume for the underlying physical device, the utility asks you to confirm the list of physical devices that comprise the associated disk group. This question is meant to remind you of the underlying storage configuration.

- Automatic Service Placement (ASP) policy to use for the service—Note that, when a DRD service consists of multiple underlying physical devices, all devices are relocated and failed over as a group.

After you enter the required information, the `asemgr` utility displays the physical devices that you selected, the DRD service name, and the device special files that will be used to provide the service.

Note that the DRD service name and the DRD device special file names are defined separately. The DRD service name is assigned by the system administrator, and the DRD device special file names are automatically assigned by the DRD subsystem. There is no direct correlation between the two names. For example, the administrator may select the name `drd2` as the DRD servce name. However, if this is the first DRD service in the cluster, it is likely that the DRD subsystem will automatically assign a device special file name of `drd1` to the service. It is also possible to have a single DRD service name that contains multiple disk devices. In this case, there is not a one-to-one mapping between a service name and a device special file name.

After it creates a unique DRD device special file for each underlying physical device or LSM volume participating in the new service (for example, `drd20004`), the `asemgr` utility ensures that member systems within the same ASE as the DRD server know the service by creating the appropriate DRD device special file (for instance, `/dev/rdrd/drd20004`) on all member systems within the same ASE as the server. In other words, the device special file is created on all member systems with the same ASE_ID as the server.

On each member system that is outside of the available server environment (ASE) providing the DRD service, you must execute the `drd_mknod` command to create the device special files. As a last step in DRD service configuration, the `asemgr` utility tells you the command you must execute on each member system that is not in the server's ASE. For example:

```
NOTE: In order to access this DRD service from cluster members outside
      of this ASE execute the following on each node which is
```

```
not a member of this ASE:
   drd_mknod -f drd20004
```

To configure a large cluster that consists of multiple ASEs with large numbers of DRD devices, first create within each ASE the DRD services you intend that ASE to provide. Next, use the `drd_ivp` utility on each cluster member. The `drd_ivp -r` command automates the creation of DRD device special files and provides a faster and less error-prone mechanism to configure large clusters than the `drd_mknod` utility. The `drd_ivp -r` command compiles a list of member systems within the cluster, polls each ASE for its DRD service configuration, and invokes the `drd_mknod` utility, as needed, to create all required DRD device special files on the member system from which it is executed.

Example 8–1 shows how to add a DRD service.

Example 8–1: Adding a DRD Service

```
# asemgr
.
.
.
        Adding a service

Select the type of service:

    1)  NFS service
    2)  Disk service
    3)  User-defined service
    4)  DRD service
    5)  Tape service

    q)  Quit without adding a service
    x)  Exit                    ?)  Help

Enter your choice [1]: 4

You are now adding a new DRD disk service to your ASE.

A DRD disk service is comprised of any number of DRDs which can be
created from a single raw disk partition or LSM volume which will
be accessible from all members in the cluster.

Note: If using a raw disk partition please be sure that the character
      device special file exists on all members which are in this ASE.

                    DRD Service Name

The name of a DRD disk service must be a unique service name.
Enter the DRD disk service name: drd_svc_1

You will now be prompted to enter a list of devices comprising
the DRD service, select q when you have completed the list.

Enter an existing character device special file for one of the following:

        a physical device (ie /dev/rrz1c)
        a LSM volume (ie /dev/rvol/dg/vol01)
        To end the list, press the Return key at the prompt.

Enter character device special file: /dev/rrz9h

Enter an existing character device special file for one of the following:

        a physical device (ie /dev/rrz1c)
```

Example 8–1: Adding a DRD Service (cont.)

```
        a LSM volume (ie /dev/rvol/dg/vol01)
        To end the list, press the Return key at the prompt.

Enter character device special file:
DRD Device Special File:    /dev/rdrd/drd10011

Underlying Storage:         /dev/rrz9h

NOTE: In order to access the DRD device[s in this service from cluster
      members outside of this ASE execute the following on each cluster
      node which is not a member of this ASE:
                drd_mknod -f drd10011

        Selecting an Automatic Service Placement (ASP) Policy

Select the policy you want ASE to use when choosing a member
to run this service:

    b)  Balanced Service Distribution
    f)  Favor Members
    r)  Restrict to Favored Members

    x)  Exit to Service Configuration    ?)  Help

Enter your choice [b]:

        Selecting an Automatic Service Placement (ASP) Policy

Do you want ASE to consider relocating this service to another
 member if one becomes available while this service is running (y/n/?): y

Enter 'y' to add Service 'drd_svc_1'  (y/n): y

Adding service...

Starting service...

Saving the updated database...


Service drd_svc_1 successfully added...
```

Example 8–1 shows how to create a DRD service named `drd_svc_1` in the local ASE. I/O requests to the DRD device in this service use the raw disk interface for the physical device `rrz9h`. The DRD special file used to access the DRD device is `/dev/rdrd/drd10011`. To allow clusterwide access to the service, you must execute one of the following commands on each member system outside of the ASE in which the DRD service's server resides:

```
# drd_mknod -f drd1
```

```
# drd_ivp -r
```

8.4 Modifying a DRD Service

To modify the properties of a distributed raw disk (DRD) service, use the `asemgr` utility. The utility's `modify` option allows you to change the

Automatic Service placement (ASP) policy as well as the service configuration description. Changes to the service configuration description include the following:

- Adding, modifying, and deleting physical disks used by services
- Adding, modifying, and deleting LSM volumes used by a DRD services. See Section 10.6.3.3 for more information.
- Changing the service name

The following behavior is specific to modifying DRD services:

- The physical disks specified in DRD services can be disk partitions. If you are modifying a DRD service to add an additional disk partition, you must run the `asemgr` utility on the member system that provides the DRD service. This allows the necessary partition overlap checks to be performed.
- When a DRD service is modified, the file permissions, ownership, and group of all of its DRD device special files are reset to the default settings. If you have modified these attributes of the device special file, you must reset them on all member systems in the ASE.

Example 8–2 shows how to modify the configuration information for a DRD service without interrupting the service's availability.

Example 8–2: Modifying an Online DRD Service

```
  Service Configuration

    a)  Add a new service
    m)  Modify a service
    o)  Modify a service without interrupting its availability
    d)  Delete a service
    s)  Display the status of a service

    x)  Exit to Managing ASE Services    ?)  Help

Enter your choice [x]: o

 Online Service Modification

Select the service you want to modify:

    1)  ase1 on fgreg1
    2)  greg on fgreg2
    3)  drd1 on fgreg1

    x)  Exit to Service Configuration

Enter your choice [x]: 3

Select what you want to modify in service 'drd1':

    g)  General service information
    a)  Automatic service placement (ASP) policy
```

Example 8–2: Modifying an Online DRD Service (cont.)

```
    x)  Exit without modifications

Enter your choice [x]: g

The following lists the current configuration for DRD service "drd1"


Enter the option you wish to modify

     )  /dev/rdrd/drd1 -> /dev/rrz21b
    a)  Add LSM volume or physical disk
    d)  Delete LSM volume or physical disk
     )  Service name -> drd1
    q)  Quit without making any changes
    x)  Exit (done with modifications)

Enter your choice [x]:  Return

Enter ’y’ to modify service ’drd1’ (y/n): y
Stopping old service information...

Deleting old service information...

Adding new service information...

Starting new service information...

Service successfully updated.

 Storage configuration for DRD service ‘drd1‘

DRD Device Special File:    /dev/rdrd/drd1

Underlying Storage:         /dev/rrz21b

NOTE: In order to access the DRD devices in this service from cluster
      members outside of this ASE execute the following on each cluster
      node which is not a member of this ASE:

                drd_mknod -f drd1
```

8.5 Deleting a DRD Service

Use the asemgr utility to delete a distributed raw disk (DRD) service in the same manner as you would delete any other available server environment (ASE) service. See Section 10.7 for instructions.

When you delete a DRD service from an ASE, the asemgr utility deletes the corresponding device special file on all member systems within the same ASE as the DRD service’s server. To delete the service from member systems in other ASEs in the cluster, you must manually execute the drd_mknod command on each member system. As a last step in DRD service deletion, the asemgr utility provides you with the command you must execute on each member system that is not in the same ASE. For example:

```
NOTE: In order to remove the device special file associated with this
      service on cluster nodes which are not a member of this ASE,
      execute the following on each node which is not a member of this ASE:
                 drd_mknod -d -f drd1
```

If you do not run the drd_mknod -d command, stale DRD device special files remain on member systems outside the serving ASE. If an application attempts to access the stale DRD file, it finds no server for the request, and the request times out with an error.

The drd_ivp -r command does not delete device special files, which correspond to DRD services that have been deleted. A cluster administrator must use the drd_mknod utility to remove the device special files.

8.6 Tuning the DRD Subsystem

You can tune the performance of the distributed raw disk (DRD) subsystem by setting any of a number of DRD-related parameters in the /etc/sysconfigtab file. The default settings of these parameters should be sufficient for most applications. See drd(7) for a list of the parameters and additional information.

8.6.1 Locating DRD Services

To maximize the performance of DRD devices, set up a DRD service on the member system that will be initiating the majority of the I/O operations that utilize the service. Local requests for a DRD service are inherently faster than remote requests for the same service, because they bypass the remote communication codepath and access the DRD device directly.

If you can identify the member system that issues the most I/O requests to a DRD device, use the asemgr utlity to identify that member system as the favored member participating in the DRD service's Automatic Service Placement (ASP) policy. When a given DRD service has many clients, it is difficult to identify a single major client, especially as the service's access patterns vary over time. For such DRD devices, specifying a favored member may not yield maximum performance.

In these cases, use the drd_balance utility for help in relocating the DRD service. The drd_balance utility periodically polls for I/O usage patterns for DRD devices, can make recommendations for optimally relocating DRD services, and can optionally attempt the recommended relocations itself.

Using the drd_balance utility on a given DRD service differs from selecting the available server environment (ASE) balanced service ASP. When the ASE balanced service ASP is selected, ASE tries to evenly distribute the number of services across member systems, as services are started. It does not take into account the actual system resources required to provide a service, and it cannot relocate a service that has already started. By contrast, the drd_balance utility will relocate a DRD service that has already started, based on its I/O access pattern.

Note

> When a DRD service consists of multiple underlying physical devices, the devices will not be considered for relocation by the drd_balance utility.

The actual observed performance benefits of using the drd_balance utility vary based on your operating environment. In cases where the DRD device access varies considerably over short periods of time, the benefits may be small. However, if access patterns are relatively constant, there could be considerable benefit in environments encountering system constraints. A suggested approach is to benchmark your application with and without running the drd_balance utility. You could then use the drd_balance utility once to obtain information about the DRD device usage patterns, and then use that information in designating favored members as the ASE servers.

See drd_balance(8) for additional information.

You must not run the drd_balance utility while a Logical Storage Mananger (LSM) volume is in the middle of a volsave or volrestore operation. The volume save and restore operations can interfere with the cluster's ability to properly relocate DRD services.

8.6.2 Analyzing Tunable Parameters

Invoking the drd_ivp utility with the –t flag collects and analyzes selected DRD performance statistics, which can reveal the cause of performance bottlenecks. It identifies potential performance problems and suggests a resolution for each. To obtain an optimal analysis of the DRD subsystem, run the drd_ivp utility with the –t flag while the cluster is under peak load. See drd_ivp(8) for additional details.

To change the default values of DRD attributes, specify entries in the /etc/sysconfigtab file. After specifying new values for these attributes, you must reboot the system for them to take effect. See drd(7) for a list of tunable DRD attributes and instructions for viewing their values and modifying them.

8.6.3 Enabling Peer-to-Peer DMA Support for DRD

Peer-to-peer DMA (direct memory addressing) is a performance enhancement that can be used on a DRD server machine that meets certain configuration restrictions. When peer-to-peer DMA is enabled on a DRD server, the data read from a disk is sent directly from the host storage controller on the peripheral component interconnect (PCI) bus to the MEMORY CHANNEL controller on the same PCI bus without transferring via main memory on the server machine. The CPU load on the server machine is thus diminished. (Peer-to-peer DMA may be used when reading a remote disk, but never when writing to it due to lack of hardware support in the MEMORY CHANNEL.)

In order for peer-to-peer DMA to be enabled on a DRD server, the host storage controller and the MEMORY CHANNEL controller must on the same PCI bus. Peer-to-peer DMA is a global attribute for DRD; that is, all host storage (for example, SCSI) and MEMORY CHANNEL controllers used by DRD must be on the same PCI bus. The `drd_dma` utility runs at boot time before the TruCluster software starts. This utility analyzes the hardware configuration and automatically enables peer-to-peer DMA, if the configuration restriction is met.

In some cases, the `drd_dma` utility does not enable peer-to-peer DMA when the hardware configuration would actually support it. You can manually enable peer-to-peer DMA by setting the value of the `drd-bss-rm-peer2peer` parameter in the `/etc/sysconfigtab` file. Make sure no DRD disks are active before making the modification to the `/etc/sysconfigtab` file; otherwise, the system may panic. See `drd_dma`(7) for more information.

8.7 Testing and Measuring the Performance of DRD

You can use the `diskx` utility, provided in the optional System Exercisers subset of the DIGITAL UNIX operating system, to test and measure the performance of distributed raw disk (DRD) devices. The `diskx` utility performs testing in the following functional areas, depending on the flags that are specified in its command line:

- Read testing
- Write testing
- Seek testing
- Performance analysis

- disktab entry verification

Other flags determine how the selected tests are run and specify test parameters.

You need root privilege to run the diskx utility. For a complete description of the tests it performs and a full list of the flags it accepts, enter the following command:

```
# /usr/field/diskx -h
```

The following example invokes the diskx utility to perform write testing on the /dev/rdrd/drd3 DRD device:

```
# cd /dev/rdrd

# /usr/field/diskx -f drd3 -w -X -x -max_xfer 8k -num_blocks 10000 -debug 1


DISKX - DEC OSF/1 Disk Exerciser.

Testing disk device drd3.

Program output level is 1.

Wed Mar  12 10:09:40 1997

-------------------------------------------------------------------------

Write Transfer Testing

This test verifies that writes will succeed.  The data is first
written to disk. After all writes have completed the data will be
read back for validation.  Since this test writes to the disk
there is potential for file system corruption if a file system
exists on the disk that is being tested.

Writes will be done using random size transfers.  The write
size will be randomly selected from the range 512 to 8192 bytes.
Writes will be issued to random locations on the disk.  To accomplish
this a seek will be issued before each write to force a write of a
different disk region.

Testing will continue until an interrupt signal is received.

Sequentially write to partition A.
Sequentially read verify partition A.
The initial write and read verification has succeeded.
Perform random writes to partition A
Random writes completed without error.
Perform random reads to partition A
Random reads completed without error.

Stopping testing due to receipt of a termination signal.
```

```
Disk Transfer Statistics

Part Seeks Seek_Er Writes Writ_Er MB_Write  Reads Read_Er MB_Read Data_Er
     29850       0  30000       0     39.1  29833       0    38.4       0
-------------------------------------------------------------------------
Wed Mar  12 10:18:47 1997
Terminating disk exerciser.
```

The `diskx` utility is useful for measuring observed DRD performance, but it does not produce the maximum possible read and write throughput. To achieve maximum read and write throughput, an application should use asynchronous I/O operations instead of synchronous `read` and `write` system calls. For a complete description of asynchronous I/O, see `aio_read`(3), `aio_write` (3), and the DIGITAL UNIX *Guide to Realtime Programming*.

8.8 Other DRD Administrative Concerns

You may need to take into account the following operational concerns when preparing the distributed raw disk (DRD) services in a cluster:

- A device contributing to a DRD service cannot contribute to any other available server environment (ASE) service. For example, you cannot specify a DRD device as part of another DRD service, disk service, tape service, or Network File System (NFS) service.
- Like the underlying device drivers on which it is layered (such as the SCSI CAM driver and the Logical Storage Manager (LSM) driver), the DRD subsystem itself does not guarantee to service requests in any given order. For example, if two cluster applications (executing either on the same member system or on separate member systems) issue writes to the same disk block at the same time, the DRD subsystem may complete the writes in any order, possibly with unintended results.

  If an application must share access to a given set of disk blocks with another application, and the ordering of their I/O requests is important, both applications should use distributed lock manager (DLM) services to lock and synchronize their access to the disk. DLM services are discussed in the TruCluster Production Server Software *Application Programming Interfaces* manual.
- After creating a DRD service, all I/O operations to the device should be performed using the DRD device special file name (for example, `/dev/rdrd/drd1`). Do not access the device by means of its underlying physical device name (for example, `/dev/rrz17c`). To protect against data corruption resulting from unsynchronized simultaneous access to a device, ensure that cooperating applications use DLM services to coordinate access to DRD device special files.

- The underlying physical device used in a DRD service must not be enabled for use by the Prestoserve™ I/O acceleration hardware. Prestoserve hardware consists of a local disk cache on the system providing the disk service. Other cluster members cannot directly access this cache. If the cluster member providing a DRD service fails, the service cannot be relocated to another member system without risk of data corruption because that system cannot access the cache contents.
- The underlying physical device used in a DRD service must not be attached to a disk controller that uses a volatile writeback cache. Use of a volatile writeback cache optimizes performance at the expense of fault tolerance. Certain failures, such as a power loss, will cause data to be lost, inasmuch as it had not been preserved in nonvolatile storage on the disk device.

8.9 Troubleshooting the DRD Subsystem

This section provides information to assist you in troubleshooting the distributed raw disk (DRD) subsystem of a cluster.

8.9.1 DRD Extensions to the kdbx Debugger

TruCluster software extends the `kdbx` debugger to allow it to display the contents of the DRD map table. You can use this feature on both crash dumps and the running system. (Note, though, that the `asemgr` utility is the preferred means of obtaining status information on DRD services.)

Note

The `kdbx` debugger is included the optional base system subset titled "Kernel Debugging Tools". You can use the DRD extension to the `kdbx` debugger only if you have previously installed that subset.

The `drd` extension to the `kdbx` debugger is defined as follows:

`drd` [*flags*] [*number*] Displays the DRD map table. Valid *flags* for the `drd` extension are as follows:

`–full` Displays all map entries in long form.

`–terse` Displays all map entries in brief form. This is the default.

The *number* argument causes the DRD map for only the specified DRD number to be displayed.

The following example shows the short form display of a DRD map listing:

```
(kdbx) pr /var/ase/sbin/drd

Name                Server   Local Name             Struct address

/dev/rdrd/drd1      rclu4    /dev/rvol/dg1/vol01    0xfffffc000ff54c80

/dev/rdrd/drd10001  rclu12   /dev/rrz9h             0xfffffc0005202000

/dev/rdrd/drd2      rclu4    /dev/rvol/dg1/vol02    0xfffffc0005202640

/dev/rdrd/drd10002  rclu12   /dev/rrz10h            0xfffffc000ff552c0
```

The following example shows the short form display of a map listing for a specific DRD number:

```
(kdbx)
pr /var/ase/sbin/drd 7

/dev/rdrd/drd7      rclu3    /dev/rvol/dg4/vol01    0xfffffc000d5aec80
```

The following example shows the long form display of a map listing for a specific DRD number:

```
(kdbx) pr /var/ase/sbin/drd 7 -full

-----------------------------------------------------

Name: /dev/rdrd/drd7

Minor Number: 7

Structure address: 0xfffffc000d5aec80

drd_local_devt: 0x0, (0, 0)

drd_local_bdev: 0x4200007, (66, 7)

Local Device Name: /dev/rvol/dg4/vol01

Server Hostname: rclu3

State Flags: 0x2

Management State Flags: 0x501

Ref Count: 0

Drain Pending: 0

Delete Count: 0

MC Node Number: 4

Maxphys: 65536

Spare: 0
```

8.9.2 Debugging a Nonoperational DRD Service

A DRD service can fail to work properly for a number of reasons as follows:

- The DRD subsystem was not properly installed or configured. See the TruCluster Software Products *Software Installation* manual for a discussion about how you can verify correct installation and configuration.
- A failure has occurred in another cluster subsystem.
- A failure has occurred in the DRD subsystem.
- The underlying physical disk (or Logical Storage Manager (LSM) volume) is nonoperational.

8.9.2.1 Verifying Cluster Operation

Because DRD errors are often a symptom of problems within other cluster subsystems, look for related problems, as follows:

- Check the console messages and error log files on each cluster member. Log files are typically found in the `/var/adm/syslog.dated` directory.
- Run the cluster installation verification procedure (`clu_ivp`) to see if it points out any abnormalities. See `clu_ivp`(8) for more information on the `clu_ivp` utility.

8.9.2.2 Verifying DRD Subsystem Operation

To verify that the various DRD system components are operational, use the following command:

```
# drd_ivp -p -v -c

Cluster Configuration Information

Hostname               ASE_ID   BSSD   BSSD    DRD    Lic
                                 Reg   Resp   Conf    Reg
-----------------------------------------------------------
mcclu11                     0    Yes    Yes    Yes    Yes
mcclu12                     0    Yes    Yes    Yes    Yes
mcclu3                      1    Yes    Yes    Yes    Yes
mcclu4                      1    Yes    Yes    Yes    Yes

DRD configuration validation tests succeeded.
ASE_ID validation tests succeeded.
```

For more information, see `drd_ivp`(8).

8.9.2.3 Checking the Status of DRD Services

Use the asemgr utility to query the status of a DRD service. Select the "Display the status of a service" option from the Managing ASE Services menu. The following example shows the status display for a DRD service named drd_svc_4:

```
        Status for DRD service ‘drd_svc_4‘

 Status:             Relocate:  Placement Policy:       Favored Member(s):
 on mcclu12           yes       Balanced_Services         None

        Storage configuration for DRD service ‘drd_svc_4‘

DRD Device Special File:    /dev/rdrd/drd4
Underlying Storage:         /dev/rrz13g

NOTE: In order to access the DRD devices in this service from cluster
      members outside of this ASE execute the following on each cluster
      node which is not a member of this ASE:
                drd_mknod -f drd4
```

Of particular importance in the asemgr utility’s output is the Status field. This field indicates that member system mcclu12 is the server. If the Status field indicates that the service is off line or unassigned, use the asemgr utility to try to bring the service on line.

Keep in mind that the status of a service can change from one moment to the next, given the load balancing and failover that can occur within an available server environment (ASE). For example, DRD service drd4 may be served by mcclu12 at the time of the status request, but it may be relocated to mcclu11 at the very next instant.

8.9.2.4 Tracking the Failure of a Specific Service

If a specific DRD service is not working properly, follow these steps to identify the problem:

1. Run the clu_ivp utility. This utility checks a wide range of cluster functions. See Section 2.6 for more information.
2. Verify network connectivity. Verify on each member system that you can ping the other member systems over the MEMORY CHANNEL interface (mc0). For example:

   ```
   # ping mcclu11

    PING mcclu11.sun.ra.com (4.0.0.11): 56 data bytes
    64 bytes from 4.0.0.11: icmp_seq=0 ttl=255 time=0 ms

    ----mcclu11.sun.ra.com PING Statistics----
    1 packets transmitted, 1 packets received, 0% packet loss
    round-trip (ms)  min/avg/max = 0/0/0 ms
   ```

3. Verify that the physical devices participating in the service are working. For example, in Section 8.9.2.3, the `asemgr` utility showed the following disk as participating in DRD service `drd4`:

   ```
   Underlying Storage:        /dev/rrz13g
   ```

   Use the `file` command on the member system that is serving the DRD service to verify that the disk device is at least minimally operational:

   ```
   # file /dev/rrz13g
   ```

   ```
   /dev/rrz13g: character special (8/21510) SCSI #1 RZ28 disk #104
                                                      (SCSI ID #5)
   ```

   This output shows that the device can identify itself as a SCSI disk of type RZ28. It is likely that the disk itself is operational. If the device were not present, or if a **SCSI device** reservation is being held by another member of the ASE, the `file` command would display output like the following:

   ```
   /dev/rrz30a:    character special (8/55296)
   ```

   Because this output does not show the disk type, the disk may not be configured. Check the system startup messages to see if the disk was identified along with the other disks.

4. Verify DRD device special file access. Use the `file` command on the member system that is serving the DRD service to verify that the DRD device special file name is properly recognized:

   ```
   # file /dev/rdrd/drd4
   ```

   ```
   /dev/rdrd/drd4: character special (65/4) SCSI #1 RZ28 disk #104
                                                 (SCSI ID #5)
   ```

   If the underlying physical device is an LSM volume, the device identification is different than the SCSI identification. For example:

   ```
   # file /dev/rdrd/drd2
   ```

   ```
   /dev/rdrd/drd2:  character special (66/2) special_device #255
   ```

   In the following example, the `file` command fails to identify the DRD service:

   ```
   # file /dev/rdrd/drd99
   ```

   ```
   /dev/rdrd/drd99: character special (63/99)
   ```

   In this case, it may take a long time before the `file` command completes. This type of output indicates that a user-level program has attempted to access a DRD service that has no server. The DRD subsystem attempts to determine which member is the nonexistent server, but it eventually gives up. This situation usually indicates the existence of stale files in the /dev/rdrd directory. You may also see the following error from the file command:

```
# file /dev/rdrd/drd99

file: Cannot get file status on /dev/rdrd/drd99.
/dev/rdrd/drd99: cannot open for reading
```

This message indicates that the device special file does not exist. For DRD services within the same ASE, the asemgr utility creates the appropriate device special files. If your cluster is composed of multiple ASEs, you must explicitly create the device special files on those member systems that are in other ASEs. To do this, enter a drd_mknod or drd_ivp –r command on these member systems.

5. After you verify access to the DRD devices in the DRD service from the server, enter the file command on the member systems to verify that they, too, have access to the DRD devices.

8.9.2.5 Examining DRD Device Special File Permissions

The asemgr utility creates DRD device special files with the permissions, ownership, and group as shown in the following example:

```
# ls -l /dev/rdrd/drd2

crw-r--r--   1 root     system    66,  2 May 19 07:17 /dev/rdrd/drd2
```

You may use the chmod, chown, or chrgp command to modify the permissions of the DRD device special files to make them more accessible. Note that if you change the permissions of the DRD device special files on one member system, these changes are not automatically propagated to other member systems. To allow highly available applications to fail over to other member systems and keep the same access permissions, you must modify the permissions in the same way on the device special files on each member system.

When a DRD service configuration is modified using asemgr, the DRD device special file permissions are reset to the default value.

8.9.2.6 Reading from the DRD Disk

After successfully completing the steps in the previous sections, try to read from the DRD disk. Enter the dd command first on the server and then on each other member system. The following example issues 20 8–KB read requests:

```
# dd if=/dev/rdrd/drd4 of=/dev/null bs=8k count=20

20+0 records in

20+0 records out
```

In this example, the 20 read operations succeeded.

In the following example, the dd command attempts to read from a DRD device for which there is no server. This command takes a long time to return with the error message, because the DRD retries the command in order to determine which member system is the server until it times out. An error such as this may occur if the DRD device special file (for example, /dev/rdrd/drd99) corresponds to a DRD service that is provided by another ASE that is not up, or a service that has been deleted from another ASE.

```
# dd if=/dev/rdrd/drd99 of=/dev/null bs=8k count=20

/dev/rdrd/drd99: No such device
```

8.9.2.7 Writing to the DRD Disk

In general, performing read requests is enough to verify correct DRD operation.

Caution

Although reading from a DRD disk is a rather harmless operation, writing to a disk can be destructive, unless you exercise the appropriate caution. Before you attempt to write to the disk, ensure that you will not be writing over valid data.

The following example performs 20 8–KB write requests:

```
# dd if=/dev/zero of=/dev/rdrd/drd4 bs=8k count=20

20+0 records in

20+0 records out
```

In this example, the 20 write operations succeeded.

A write to a DRD device may fail if there is a disk label on the underlying physical device. For example:

```
# dd if=/dev/zero of=/dev/rdrd/drd4 bs=8k count=20

dd write error: Read-only file system

8+0 records in

0+0 records out
```

In this example, none of the write operations succeeded.

The `Read-only file system` error message indicates that the cause of this problem was an attempt to write to the first block of a disk with a disk label. To write to block 0, you must delete the disk label by first placing the DRD service off line, and then zeroing the label, as follows:

```
# disklabel -z /dev/rrz13c
```

Note that this is not a DRD-specific behavior. The same error would have occurred had you specified `/dev/rrz13c` in the `dd` command line.

The `diskx` utility is useful for performing more comprehensive read/write data validation testing. See the description of using `diskx` with DRD in Section 8.7.

9

Setting Up a Shared Tape Service

An available server environment (ASE) tape service depends on a set of one or more tape devices. It may also include media changer devices and file systems. A tape service enables the system administrator to configure the POLYCENTER NetWorker server and the servers for other client/server-based applications, for failover. The tape drive(s), media changer(s), and file systems all fail over as a unit.

Eligible servers must be written to react appropriately to certain events on the shared SCSI bus, such as bus and bus device resets. Bus and device resets cause any tape device on the shared SCSI bus to rewind. Therefore, a tape server application will inspect the `errno` value and extended error information returned from its I/O call and reposition the tape. Note that because the commonly used utilities `tar`, `cpio`, `dump`, and `vdump` are not designed in this way, they may unexpectedly terminate when used within an ASE, due to a node state transition or other normal ASE event.

This chapter describes how to use the `asemgr` to add a tape service to an ASE. It describes the components of the service and includes an example of setting up and modifying a tape service.

9.1 Tape Service Requirements

Tape services have the following requirements:

- A tape service that has an Internet Protocol (IP) name must be included in the `/etc/hosts` file on each member system before you set up the tape service.
- Service names and member system names must be unique; you cannot use a name for the tape service that is the same as a system name.
- A tape service IP name must adhere to the conventions for naming a system, as described in the DIGITAL UNIX *Installation Guide*.
- A tape service IP name can be associated with any subnet directly connected to all the member systems. If you are using a distributed database lookup service, such as the Network Information Service (NIS), be sure that the service name information is local to all the member systems by making all the member systems either master or slave servers, or by specifying the service name information in the local

`/etc/hosts` file. Ensure that the `/etc/svc.conf` file specifies `local` as the first entry.

In addition to these requirements, consider adjusting the timeouts used by the tape driver. Some tape motion operations, such as rewinds, forward and backward spacing files and records, and writing file marks, may cause the tape driver to time out for long periods of time. As a result, certain failures that would trigger a service relocation may not be detected until the timeout expired. To avoid this problem, the administrator can set the timeouts used by the tape driver for particular tape device models in the`/etc/ddr.dbase` file.

When adjusting tape driver settings, ensure that raw tape devices are consistently named. See the DIGITAL UNIX *Installation Guide* for recommendations on creating consistent device special files for ASE shared tape service.

To adjust the timeouts used by a tape driver, follow these steps:

1. Modify the device type entries in the `/etc/ddr.dbase`file.
2. After making these changes to the `/etc/ddr.dbase`file, run `ddr_config -c` to make the changes take effect without a reboot. These settings will be picked up automatically on future reboots.

The following example shows sample entries in the `/etc/ddr.dbase` file. These changes would set the timeout values to 300 seconds (5 minutes). These settings should be placed after the PARAMETERS section of the particular tape entry you want to modify.

```
SCSIDEVICE
    #
    Type = tape
    Name = "DEC" "TZ30"
    #
    PARAMETERS:
        TypeSubClass        = tk
        MaxTransferSize     = 64512
        SyncTransfers       = disabled
        TagQueueDepth       = 0
        ReadyTimeSeconds    = 45                # seconds

    # Use "default" tape densities

    ATTRIBUTE:
        AttributeName = "BDRatBoot"
        Length        = 1
        ubyte[0]      = 1

    ATTRIBUTE:
        AttributeName = "REW_TIMEOUT"
        Length        = 8
        ubyte[0]      = 300

    ATTRIBUTE:
```

```
    AttributeName = "LOAD_TIMEOUT"
    Length        = 8
    ubyte[0]      = 300


ATTRIBUTE:
    AttributeName = "SPACE_TIMEOUT"
    Length        = 8
    ubyte[0]      = 300

ATTRIBUTE:
    AttributeName = "SEOD_TIMEOUT"
    Length        = 8
    ubyte[0]      = 300
```

9.2 Tape Service Components

When you add a tape service, you can specify the following information:

- Service name—Must be unique and cannot contain a slash (/). Optionally, you can specify a tape service name that is also an IP host name. The IP host name must be specified in each member system's `/etc/hosts` file. See `hosts`(4) for more information. Service names that are IP host names must adhere to the conventions for naming a system, as described in the DIGITAL UNIX *Installation Guide*.
- Tape Information— One or more character device special files to define the tape storage for this service.
- Media Changer— Character device special files to define the media changer(s) for this service (if applicable).
- Disk Information— One or more device special files, Advanced File System (AdvFS) filesets, or Logical Storage Manager (LSM) volumes to define the disk storage for this service.
- Automatic Placement Policy (ASP)— Select the policy you want ASE to use when choosing a member to run this service: Balanced Service Distribution, Favor Members, or Restrict to Favored Members. See Chapter 4 for information about the ASP policies.
- User-defined action scripts— Add any action scripts necessary to fail over the application. See Chapter 4 for information about creating action scripts.

9.3 Adding a Tape Service

Example 9–1 shows how to add a tape service. No media changer is considered in this example. The tape service will be modified in Example 9–2 to include a media changer.

Example 9–1: Adding a Tape Service

```
# asemgr
.
.
.
  Adding a service

Select the type of service:

    1)  NFS service
    2)  Disk service
    3)  User-defined service
    4)  DRD service
    5)  Tape service

    q)  Quit without adding a service
    x)  Exit                             ?)  Help

Enter your choice [1]: 5

You are now adding a new tape service to your ASE.

A tape service consists of one or more tape devices, zero or more media
changer devices, and an optional disk configuration that are failed over
together. The disk configuration can include UFS filesystems, AdvFS
filesets, LSM volumes, or raw disk information.

                     Tape Service Name

The name of a tape service must be a unique service name within this ASE.
Optionally, an IP address may be assigned to a tape service.  In this
case, the name must be a unique IP host name set up for this service and
present in the local hosts database on all ASE members.

Enter the tape service name ('q' to quit): sh-tape01

Assign an IP address to this service? (y/n): n

                     Specifying Tape Information

Enter one or more character device special files to define the tape storage
for this service.

    For example: Rewind on close, high density:      /dev/rmt0h
   No rewind on close, medium density:  /dev/nrmt1m

To end the list, press the Return key at the prompt.  To quit, enter 'q'.

Enter a tape special file name (press 'Return' to end): /dev/rmt0h

Enter a tape special file name (press 'Return' to end):  Return

                     Specifying Media Changer Information

Enter zero or more character device special files to define the media changers
for this service.

    For example: /dev/mc16

To end the list, press the Return key at the prompt.  To quit, enter 'q'.

Enter a media changer special file name (press 'Return' to end):  Return

                     Specifying Disk Information
```

Example 9–1: Adding a Tape Service (cont.)

```
Enter one or more device special files, AdvFS filesets, or LSM volumes
to define the disk storage for this service.

    For example: Device special file:      /dev/rz3c
   AdvFS fileset:            domain1#set1
   LSM volume:               /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.

Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press 'Return' to end): /dev/rrz9c

                  Mount Point

The mount point is the directory on which to mount '/dev/rz9c'.
If you do not want it mounted, enter "NONE".

Enter the mount point or NONE: NONE

                     Specifying Disk Information

Enter one or more device special files, AdvFS filesets, or LSM volumes
to define the disk storage for this service.

    For example: Device special file:      /dev/rz3c
   AdvFS fileset:            domain1#set1
   LSM volume:               /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.

Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press 'Return' to end):  Return


Modifying user-defined scripts for 'sh-tape01':

    1)  Start action
    2)  Stop action
    3)  Add action
    4)  Delete action

    x)  Exit - done with changes

Enter your choice [x]:  Return

 Selecting an Automatic Service Placement (ASP) Policy

Select the policy you want ASE to use when choosing a member
to run this service:

    b)  Balanced Service Distribution
    f)  Favor Members
    r)  Restrict to Favored Members

                                         ?)  Help

Enter your choice [b]: b
```

Example 9–1: Adding a Tape Service (cont.)

```
 Selecting an Automatic Service Placement (ASP) Policy

Do you want ASE to consider relocating this service to another member
if one becomes available while this service is running (y/n/?): n

Enter 'y' to add Service 'sh-tape01' (y/n): y
Adding service...
Starting service...
Service sh-tape01 successfully added...
#
```

In Example 9–2 the tape service `sh-tape01` will be modified to add a media changer. The tape drive, media changer, and disk will all be failed over as a unit.

Example 9–2: Modifying a Tape Service

```
# asemgr
⋮
  Service Configuration

    a)  Add a new service
    m)  Modify a service
    o)  Modify a service without interrupting its availability
    d)  Delete a service
    s)  Display the status of a service
    c)  Display the configuration of a service

    q)  Quit (back to Managing ASE Services)
    x)  Exit                             ?)  Help

Enter your choice [q]: m

 Modifying a Service

Select the service you want to modify:

    1)  tiebreaker1 on rtcr3b
    2)  sh-tape01 on rtcr4b

    q)  Quit without modifying a service
    x)  Exit                             ?)  Help

Enter your choice [q]: 2

Select what you want to modify in service 'sh-tape01':

    g)  General service information
    a)  Automatic service placement (ASP) policy

    q)  Quit without modifications
    x)  Exit                             ?)  Help

Enter your choice [g]: g

                Tape Service Modification
```

Example 9–2: Modifying a Tape Service (cont.)

```
The following menu lists the storage configuration for the tape service
‘sh-tape01‘.

You can modify the following storage configuration, add more storage, or
perform miscellaneous modifications.

Select what to modify in tape service ‘sh-tape01‘:


    1)  /dev/rmt0h (tape device)
    a)  Add a tape device, media changer, UFS file system, Advfs fileset,
        LSM volume, or raw disk
    m)  Miscellaneous modifications for ‘sh-tape01‘
    q)  Quit without making any changes
    x)  Exit (done with modifications)


Enter your choice [x]: a

                     Specifying Tape Information

Enter one or more character device special files to define the tape storage
for this service.

    For example: Rewind on close, high density:      /dev/rmt0h
   No rewind on close, medium density:  /dev/nrmt1m

To end the list, press the Return key at the prompt.  To quit, enter ’q’.

Enter a tape special file name (press ’Return’ to end):  Return

                     Specifying Media Changer Information

Enter zero or more character device special files to define the media changers
for this service.

    For example: /dev/mc16

To end the list, press the Return key at the prompt.  To quit, enter ’q’.

Enter a media changer special file name (press ’Return’ to end): /dev/mc20b

Enter a media changer special file name (press ’Return’ to end):  Return

                     Specifying Disk Information

Enter one or more device special files, AdvFS filesets, or LSM volumes
to define the disk storage for this service.

    For example: Device special file:      /dev/rz3c
   AdvFS fileset:            domain1#set1
   LSM volume:               /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.

Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press ’Return’ to end):   Return

Select what to modify in tape service ‘sh-tape01‘:
```

Example 9–2: Modifying a Tape Service (cont.)

```
    1)  /dev/rmt0h (tape device)
    2)  /dev/mc20b (media changer)
    a)  Add a tape device, media changer, UFS file system, Advfs fileset,
        LSM volume, or raw disk
    m)  Miscellaneous modifications for ‘sh-tape01‘
    q)  Quit without making any changes
    x)  Exit (done with modifications)
Enter your choice [x]: x

NOTE: Modifying a service causes it to stop and then restart.  If you do
not want to interrupt the service availability, do not modify the service.

Enter ’y’ to modify service ’sh-tape01’ (y/n): y
Stopping service...
Deleting service...
Adding service...
Starting service...
Service successfully updated.
#
```

10

Managing ASE Services

This chapter describes how to use the asemgr utility to manage the services in an available server environment (ASE). Specifically, you can perform the following tasks with the asemgr utility:

- Display service status (Section 10.1)
- Display the contents of a single ASE service or an entire ASE database (Section 10.2)
- Place a service off line (Section 10.3)
- Restart unassigned services (Section 10.4)
- Manually relocate a service to a specific member system (Section 10.5)
- Modify services (Section 10.6)
- Delete services (Section 10.7)
- Rereserve a Logical Storage Manager (LSM) device (Section 10.8)

10.1 Displaying Service Status

To display the status of available server environment (ASE) services, choose the "Display the status of a service" item from either the Managing ASE Services menu or the Obtaining ASE Status menu. Service status information includes the following:

- Type of service, either distributed raw disk (DRD), Network File System (NFS), disk, tape, or user-defined
- Service name
- Member on which the service is running or off line if the service is off line
- Automatic Service Placement (ASP) policy
- Disk configuration that the service uses

Example 10–1 shows how to display the status of the NFS service named ase4.

Example 10–1: Displaying Service Status

```
  Managing ASE Services

    c)  Service Configuration    -->
    r)  Relocate a service
   on)  Set a service on line
  off)  Set a service off line
  res)  Restart a service
    s)  Display the status of a service
    a)  Advanced Utilities       -->

    x)  Exit to the Main Menu           ?)  Help

Enter your choice [x]: s


        Service Status

Select the service whose status you want to display:

    1)  ase4 on toto
    2)  aseba2 on daffy
    3)  disk1  on gideon

    x)  Exit                            ?)  Help

Enter your choice [x]: 1

        Status for NFS service ‘ase4‘

 Status:      Relocate:    Placement Policy:    Favored Member(s):
 on toto      yes          Balance services     None

        Storage Configuration for NFS service ‘ase4‘

NFS Exports list
 /nfstest tregtest

Mount Table (device, mount point, type, options)
 treg#fset1 /var/ase/mnt/ase4/nfstest advfs rw

Advfs Configuration
 Domain:          Volume(s):
 treg             /dev/vol/dg3/vol01
LSM Configuration
 Disk Group:      Device(s):
 dg3              rz19g rz27g
```

10.2 Displaying Service Information from the ASE Database

You can obtain the full contents of the available server environment (ASE) database (/var/ase/asecdb) from disk in ASCII format by selecting the "Display the configuration of the ASE database" item from the Managing the ASE menu of the asemgr utility. Similarly, you can obtain the contents of the database for any given service by selecting the "Display the configuration of a service" item from the "Service Configuration" item of the Managing ASE Services menu.

Inspecting the contents of the ASE database can be helpful when you are staging a new ASE service, verifying the specific details of a service, or debugging unexpected service behavior. Because the format of the database output is in ASCII format, and tokens delimit the start and end of each action script, you can write scripts (such as a service monitor) that process it.

The `asemgr` command provides command-line flags that also allow you to dump ASE database contents. To display the information for a single ASE service, use the following command-line syntax:

asemgr -d -c [*service*]

To display the information for the current ASE database, use the following syntax:

asemgr -d -C

By default, the `asemgr` utility dumps the contents of the ASE database from the `/var/ase/config/asecdb` file. To display the information in a specific ASE database (for example, a saved version), use the following syntax:

asemgr -d -C /var/ase/config/asecdb.backup

The ASE database on disk (rather than the one in memory) is read by these commands. The output from these commands can be read by a shell or `awk` script and parsed into actions, depending on the data.

For example, to dump the description of the `drd1` service, enter the following command:

```
# asemgr -d -c drd1

!! ----------------------------------------------------------------------
!! ASE service configuration for drd1
!! ----------------------------------------------------------------------

@startService drd1
Service name: drd1
Service type: DRD
Relocate on boot of favored member: yes
Placement policy: balanced

DRD Device Special File:    /dev/rdrd/drd1
Underlying Storage:         /dev/rvol/dg1/vol01
LSM disk group: dg1
  dg1 disks: rz10 rz9
@endService drd1
```

The following is an example of the type of information displayed when you enter an `asemgr -d -C` command or select the "Display the configuration of the ASE database" item from the Managing the ASE menu. Callout numbers refer to the descriptions listed after the example.

```
                Managing the ASE

    a)  Add a member
    d)  Delete a member
    n)  Modify the network configuration
    m)  Display the status of the members
    C)  Display the configuration of the ASE database
    l)  Set the logging level
    e)  Edit the error alert script
    t)  Test the error alert script
     )  Enable ASE V1.5 functionality

    q)  Quit (back to the Main Menu)
    x)  Exit                             ?)  Help

Enter your choice [q]: C

!! ---------------------------------------------------------------------------
!!  ASE database /var/ase/config/asecdb
!! --------------------------------------------------------------------------- 1

ASE functionality version: V1.5 2
ASE logging level: Notice 3
Number of ASE members: 2 4
Primary network default retry settings: (see /etc/hsm.conf for overrides) 5
  Primary network max retries: 4 6
  Primary network time between retries: 10 7
  Primary network time between successes: 30 8
  Primary network time between failures: 50 9
Backup network default retry settings: (see /etc/hsm.conf for overrides) 10
  Backup network max retries: 4
  Backup network time between retries: 10
  Backup network time between successes: 300
  Backup network time between failures: 300
AM default retry settings: (see /etc/hsm.conf for overrides) 11
  AM max retries: 2
  AM time between retries: NA
  AM time between successes: 20
  AM time between failures: 300
Network interface default retry settings: (see /etc/hsm.conf for overrides) 12
  Network interface max retries: 4
  Network interface time between retries: 10
  Network interface time between successes: 30
  Network interface time between failures: 50

!! ---------------------------------------------------------------------------
!! ASE database information
!! ---------------------------------------------------------------------------

ASE database format: V1.5 13
ASE database created by product: TruCluster Production Server 14
ASE database created by member: lildogrm 15
ASE database timestamp version: 870356729 16
ASE database last updated on: Thu Jul 31 09:45:29 1997 17
ASE database last updated by: lildogrm 18
ASE database revision number: 16 19

!! ---------------------------------------------------------------------------
!! ASE network interface configuration
!! ---------------------------------------------------------------------------
```

```
Member name: lildogrm 20
  Member IP address: 10.0.0.2 21
  Daemon communication use: Primary 22
  Member aliveness ping use: Primary 23
  Monitor interface setting: ignore 24
Member name: lilcatrm 20
  Member IP address: 10.0.0.1 21
  Daemon communication use: Primary 22
  Member aliveness ping use: Primary 23
  Monitor interface setting: ignore 24

!! ---------------------------------------------------------------------------
!! ASE alert script information 25
!! ---------------------------------------------------------------------------

System alert script name: /var/ase/sbin/ase_run_sh 26
System alert script timeout: 20 27

@startText alert_script_0 28
# *****************************************************************
# *                                                               *
# *    Copyright (c) Digital Equipment Corporation, 1991, 1997    *
# *                                                               *
# *   All Rights Reserved.  Unpublished rights  reserved  under   *
# *   the copyright laws of the United States.                    *
# *                                                               *
# *   The software contained on this media  is  proprietary  to   *
# *   and  embodies  the  confidential  technology  of  Digital   *
# *   Equipment Corporation.  Possession, use,  duplication  or   *
# *   dissemination of the software and media is authorized only  *
# *   pursuant to a valid written license from Digital Equipment  *
# *   Corporation.                                                *
# *                                                               *
# *   RESTRICTED RIGHTS LEGEND   Use, duplication, or disclosure  *
# *   by the U.S. Government is subject to restrictions  as  set  *
# *   forth in Subparagraph (c)(1)(ii)  of  DFARS  252.227-7013,  *
# *   or  in  FAR 52.227-19, as applicable.                       *
# *                                                               *
# *****************************************************************
# @(#)$RCSfile: ase_logcrit_sh.sh,v $ $Revision: 1.2.5.3 $ (DEC)
# $Date: 1996/06/26 21:19:55 $

ADMIN="root"

PATH=/sbin:/usr/sbin:/usr/bin
export PATH

TIME=‘date +"%D %T"‘

ERR_FILE=/var/ase/tmp/alertMsg
HSM_STATUS=‘awk -F: ’{print $2}’ ${ERR_FILE} | sed ’s/ //g’‘

case    "${HSM_STATUS}" in
                HSM_PATH_STATUS)
                        awk -f /var/ase/lib/path_status_awk ${ERR_FILE}
                        ;;
                HSM_NI_STATUS)
                        awk -f /var/ase/lib/ni_status_awk ${ERR_FILE}
                        ;;
esac

if [ -n "${ADMIN}" ]; then
      if [ ! -f "${ERR_FILE}" ]; then
        echo "Critical ASE error or status change detected on ‘date‘" > ${ERR_FILE}
      fi

        mailx -s "***Critical ASE error or status change -
 ${TIME}" ${ADMIN} < ${ERR_FILE}
fi

rm -f ${ERR_FILE}


@endText alert_script_0 29


!! ----------------------------------------------------------------------------
!! ASE service configuration for drd_service_1
!! ----------------------------------------------------------------------------

@startService drd_service_1 30
Service name: drd_service_1 31
Service type: DRD 32
Relocate on boot of favored member: yes 33
Placement policy: favored 34
Favored member(s): milesd35
DRD Device Special File:    /dev/rdrd/drd1 36
Underlying Storage:         /dev/rrz19c 37
@endService drd_service_1 38


!! ----------------------------------------------------------------------------
!! ASE service configuration for disk_service_1
!! ----------------------------------------------------------------------------

@startService disk_service_1 30
Service name: disk_service_1 31
Service type: DISK 32
Relocate on boot of favored member: yes 33
Placement policy: favored 34
Favored member(s):  milesd35
Device: /dev/rz20c 39
  /dev/rz20c mount point: NONE/dev/rdrd/drd1
@endService disk_service_1 38


!! ----------------------------------------------------------------------------
!! ASE service configuration for aseqa236
!! ----------------------------------------------------------------------------

@startService aseqa236 30
Service name: aseqa236 31
Service type: NFS 32
Relocate on boot of favored member: yes 33
Placement policy: favored 34
Favored member(s): 35
NFS locking file: /var/ase/mnt/aseqa236/ase/aseqa236/.ase/nfs_lock 40
IP address: 16.141.112.236 41
Device: dom1#set1 39
  dom1#set1 mount point: /var/ase/mnt/aseqa236/ase/aseqa236
  dom1#set1 filesystem type: advfs
  dom1#set1 mount options: rw
  dom1#set1 mount point mode: 777
AdvFS domain: dom1
  dom1 volumes: /dev/vol/dg1/vol01
LSM disk group: dg1
  dg1 disks: rz10 rz9
```

```
!! ------------------------------------------------------------------------
!! NFS exports list for aseqa236
!! ------------------------------------------------------------------------

@startText aseqa236_NFS_exports 42
 /ase/aseqa236 43
@endText aseqa236_NFS_exports 44
@endService aseqa236
```

1 The exclamation points are comment characters.

2 Identifies the version of ASE installed.

3 Identifies the condition for which ASE logging is enabled. Values include:

- Informational
- Notice
- Warning
- Error
- Alert

4 Identifies the number of members in the ASE

5 Identifies network retry settings for the primary network, as defined in the `/etc/hsm.conf` file. See Section 1.2.3 for more information on the HSM daemon.

6 The number of additional attempts to get a response from a member system before a member is determined to be down.

7 The amount of time in seconds between attempts to get a response from a member system.

8 The time in seconds between a successful network or SCSI interconnect ping to a member and the subsequent ping.

9 The time in seconds between a failed network or SCSI interconnect ping to a member and the next attempt to ping the member.

10 Identifies network retry settings for the backup network, as defined in the `/etc/hsm.conf` file.

11 Identifies network retry settings for the Availability Manager (AM) driver, as defined in the `/etc/hsm.conf` file. See Section 1.2.4 for more information on the AM driver.

12 Identifies network retry settings for the backup network, as defined in the `/etc/hsm.conf` file.

13 Identifies the format version of the ASE database.

14 Identifies the product that was used to create the ASE database.

15 Identifies the ASE member that was used to create the database.

16 The timestamp for the version of the ASE database.

17 Identifies when the last update to the database was made.

18 Identifies the ASE member that was used last to update the database.

19 Identifies how many times the ASE database has been updated.

20 Identifies the ASE member name.

21 Identifies the ASE member's Internet Protocol (IP) address.

22 Identifies the network path used by the ASE daemons. Values for this field include:

- Never

  Network is not fully connected to all ASE members.
- No

  Network was not configured, or administrator requested that it be ignored.
- Primary

  Primary network.
- Backup

  Backup network.

23 Specifies whether ASE will use this network to send pings to members to determine whether they are alive. Values for this field include:

- Never

  Network is not fully connected to all ASE members.
- No

  Network was not configured, or administrator requested that it be ignored.
- Primary

  Primary network.
- Backup

  Backup network.

24 Specifies whether ASE will monitor the interface. Values for this field include:

- Ignore

  Do not monitor.
- Monitor

  Monitor.

25 The default error alert script which is called via alert.

26 Identifies the script the ASE agent uses to invoke the alert script.

27 Specifies the alert script's timeout value.

28 Defines the start of a script.

29 Defines the end of a script.

30 Start token for service information.

31 Name of the ASE service being defined.

32 Type of service. Values include:

- DRD
- NFS
- DISK
- USER
- TAPE

33 Specifies whether the service should relocate to a more favorable member when such a member becomes available (for instance, it boots).

34 Identifies the ASE placement policy. Values include:

- balanced
- favored
- restricted or restricted

35 Identifies the members on which the service can run (that is, those that have been specified as favored or restricted).

36 Identifies the DRD device special file corresponding to a DRD service.

37 Identifies the raw device(s) underlying a DRD service.

38 End token for service information.

39 Details on the underlying storage, disk, Advanced File System (AdvFS), and Logical Storage Manager (LSM) for Network File System (NFS) and disk services.

40 NFS locking file for NFS services.

41 IP address of disk, tape, or NFS service.

42 Start token for NFS exports list for NFS services.

43 List of exported NFS file systems.

44 End token for NFS exports list for NFS services.

10.3 Placing a Service Off Line

You can use the `asemgr` utility to temporarily stop a service by placing it off line. While a service is off line, it is unavailable to clients.

When you place a Network File System (NFS) service off line, the file systems are automatically unmounted. When you place the service on line to start it, the file systems will be mounted.

When you place off line a service that uses the Logical Storage Manager (LSM), the disk groups are deported. Deporting the disk groups only makes them inaccessible; the disk groups, the volumes, and the data in the volumes are not deleted. When you set the service on line to start it, the disk groups will be imported.

To temporarily stop a service, choose the "Set a service off line" item from the Managing ASE Services menu and choose the service you want to place off line.

If you try to set a service off line and the service cannot be stopped, the `asemgr` utility displays the following message:

```
ASE was unable to stop service ‘disk1‘.  Check the syslog’s
daemon log to determine why the stop action failed.

Two common reasons for the stop action to fail are:

(1) One of the service’s filesets is in use
    (‘umount‘ fails with Device Busy error)

(2) The user-defined stop script returns an error

You can fix the problem now and let ASE try to stop the
service, or you can ignore this failure and let ASE take the
service off line.

Enter ’r’ for ASE to RETRY the stop action or ’o’ take
the service OFFLINE [r]:
```

If you choose to retry the stop action by choosing `r` and it is successful, the service is placed off line. If you retry the stop action and it is unsuccessful, the `asemgr` utility displays the previous message again.

If you choose to continue the offline operation by choosing `o`, the `asemgr` utility sets the service off line. You must then manually stop all service processes, unmount the file systems or filesets, deport any imported LSM disk groups, and set the LSM disks off line.

If you added a user-defined stop script to the service, you must also manually perform the steps specified by this script. You can use the `-d` and `-c` options with the `asemgr` utility to dump the description of an ASE service, including the contents of any user-defined stop script (see Section 10.2).

To start a service that has been temporarily stopped, choose the "Set a service on line" item from the Managing ASE Services menu and then choose the service that you want to place on line.

The following list summarizes actions taken by the `asemgr` utility to stop a disk, Network File System (NFS), or shared tape service. If you must manually stop a service, perform these steps in sequence.

- Stop advertising the internet protocol (IP) address, as follows:

  ```
  /var/ase/sbin/nfs_ifconfig 0 stop service
  ```

  The 0 directs errors to the `/var/ase/tmp/ChildLog.0` file.

- Undo any Advanced File System (AdvFS) and UNIX File System (UFS) mounts (if applicable), as follows:

  ```
  /usr/sbin/umount /var/ase/mnt/usr/staff/dsk1
  /usr/sbin/umount /var/ase/mnt/usr/staff/dsk4/dsk5
  /usr/sbin/umount /var/ase/mnt/usr/staff/dsk4
  ```

  NFS file systems are mounted under `/var/ase/mnt/...`. Hierarchical mount points must be unmounted in reverse order.

- Stop AdvFS (if applicable) as follows:

  ```
  /sbin/rm -rf /etc/fdmns/domain
  ```

- Stop the Logical Storage Manager (LSM) (if applicable):

  ```
  /sbin/voldg deport disk_group
  /sbin/voldisk offline disk1,disk2,disk3...
  ```

To stop a distributed raw disk (DRD) service, enter the following command:

```
/var/ase/sbin/drdmgr 0 stop service
```

The 0 directs errors to the `/var/ase/tmp/ChildLog.0` file.

10.4 Restarting Unassigned Services

If the status of a service is unassigned, you can manually restart the service. To do this, choose “Restart a service” item from the Managing ASE Services menu and then choose the service you want to restart.

When you restart an unassigned service, the TruCluster software performs the following steps to completely stop the service:

1. Executes any user-defined stop script for the service.
2. Removes the Internet Protocol (IP) alias for the service.

3. Unmounts file systems, tapes, raw disks, or Advanced File Server (AdvFS) filesets used by the service.
4. Stops Logical Storage Manager (LSM) volumes and deports associated disk groups.
5. Unreserves disks.

If this sequence completes successfully, the TruCluster software attempts to start the service. If any of the steps to stop a service fails, the TruCluster software does not attempt to start the service, giving you an opportunity to fix the problem before you manually restart the service. For example, LSM volumes may be left configured on the system because the step to deconfigure LSM did not run. The `asemgr` utility will tell you that the stop has failed. Look at the `daemon.log` file to see if the disk groups were deported. (You can also use LSM tools from the Cluster Monitor to get this information.) If cleanup is required, you can do this yourself or use the `asemgr` utility to place the service off line, which performs cleanup for you.

If the TruCluster software encounters a problem when trying to start a service after completing the stop sequence, the stop scripts are run again to try to clean up the problem. This has the same effect as placing the service off line, except that the status does not show as off line when you use the `asemgr` status menu.

When a service restart fails, look in the `daemon.log` file for error messages, which may help you determine why a service was unable to start. If possible, fix the problem that prohibited the service from being started and try again to restart it. You may need to place the service off line to get to a state so that the service can be successfully restarted.

Note

If an ASE service becomes unassigned to a member system, you may not be able to use the "Restart a service" item from the Managing ASE Services menu to bring the service back on line after the error has been corrected. To make the service operational, you may need to use the "Set a service off line" and "Set a service on line" items to disable and reenable the service.

10.5 Manually Relocating a Service

You can use the `asemgr` utility to manually relocate a service. For example, you may want to relocate a service if the member system currently providing the service needs maintenance. When you relocate a service, the TruCluster software stops the service on the member currently running the service and starts the service on another member.

When you use the asemgr utility to relocate a service, you can override the service's Automatic Service Placement (ASP) policy, which may restrict the service to specific members. See Chapter 4 for more information on ASP policies.

To relocate a service, follow these steps:

1. Choose the "Relocate a service" item from the Managing ASE Services menu.
2. Choose the service you want to relocate.
3. Choose the member system that you want to run the service.

When you know the name of the service you want to relocate, you can do it from the command line. For example, to relocate the service named disk1 to member gideon, enter the following command:

```
 # asemgr -m disk1 gideon
```

Example 10–2 shows how to relocate a service using the menu interface.

Example 10–2: Relocating a Service

```
Select the service you want to relocate:

Services:

    1)  aseba1 on daffy
    2)  aseba2 on gideon
    3)  disk1 on toto

    x)  Exit to Managing ASE Services        ?)  Help

 Enter your choice [x]: 2

Select member to run ’aseba2’ service:

    1)  toto
     )  gideon
    3)  daffy

    x)  Exit without making changes      ?)  Help

 Enter your choice: 1

relocating service ‘aseba2‘ to member‘toto‘...
relocation successful...
```

10.6 Modifying a Service

The "Service Configuration" item of the Managing ASE Services menu of the asemgr utility provides two options that allow you to change service parameters in the ASE database.

You can make some changes to a service while it remains on line to its clients by selecting the "Modify a service without interrupting its availability" item, and following the instructions in Section 10.6.1. Other changes require minimal interruption of a service's availability to clients. To perform these operations, select the "Modify a service" item, and follow the instructions in Section 10.6.2.

Table 10–1 lists indicates which operations can be performed from the "Modify a service without interrupting its availability" menu item. You can perform all of the operations listed from the "Modify a service" item, but the asemgr utility will put the service off line while it accomplishes them.

Table 10–1: Service Modification Impact on Service Availability

Operation	"Modify a service without interrupting its availability"	"Modify a service"
Change the Automatic Service Placement (ASP) policy—See Chapter 4 for information about the ASP policies you can assign to a service.	Yes	Yes
Update the ASE database to reflect the addition of raw devices (/dev/rz*, or /dev/*mt* for tape services) to, or the removal of raw devices from, a Logical Storage Manager (LSM) disk group participating in a disk-based service.	Yes	Yes
Update the ASE database to reflect the addition of raw devices to, or the removal of raw devices from, an Advanced File System (AdvFS) domain participating in a disk-based service.	Yes	Yes

Table 10–1: Service Modification Impact on Service Availability (cont.)

Operation	"Modify a service without interrupting its availability"	"Modify a service"
Update the ASE database to reflect the addition of LSM volumes (`/dev/vol/*/vol*`) to, or the removal of LSM volumes from, an AdvFS domain participating in a disk-based service.	Yes	Yes
Modify the exports file for a Network File System (NFS) service.	Yes	Yes
Add raw devices to, or remove raw devices from, a distributed raw disk (DRD) service.	Yes	Yes
Add LSM volumes (`/dev/vol/*/vol*`) to, or remove LSM volumes from, a DRD service.	Yes	Yes
Add UNIX file systems or AdvFS filesets to a disk-based service.	No	Yes
Change the name of a file system, fileset, or volume participating in a service.	No	Yes
Change the mount point of a service, including its mount options, owner, and mode.	No	Yes
Change the disk access mode (either read/write or read-only) or quotas.	No	Yes
Change a service's name.	No	Yes
Replace, edit, modify, or delete a user-defined action script.	No	Yes
Change the timeout value for a user-defined action script.	No	Yes

DIGITAL recommends using either of these options whenever you change a service's configuration information in the ASE database. Both allow the service to remain on line in its ASE after the modifications are performed expeditiously under ASE control.

You can also modify a service while it is off line to the ASE (for example, if you have set it off line as described in Section 10.3) and outside of the utility's control. However, this method requires more manual intervention, is most disruptive to service availability, and is prone to error, especially in complex storage configurations.

Notes

Carefully read Section 10.6.3 before modifying AdvFS or LSM configuration information for a service that you have set off line. This section also describes restrictions to the use of AdvFS, LSM, and DIGITAL UNIX disk management commands in an ASE.

If you modify an NFS service in an ASE, your first attempt to access the service may cause a `stale file handle` message to be displayed

After you modify the exports file of an NFS service, clients may receive `stale file handle` messages and be unable to mount NFS file systems associated with the service. If this problem occurs, send the `SIGHUP` signal to the `mountd` process on the ASE member that is running the service.

10.6.1 Modifying a Service Without Interrupting Its Availability

For those changes to a service that do not require a disruption in service availability (as described in Table 10–1), follow these steps:

1. Run the `asemgr` utility on the member system on which the service is currently running.
2. Choose the "Modify a service without interrupting its availability" item from the Service Configuration menu and then choose the service that you want to modify:

```
 Service Configuration

   a)  Add a new service
   m)  Modify a service
   o)  Modify a service without interrupting its availability
   d)  Delete a service
   s)  Display the status of a service

   x)  Exit to Managing ASE Services    ?)  Help

Enter your choice [x]: o
```

```
 Online Service Modification

Select the service you want to modify:

   1)  ase1 on fgreg1
   2)  greg on fgreg2
   3)  drd1 on fgreg2

   x)  Exit to Service Configuration

Enter your choice [x]: 2
```

The asemgr utility displays a list of modifications that can be performed without disrupting the selected online service.

3. For example, choose option a to modify the Automatic Service Placement (ASP) policy:

```
Select what you want to modify in service ‘greg‘:

    a)  Automatic service placement (ASP) policy
    u)  Update with storage configuration changes (LSM and AdvFS only)

    x)  Exit without modifications

Enter your choice [x]: a
```

   At this point, proceed to specify a new ASP policy in the same manner as when adding a new service. For more information on specifying an ASP policy, see Chapter 4.

Example 10–3 shows how to update disk configuration information for an online AdvFS service. In this example, the administrator added a new volume to an AdvFS domain and runs the asemgr utility to modify the service (ase1) that uses this domain. The asemgr utility checks the AdvFS configuration information and updates the ASE database while service ase1 continues to run without interruption.

Example 10–3: Modifying an AdvFS Service with No Disruption to Its Availability

```
  Service Configuration

    a)  Add a new service
    m)  Modify a service
    o)  Modify a service without interrupting its availability
    d)  Delete a service
    s)  Display the status of a service

    x)  Exit to Managing ASE Services    ?)  Help

Enter your choice [x]: o

 Online Service Modification
```

Example 10–3: Modifying an AdvFS Service with No Disruption to Its Availability (cont.)

```
Select the service you want to modify:

    1)  ase1 on fgreg1
    2)  greg on fgreg2
    3)  drd1 on fgreg2

    x)  Exit to Service Configuration

Enter your choice [x]: 1

Select what you want to modify in service ‘ase1‘:

    a)  Automatic service placement (ASP) policy
    u)  Update with storage configuration changes (LSM and AdvFS only)

    x)  Exit without modifications

Enter your choice [x]: u

Checking AdvFS domain val ...

AdvFS domain ‘val‘ shows a change in the volumes configured:

Old volume list:
 /dev/rz20b

New volume list:
 /dev/rz16b
 /dev/rz20b

Is this correct (y/n) [y]: y
Enter ’y’ to modify service ’ase1’ (y/n): y
Service successfully updated.
```

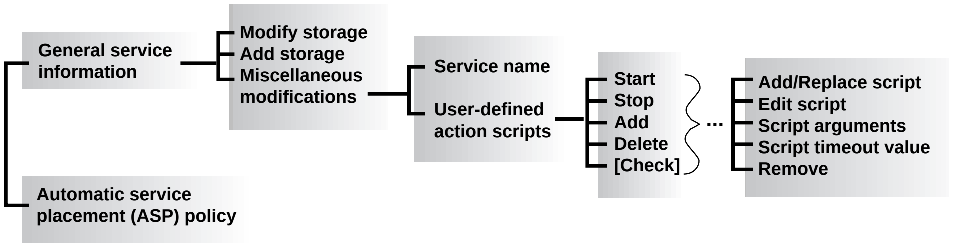
If the service cannot be started with the new storage configuration information, the service is placed off line. You are then given the opportunity to correct the problem or to discard the configuration modifications.

10.6.2 Modifying a Service with Minimal Disruption to Its Availability

For those changes to a service that require a slight disruption in service availability (as described in Table 10–1), select the "Modify a service" item from the Service Configuration menu of the `asemgr` utility.

Figure 10–1 shows a map of the `asemgr` utility’s Modifying a Service menu.

Figure 10–1: Map for Modifying a Service Menu



ZK−1154U−AI

When you select the "Modify a service" item from the Service Configuration menu of the `asemgr` utility, the TruCluster software automatically performs the following tasks:

1. Stops the service on the member system.
2. Deletes the service from all the members.
3. Adds the service to the ASE database.
4. Starts the modified service on a member.
5. Propagates the database changes to all the members.

If you try to modify a service and the service cannot be stopped, the `asemgr` utility displays the following message:

```
ASE was unable to stop service ‘disk‘.  Check the syslog’s
daemon log to determine why the stop action failed.

Two common reasons for the stop action to fail are:

(1) one of the service’s filesets is in use
    (‘umount‘ fails with Device Busy error), or

(2) the user-defined stop script returns an error.

You can fix the problem now and let ASE try again to stop the
service or you can let ASE take the service OFFLINE so you can
stop the service manually.  If you choose to stop the service
manually and cannot stop it completely, you will need to reboot
the system to avoid a system panic or data corruption.

If the stop action failed because one of the service’s filesets
is in use, you should try to kill the processes that are using
the fileset and let ASE try again to stop the service.  If ASE
is unable to stop the service, you should let ASE take the
service offline.  If the stop action failed because the
user-defined stop script returned an error, you should let ASE
```

```
continue with the modify operation.


Enter 'r' for ASE to RETRY the stop action or 'm' for you to
MANUALLY stop the service [r]:
```

If you retry the stop action and it is successful, the service is modified. If you retry the stop action and it is unsuccessful, the `asemgr` utility displays the previous message again.

If you continue the stop operation without fixing the problem, the `asemgr` utility displays the following message:

```
You must manually stop all service processes, unmount any mounted
filesets, deport any imported LSM disk groups, and set the LSM
disks offline.

Failure to stop the service could cause the system to panic when
the service is restarted.  If you cannot stop the service, you
should reboot the member running the service.

 Press 'Return' to continue:  Return

You can now exit the modify operation or continue with
the modify.  If you continue with the modify, the old service
will be deleted and then the new service will be added.  Once the
old service is deleted, if the ASE cannot add the new service or
restore the old service, the service will remain deleted.

Enter 'x' to exit the modify operation or 'c' to continue [x]:
```

The `asemgr` utility allows you to either continue to modify the service or to exit the operation. If you choose to exit the operation, the service is placed off line, and the `asemgr` utility returns to the Service Configuration menu. If you continue with the modification, the original service is deleted and the new service is added and then started, if possible.

If a modified service cannot be added to the ASE, the `asemgr` utility displays the following message:

```
Add failed - Unable to add service.
Check syslog's daemon log to determine the error.

Enter 'o' to restore the old service configuration, 'n' to retry
the new service configuration, or 'd' to delete the service [n]:
```

For example, the mount point specified in an NFS service may not have been created on the system. You can create the mount point, then type **n** to retry.

If you retry the service modification and it succeeds, the modification proceeds. It the retry fails, the `asemgr` utility displays the previous message.

If you successfully restore the original service, it remains configured but is placed off line. If the original service cannot be restored, the `asemgr` utility displays the previous message.

If you cannot successfully restore the original service or modify the service, you can delete the service. See Section 10.7 for information about deleting services.

Example 10–4 shows how to modify an NFS service by adding a new area to be exported, and by editing the ASE exports file to restrict the service to two clients.

Example 10–4: Modifying an NFS Service

```
 # asemgr
⋮

 Modifying a Service

Select the service you want to modify:


    1)  ase3
    2)  aseba2

    x)  Exit                             ?)  Help


 Enter your choice [x]: 1

Select what you want to modify in service ‘ase3‘:

    g)  General service information
    r)  Automatic service placement (ASP) policy

    x)  Exit without modifications           ?)  Help

 Enter your choice [g]: g

                NFS Service Modification

The following menu lists the disk storage configuration for
the service "ase3."

You can modify the following storage configurations, add more storage,
or perform miscellaneous modifications (for example, modify the
exports file).

Select what to modify in the NFS service ‘ase3‘:

    1)  /dev/vol/dg3/vol03      (UFS)
    2)  dom1#fset1     (ADVFS)
    a)  Add a UFS file system, AdvFS fileset, or LSM volume
    m)  Miscellaneous modifications for ‘ase3‘
```

Example 10–4: Modifying an NFS Service (cont.)

```
    q)  Quit without making any changes
    x)  Exit (done with modifications)

 Enter your choice [x]: a

                   Specifying Disk Information

Enter one or more UFS device special files, AdvFS filesets, or LSM
volumes to define the disk storage for this service.

    For example:        Device special file:      /dev/rz3c
                        AdvFS fileset:            domain1#set1
                        LSM volume:               /dev/vol/dg1/vol01

To end the list, press the Return key at the prompt.

Enter a device special file, an AdvFS fileset, or an LSM volume as
 storage for this service (press 'Return' to end): /dev/rz28c

Enter the directory pathname(s) to be NFS exported from the storage
area "/dev/rz28c".  Press 'Return' when done.

 Enter a directory pathname: /usr/staff

  Enter a host name, NIS netgroup, or IP address for the NFS
   exports list (press 'Return' for all hosts): staff_group

Enter a directory pathname: Return

              UFS File System Read-Write Access
Mount `/dev/rz28c` file system with read-write or read-only access?

    1)  Read-write
    2)  Read-only

 Enter your choice [1]: 2

              UFS Mount Options Modification

Enter a comma-separated list of any mount options you want to use for
`/dev/rz28c` (in addition to the UFS-specific defaults listed
in the mount.8 reference page).  If none are given, only the default
mount options are used.

 Enter options (Return for none): Return

Enter a device special file, an AdvFS fileset, or an LSM volume as
 storage for this service (press 'Return' to end):  Return

Select what to modify in the NFS service `ase3`:

    1)  /dev/vol/dg3/vol03      (UFS)
    2)  dom1#fset1     (ADVFS)
    a)  Add a UFS file system, AdvFS fileset, or LSM volume
    m)  Miscellaneous modifications for `ase3`
    q)  Quit without making any changes
    x)  Exit (done with modifications)

 Enter your choice [x]: m
```

Example 10–4: Modifying an NFS Service (cont.)

```
Miscellaneous modifications for service ‘ase3‘:

    e)  Export file input for ‘ase3‘
    n)  Service name [ase3]
    s)  NFS file system area [dom1#fset1]
    u)  User-defined action scripts
    x)  Exit to previous menu


 Enter your choice [x]: x

Select what to modify in the NFS service ‘ase3‘:

    1)  /dev/vol/dg3/vol03      (UFS)
    2)  dom1#fset1     (ADVFS)
    a)  Add a UFS file system, AdvFS fileset, or LSM volume
    m)  Miscellaneous modifications for ‘ase3‘
    q)  Quit without making any changes
    x)  Exit (done with modifications)

 Enter your choice [x]: x

 Enter ’y’ to modify Service ’ase3’ (y/n): y

Stopping service...
Deleting service...
Adding service...
Starting service...
Saving the updated database...
Service successfully updated...
```

10.6.3 Managing Storage Participating in ASE Services

Managing storage in an ASE requires a detailed understanding of the TruCluster software and storage subsystem operation.

This section provides information about managing AdvFS and LSM subsystems, replacing a failed disk, and using DIGITAL UNIX storage management commands.

10.6.3.1 Introduction to ASE Storage Configuration Management

Before you set up a service, you must have configured the disks that you will use in the service. For example, you must have set up your file systems, AdvFS filesets, and LSM volumes. When you run the `asemgr` utility to add a service, you specify information about the disk configuration at the prompts. This information is included in the ASE database on each member system.

However, after a disk is used in an ASE service, you must use special procedures to manage the disk. This is because the ASE must always maintain control of the disk while a service is running. For example, you do not manually unmount a file system that is being used by an online ASE service.

As discussed in Section 10.6, the TruCluster software allows you to modify an ASE service and change its storage configuration. The "Service Configuration" item of the Managing ASE Services menu of the `asemgr` utility provides two options that allow you to change service parameters while the service remains on line within its ASE and under ASE control. You can modify any of the information that was specified when the service was added to the ASE. Although most common changes (such as adding a disk to an AdvFS domain or an LSM volume) can be made to a service without disrupting its availability to clients, others (such as changing disk attributes) can be made only with a slight disruption to its availability. Table 10–1 indicates which modifications require a disruption of service availability.

DIGITAL recommends using either of these options whenever you change a service's configuration information in the ASE database. Both allow the service to remain on line in its ASE while the modifications are performed expeditiously under ASE control. You can also modify a service while it is off line to the ASE and outside of the utility's control. However, this method requires more manual intervention, is most disruptive to service availability, and is prone to error, especially in complex storage configurations. Carefully read Section 10.6.3 before modifying AdvFS or LSM configuration information for a service that you have set off line. This section also describes restrictions on the use of AdvFS, LSM, and DIGITAL UNIX disk management commands in an ASE.

If you make modifications that affect a service's AdvFS or LSM storage configuration and you want to modify the service while it is off line to the ASE, you must set the service off line as described in Section 10.3 and follow the special procedures described in the following sections. These procedures ensure that the ASE recognizes and incorporates the configuration changes, and that the ASE database is consistent with any AdvFS and LSM subsystem databases.

10.6.3.2 Understanding Storage Configuration Databases

The AdvFS and the LSM subsystems utilize databases that contain records of the subsystems' disk configuration. AdvFS and LSM subsystems consist of a disk organization built on top of physical disks. For example:

- In an AdvFS subsystem, a file domain contains one or more physical disks (volumes) and is used as a shared storage pool for one or more filesets.

- In an LSM subsystem, a disk group is a collection of physical disks and is used as a shared storage pool for LSM volumes. You can create file systems or AdvFS domains and filesets on top of LSM volumes.

Before you add an ASE service, you must set up the storage configuration for the service. When you create the service, the `asemgr` utility prompts you for information about the storage configuration, including physical disks, UNIX file systems, AdvFS filesets and domains, LSM volumes, and mount points. This information is included in the ASE database file on each member system and is used to configure, start, and stop the ASE service.

For proper ASE operation, the information in the ASE database must be consistent with any AdvFS or LSM configuration database. If you modify a service's AdvFS or LSM configuration and those modifications cause the subsystem database to change, you must be sure to incorporate those changes into the ASE database.

10.6.3.3 Modifying an Offline AdvFS or LSM Service

If you modify a service that has been placed off line, the service will be unavailable during the entire service modification procedure.

To modify an AdvFS or LSM storage configuration used in a service, you must follow special procedures:

- If you are modifying a service on line, make the modifications on the member system running the service and make sure that the service does not relocate. (See Section 10.6 for instructions.)
- If you are modifying a service off line, make the modifications on the member system on which the disks are configured or imported.

Regardless of whether you are modifying an online or an offline service, some AdvFS and LSM storage modifications require you to update the ASE database with the storage configuration changes. You also must adhere to the AdvFS and LSM command restrictions listed in Section 10.6.3.4 and Section 10.6.3.5.

To modify the storage configuration for a service that has been placed off line, follow these steps:

1. Optionally, run the `asemgr` utility to display the status of the service you want to modify and the storage configuration for the service.
2. Run the `asemgr` utility and set the service off line.

3. If you are modifying an LSM storage configuration or an AdvFS configuration on top of an LSM volume, perform the following tasks on the member system on which you will modify the LSM configuration:

    a. Place on line the disks used in the disk group you want to modify, using the following command syntax:

    ```
    voldisk online disk . . .
    ```

    b. Import the disk group, using the following command syntax:

    ```
    voldg import disk_group
    ```

    c. Restart the volumes, using the following command syntax:

    ```
    volrecover -sb disk_group
    ```

    If you are modifying an AdvFS storage configuration, perform the following tasks:

    a. On the member system on which you will modify the AdvFS configuration, re-create the domain directory, using the following command syntax:

    ```
    mkdir -p /etc/fdmns/ domain_name
    ```

    b. Change your directory to the domain directory, using the following command syntax:

    ```
    cd /etc/fdmns/ domain_name
    ```

    c. Re-create the device links, using the following command syntax for all the volumes in the domain:

    ```
    ln –s device
    ```

    The *device* variable can be a UNIX file system or an LSM volume (for example, `/dev/rz20c` or `/dev/vol/dg3/vol04`).

4. Modify the AdvFS or LSM storage configuration. See Section 10.6.3.4, Section 10.6.3.5, and Section 10.6.3.6 for restrictions on using AdvFS, LSM, and DIGITAL UNIX disk management commands in an ASE.

5. If you modified the LSM storage configuration, deport the disk group and take its disks off line. Use the following command syntax:

    ```
    voldg deport disk_group
    ```

    ```
    voldisk offline disk . . .
    ```

    If you modified the AdvFS storage configuration, remove the domain directory that you created, using the following command syntax:

    ```
    rm -rf /etc/fdmns/domain_name
    ```

6. If necessary, update the ASE database by running the `asemgr` utility. Select the "Modify a service" or "Modify a service online" item, as appropriate, from the Service Configuration menu, specify the name of the service, and then select the "General service information" menu item.

   Choose the menu item that corresponds to the storage modification you made. For example:

   ```
     Select what to modify in the NFS service ‘ase3‘:

       1)  /dev/vol/dg3/vol03      (UFS)
       2)  dom1#fset1     (ADVFS)
       a)  Add a UFS file system, AdvFS fileset, or LSM volume
       m)  Miscellaneous modifications for ‘ase3‘
       q)  Quit without making any changes
       x)  Exit (done with modifications)
   ```

   In the previous example:

   - Choose `1` if you changed the LSM configuration (for example, if you added disks to disk group `dg3`).
   - Choose `2` if you changed the AdvFS configuration (for example, if you added a volume to domain `dom1`).
   - Choose `a` if you changed the storage configuration and you also want to add a new file system, AdvFS fileset, or LSM volume to the service.

   After you choose a menu item, another menu is displayed, depending on your previous choice. Select the menu item that corresponds to the storage modifications you made.

   If you chose the `/dev/vol/dg3/vol03` item from the previous menu, the following menu might be displayed:

   ```
   Modify/Delete file system ‘/dev/vol/dg3/vol03‘:

       m)  Modify UFS file system information
       n)  Change file system device special file
              [/dev/vol/dg3/vol03]
       e)  Modify the exports list
        )  Modify AdvFS domain information
       d)  Delete ‘/dev/vol/dg3/vol03‘
       x)  Exit - done with changes
   ```

   After you choose the appropriate menu item, information about the modified configuration is displayed, and you are prompted to confirm that the information is correct.

   If you chose item `m` in the previous example, the following information might be displayed:

```
LSM physical list does not match with what LSM thinks

        New list: rz10g rz11g
        Old list: rz10g

 Following is a list of devices and pubpath for disk group dg3

        DEVICE  PUBPATH

        rz10g   /dev/rz10g
        rz11g   /dev/rz11g

Is this correct (y/n) [y]:
```

When you enter `y`, the TruCluster software updates the ASE database with the new storage configuration on all the member systems.

7. Run the `asemgr` utility and place the service on line. If you are unable to restart the service because of an error, the service remains off line, giving you the opportunity to correct the problem.

10.6.3.4 Restrictions on Using AdvFS Commands in an ASE

The following list contains AdvFS commands that have restrictions when used in an ASE or that require you to update the ASE database:

- `addvol`

  To add a volume to a domain, use the `addvol` command. Then, run the `asemgr` utility and use the `o` option on the Service Configuration menu to update the ASE database.

- `balance`

  Before you can balance volumes in a file domain, all filesets in the file domain must be mounted. If there are filesets in the domain that the ASE service does not mount, you must use the `asemgr` utility to modify the service so that the filesets are mounted. Unless the service is off line, do not manually mount the filesets.

- `defragment`

  Before you can defragment a file domain, all filesets in the file domain must be mounted. If there are filesets in the domain that the ASE service does not mount, you must use the `asemgr` utility to modify the service so that the filesets are mounted. Unless the service is off line, you cannot manually mount the filesets.

- `dxadvfs`

  There are no restrictions on running the `dxadvfs` command to display information. However, if you are using the graphical maintenance utilities, you must adhere to the restrictions for the equivalent command-line operations.

- `mkfdmn`

  If one partition of a volume is assigned to an ASE service, the other partitions must be assigned to the same service or remain unused. This is because only one ASE service can use a disk.

- `mkfset`

  If one fileset in a domain is assigned to an ASE service, the other filesets must be assigned to the same service or remain unused. This is because only one ASE service can use a disk. Do not include any of the filesets in a domain that is used in an ASE service in the `/etc/fstab` file. Loopback NFS-mounting of the filesets exported by an ASE service is supported.

- `renamefset`

  To rename a fileset that is used in an ASE service, use the `renamefset` command. Then, run the `asemgr` utility to place the service off line, and update the ASE database by using the `m` option on the Service Configuration menu.

- `rmfdmn`

  To remove a domain belonging to an ASE service, use the `asemgr` utility to modify the service and remove all of the domain's filesets. If there are dependencies between the filesets' mount points, remove the filesets in the order that they would unmounted. Then, invoke the `rmfdmn` command.

- `rmfset`

  To remove a fileset belonging to an ASE service, use the `asemgr` utility to modify the service and delete the fileset from the service. Then, use the `rmfset` command to remove the fileset from the domain.

- `rmvol`

  If the domain contains filesets that the ASE service does not mount, you must modify the service to mount them, or place the service off line and then manually mount the filesets. Do not manually mount the filesets while the service is on line.

  To remove a volume from a domain, use the `rmvol` command. Then, run the `asemgr` utility and use the `o` option on the Service Configuration menu to update the ASE database.

- `vquotacheck`

  The TruCluster software runs the `vquotacheck` command when it starts a service. If there is activity on a fileset used in the service, the `vquotacheck` command may fail and the `vquotaon` command will not run. If the `vquotacheck` command fails, you can manually invoke the command.

- vncheck

  To run the vncheck command on filesets used in an ASE service, you must put the service off line. Save a copy of the /etc/fstab file and add the filesets to the file. Then, invoke the vncheck command. After you run the command, you can restore the original /etc/fstab file and place the service on line.

- vquotaqon

  The TruCluster software runs the vquotaon command when it starts a service if the vquotacheck command exits with zero status. If necessary, you can manually invoke the vquotaon command after the service is started.

The following list contains the AdvFS commands that have no restrictions when used in an ASE, and do not require you to update the ASE database:

- chfile
- chfsets
- chvol
- clonefset
- migrate
- mktrashcan
- rmtrashcan
- showfdmn
- showfsets
- shtrashcan
- stripe
- vdump
- vedquota
- vquot
- vquota
- vquotaoff
- vrepquota
- vrestore

10.6.3.5 Restrictions on Using LSM Commands in an ASE

The following list contains the LSM commands that have restrictions when used in an ASE, or that require you to update the ASE database:

- `dxlsm`

  There are no restrictions on running the `dxlsm` command to display information. However, if you are using the graphical maintenance utilities, you must adhere to the restrictions that apply to the equivalent command-line operations.

- `voladvdomencap`

  See the restrictions for the `volencap` command.

- `volassist`

  If you use the `make` keyword, all volumes in a disk group must belong to one service and cannot be used outside of the ASE. There are no restrictions for the `mirror`, `move`, `grow`, `shrink`, or `snap` keywords.

- `vold`

  The TruCluster software requires that the LSM configuration daemon, `vold`, is running and stabilized. In addition, you cannot fail over or relocate an ASE service that uses LSM if a `vold` is either initializing or not running.

- `voldctl`

  Do not use the `hostid` keyword on an ASE member system, unless you stop the TruCluster software on that system. You can do one of the following tasks:

  - Delete the member system from the ASE, run the `voldctl` command with the `hostid` keyword, and then add the system to the ASE.
  - Shut down the system to single-user mode, run the `voldctl` command with the `hostid` keyword, and then boot to multiuser mode.
  - Invoke the `/sbin/init.d/asemember stop` script, run the `voldctl` command with the `hostid` keyword, and then invoke the `/sbin/init.d asemember start` script.

  If you use the `add disk` keyword, do not add a disk that is used by another ASE service.

  Do not use the `disable` keyword to disable the `vold` daemon on a member system if any service in the ASE uses LSM.

  Do not use the `stop` option to stop the `vold` daemon on a member system if any service in the ASE uses LSM.

  There are no restrictions for the `init`, `rm disk`, `list`, `enable`, and `mode` keywords.

- `voldg`

  If you use the `import` keyword, do not import a disk group that belongs to an online ASE service.

  If you use the `deport` keyword, do not deport a disk group that belongs to an online ASE service.

  If you use the `flush` keyword, do not flush a disk group belonging to an online ASE service.

  Before you add disks to a disk group, you must use the `voldisksetup` command to initialize the disks. Use the `voldg adddisk` command (or the `voldiskadd` command) to add the disks. Then, run the `asemgr` utility and choose the `o` option on the Service Configuration menu to update the ASE database. To remove disks from a disk group, use the `voldg rmdisk` command, then run the `asemgr` utility and choose the `o` option on the Service Configuration menu to update the ASE database.

  There are no restrictions for the `init`, `list`, and `free` keywords.

- `voldisk`

  If you use the `init` keyword, do not initialize a disk if any of its partitions are used in an ASE service.

  Do not use the `rm` keyword to remove a disk that is used in an online ASE service.

  Do not use the `clearimport` keyword on a disk that is used in an online ASE service.

  There are no restrictions for the `define`, `moddb`, `offline`, `online`, or `list` keywords.

- `voldiskadd`

  Use the `voldiskadd` command to add the disks. Then, run the `asemgr` utility and choose the `o` option on the Service Configuration menu to update the ASE database.

- `voldiskadm`

  This utility provides an interactive interface for other LSM utilities. See the restrictions for the `voldg` and `voldisk` commands.

- `voledit`

  To remove a volume record from a disk group used in an ASE service, run the `asemgr` utility and delete the volume from the service, and then use the `voledit rm` command to remove the record.

  To rename a volume assigned to an ASE service, run the `voledit` command to rename the volume, and then run the `asemgr` utility to modify the service, choosing the "Change the device special file" menu item.

There are no restrictions for the `set` and `cc` keywords.

- `volencap`

  To encapsulate a disk partition and add it to a service's LSM configuration, use the `asemgr` utility to delete the affected storage from the ASE service. Invoke the `volencap` or the `vol-reconfig` command to encapsulate the storage. If necessary, reboot the system. Then, modify the ASE service to add the LSM volume or volumes to the service.

- `volmend`

  Do not use the `volmend` command on storage used in an online ASE service.

- `volplex`

  If you want to use the `dis` keyword, see the restrictions for the `volassist make` command and the `mount` command.

  There are no restrictions for the `att`, `det`, `cp`, `snap`, and `mv` keywords.

- `vol-reconfig`

  See the restrictions for the `volencap` command.

- `volsetup`

  Do not include any ASE shared storage devices in the `rootdg` disk group.

- `volstartup`

  See the restrictions for the `vold` command.

- `volume`

  If you use the `init` keyword, do not initialize a volume that is used in an online ASE service.

  If you use the `stop` or `stopall` keywords, do not stop a volume that is used in an online ASE service.

  Do not use the `maint` keyword on a volume that is used in an online ASE service.

  There are no restrictions for the `rdpol`, `start`, `startall`, `resync`, and `set` keywords.

The following list contains the LSM commands that have no restrictions when used in an ASE, and do not require you to update the ASE database:

- `volinfo`
- `voliod`
- `volinstall`
- `volnotify`

- `volprint`
- `volrecover`
- `volsd`
- `volstat`
- `voltrace`
- `volwatch`

10.6.3.6 Restrictions on DIGITAL UNIX Storage Management Commands

The following commands have restrictions when used in an ASE:

- Disk quota management commands and utilities

  The disk quota utilities require entries in the `/etc/fstab` file. However, you must not edit the `/etc/fstab` in an ASE. Instead, use the `asemgr` utility to enable quota enforcement on file systems and filesets used in an ASE service.

- `mount`

  Do not manually mount a file system or fileset used in an ASE service unless the service is off line.

- `scu`

  There are no restrictions on the `show device`, `show inquiry`, or `show path-inquiry` commands. However, DIGITAL does not support any other `scu` commands with a bus or device that is used in an online ASE service, unless you are authorized to run the commands by the Customer Support Center (CSC).

- `umount`

  Do not manually unmount a file system or fileset that is used in an ASE service unless the service is off line.

10.6.3.7 Replacing a Failed Disk

The following sections describe how to replace a failed disk. See your hardware configuration and software installation manuals for information about installing disks in storage units.

10.6.3.7.1 Replacing a Nonmirrored Disk

If a disk that is not part of an LSM mirrored volume fails, the service that is using the disk will stop. To replace a disk, follow these steps:

1. Pull the failed disk from its slot and replace it with a disk that has the same unit number as the failed disk.

2. If the disk is part of an LSM disk group, use the `asemgr` utility to rereserve the service's disks.
3. Restore the data from a backup.
4. Start the service.

If you are using LSM, you can replace a failed disk with a disk that has a different unit number. To do this, you must have a spare disk that is part of the same LSM disk group as the failed disk. Having the spare disk available ensures that the list of physical disks in the disk group does not change, and you do not have to update the ASE database. However, you must keep the failed disk in its disk slot until you are ready to replace it with a viable disk. This replacement disk can now be used as the spare disk in the disk group.

10.6.3.7.2 Replacing a Disk that is Part of an LSM Mirrored Volume

You can replace a failed disk that is part of an LSM mirrored volume without interrupting the availability of the service. To do this, follow these steps:

1. Follow the procedure in the DIGITAL UNIX *Logical Storage Manager* manual for replacing a disk with the same unit number. However, while the failed disk is removed from the disk group, the service cannot be failed over.
2. Run the `asemgr` utility and rereserve the service's disks.
3. Use the `volrecover` command to recover the disk data.

You can replace a failed disk with a disk that has a different unit number without interrupting the availability of the service. To do this, you must have a spare disk that is part of the same disk group as the failed disk. Having the spare disk available ensures that the list of physical disks in the disk group does not change, and you do not have to update the ASE database. However, you must keep the failed disk in its disk slot until you are ready to replace it with a viable disk. This replacement disk can now be used as the spare disk in the disk group.

10.7 Deleting a Service

You use the `asemgr` utility to delete a service. When you delete a service, the TruCluster software stops the service on the member system, deletes the service from all the members, removes the service information from the database, and propagates the database changes to all the members.

If you delete or stop a service that uses Logical Storage Manager (LSM), the disk group is deported. Deporting the disk group only makes it inaccessible; the disk group, the volumes, and the data in the volumes are not deleted. In addition, if you delete or stop a service that uses Advanced File System (AdvFS), the domain is no longer configured. Because of this, when you delete a service that uses either LSM or AdvFS, the `asemgr` utility prompts you for a member on which to leave the LSM disk group imported or AdvFS domain configured, so you can use it for other purposes.

To delete a service, choose the "Delete a service" item from the Service Configuration menu and then choose the service you want to delete. If the service uses AdvFS or LSM, you are prompted for a member on which to keep the disks configured.

Example 10–5 shows how to delete a service.

Example 10–5: Deleting a Service

```
 # asemgr

⋮
                Service Configuration

    a)  Add a service
    m)  Modify a service
    d)  Delete a service
    s)  Display the status of a service

    x)  Back to Managing ASE Services menu   ?)  Help

 Enter your choice [x]: d

Select the service you want to delete:

    1)  aseba1 on daffy
    2)  aseba2 on gideon
    3)  disk1 on toto

    x)  Exit to Service Configuration     ?)  Help

 Enter your choice [x]: 2

This service uses either an AdvFS or an LSM storage configuration.
You must select a member on which to leave the storage configured:

    1)  toto
    2)  gideon
    3)  daffy

    x)  Exit to Service Configuration      ?)  Help

 Enter your choice [1]: 2

Member to leave the storage configuration on: gideon

 Is this correct (y/n) [y]: y

 Enter ’y’ to delete Service ’aseba2’ (y/n): y

Stopping service...
```

Example 10–5: Deleting a Service (cont.)

```
Deleting service...
Saving the updated database...
Service successfully removed...
```

If you try to delete a service and the service cannot be stopped, the `asemgr` utility displays the following message:

```
ASE was unable to stop service ‘service‘.  Check the syslog’s
daemon log to determine why the stop action failed.

Two common reasons for the stop action to fail are:

(1) One of the service’s filesets is in use
    (‘umount‘ fails with Device Busy error)

(2) The user-defined stop script returns an error

You can fix the problem now and let ASE try to stop the service,
or you can ignore this failure and let ASE continue with the
delete operation.

Enter ’r’ for ASE to RETRY the stop action or ’c’ to CONTINUE
with the delete operation [r]:
```

If you retry the stop action and it is successful, the service is deleted. If you retry the stop action and it is unsuccessful, the `asemgr` utility displays the previous message again.

If you continue the delete operation, the `asemgr` utility displays the following message:

```
You must manually stop all service processes, unmount any
mounted filesets, deport any imported LSM disk groups, and
set the LSM disks off line.

Failure to stop the service could cause the system to panic
when the service is restarted.  If you cannot stop the service,
you should reboot the member running the service.

Press ’Return’ to continue:
```

When you press the Return key, the `asemgr` utility deletes the service. You must then manually stop all service processes, unmount any mounted file systems or filesets, deport any imported LSM disk groups, and set the LSM disks off line. You can then use the disks in another service or for some other purpose.

10.8 Rereserving an LSM Device

When a failed or previously unavailable part of a Logical Storage Manager (LSM) mirrored volume becomes available again, you can reincorporate the device into the service without interrupting the service. To do this, resynchronize the mirrored volume outside of the available server environment (ASE) on the member to which the disk groups are imported. Then, rereserve the devices by using the `asemgr` utility's Advanced Utilities menu.

11

Using the Cluster Monitor

The Cluster Monitor provides a graphical view of the Production Server cluster or Available Server configuration based on event reports from the connection manager and available server environment (ASE) daemons. Use the Cluster Monitor to track service availability and connectivity among member systems. You can also use it to manage services and to start storage management applications.

The Cluster Monitor:

- Displays the status of each member of an Available Server configuration or Production Server cluster.
- Reports errors on the primary network interface, member system failures, ASE service failures, and hard and soft disk errors.
- Displays the configuration of the Available Server configuration or Production Server cluster, including all ASEs, member systems and their services, storage devices, and network interfaces.
- Displays the devices on a member system's **private SCSI buses**.
- Displays the shared storage reserved by a service.
- Starts, stops, restarts, and relocates services.
- Launches the `asemgr` utility as an external tool.
- Launches `dxadvfs`, `dxterm`, `dxlsm`, `cnxshow` (in a Production Server configuration), and Performance Manager as external tools. (Note that some of these tools require the installation of specific product licenses and subsets. See the TruCluster Software Products *Software Installation* manual for additional information.)

The Cluster Monitor does not display the following:

- Devices associated with a shared tape service.
- Problems that occur with network interconnects other than the one defined as the primary network interconnect. (In a Production Server configuration, this is always the MEMORY CHANNEL interconnect.)

The Cluster Monitor displays an updated snapshot of the state of a Production Server cluster or Available Server configuration each time an event or change occurs.

Note

The Cluster Monitor groups all Production Server cluster members that are not in an ASE into a single pseudodomain named 9999. You can display the device for this domain, but not an ASE view.

11.1 Setting Up the Cluster Monitor

To set up the Cluster Monitor, follow these steps:

1. Make sure all systems and devices are properly connected.
2. Make sure the base operating system subsets upon which the Cluster Monitor depends are installed (see the TruCluster Software Products *Software Installation* manual).
3. Make sure the Cluster Monitor subset (`TCRCMS150`) is installed on every member.
4. On each member system, set up the `/.rhosts` file to allow root access for the `rsh` command between any two members. Be sure to use the systems' member names in the `/.rhosts` file.

   Note

   To maintain a secure network, modify the `ifaccess.conf` file as explained in the `cluster_map_create`(8) reference page.

5. Make sure the `/etc/hosts` file on each system in the cluster lists the Internet Protocol (IP) name and IP address and the MC (MEMORY CHANNEL) IP name and IP address of every member system and the `cluster_cnx` IP name and "10.0.0.42" IP address.

   Note

   The IP name and IP address must be manually inserted in each cluster member's `/etc/hosts` file. Each member's own MC IP name and address and that of the `cluster_cnx` are entered automagically during the install. The other member's MC IP names and addresses must also be entered manually in each member's `/etc/hosts` file.

6. If you have not already done so, configure each available server environment (ASE) by running the asemgr utility on one member in each ASE and entering the complete member list for that ASE. (See Chapter 2 for instructions.)
7. Check that all members are up by running the asemgr utility in each ASE and displaying member status. (See Chapter 2 for more information.)
8. Create the cluster configuration map on one cluster member (see Section 11.1.1).

   After the cluster configuration map is successfully created, the value of the CMS_CONF variable in the /etc/rc.config file is set to on and the tractd and submon daemons are automatically started on all cluster members. (These daemons also automatically start on subsequent reboots.)
9. To enable remote ASE functions between ASEs in a cluster, you must make each cluster member an authorized host; you can use the xhosts command or add all cluster members to the X access list using the Host Manager utility. Be sure to use the MEMORY CHANNEL IP name for each cluster member.
10. Invoke the Cluster Monitor from the command line to verify the cluster installation and configuration (see Section 11.1.2). The output of the monitor can be directed to any display device that is compatible with the X Windows protocol.

11.1.1 Creating the Cluster Configuration Map

You must create a cluster configuration map:

- After the TruCluster software is installed on all member systems and before using the Cluster Monitor for the first time
- Whenever you add a new member system or delete an existing member system
- Whenever the hardware configuration changes

The cluster configuration map contains a record of the entire hardware configuration, including systems, interconnects, and devices. A copy of the cluster configuration map resides on each member. The cluster configuration map file, /etc/CCM, must be identical and up to date on each member system so that the Cluster Monitor can properly display configuration information.

To update or re-create the cluster configuration map, select a single cluster member on which to run the cluster_map_create utility. You can use the following flags:

- Use the –full flag to force each cluster member to reconstruct its current cluster configuration map by invoking the SCSI CAM utility (scu) to derive its SCSI bus and device configurations. When the cluster_map_create utility is used with the –full flag, it includes in the map only those members and devices it can access. If it cannot access a component, it omits it from the map. Use the –full flag only when you are certain that all member systems are up and operational. Use the cnxshow utility to determine the status of cluster members.
- Use the –append flag to force each cluster member to append new hardware components to its current cluster configuration map. Use this flag to add hardware components to an existing cluster configuration map when a full rebuild of the map would cause hardware that is temporarily down and unavailable to be removed from the map.

If a cluster configuration map file (/etc/CCM) does not already exist in the cluster, the cluster_map_create utility creates it, distributes a copy to each cluster member, and starts the submon process and trigger-action daemon (tractd) on each cluster member.

If a cluster configuration map already exists, the utility does nothing. However, when the –full flag is specified, the cluster_map_create utility always rebuilds and redistributes the cluster configuration map. (Note that the cluster_map_create utility only starts the submon process and tractd daemon if they are not currently running.)

You must perform the following tasks before entering the cluster_map_create command:

- You must configure all cluster members, ASEs, and shared storage in the cluster.
- You must add the names of all members' cluster interconnect interfaces to each member's /.rhosts file. This enables the cluster_map_create utility root access to all cluster members from any member. (To protect your system against distributed security attacks, remove these names from the .rhosts files after the cluster_map_create utility completes. However, if you do so, you will not be able to run external tools, such as the asemgr utility, on individual members by dragging and dropping a tool icon on a member icon.)

When the cluster_map_create utility completes successfully, it displays a series of messages. The utility displays one or more dots (.) to indicate that work is in progress. The length of time required for the cluster_map_create utility to complete depends on the number of shared devices in the cluster. In the following example, the cluster_map_create utility completed successfully on a two-system cluster:

```
# /usr/sbin/cluster_map_create cluster1 -full
Members running are ( clu13, clu14 )
Doing device table scans
...
Doing symmetry checks
...
Processing map input file
Calling makeclmap to create /etc/CCM
Distributing cluster map to all members
Processing member clu13
Processing member clu14
Successful cluster map creation and distribution.
```

The `cluster_map_create` utility collects, checks, and merges the configuration information into the cluster configuration map file, `/etc/CCM`, which is distributed to the kernels of all member systems. If configuration errors (such as missing devices or asymmetry) are discovered or if any cluster members go down (with a return status of DOWN) during this process, the `cluster_map_create` utility displays error messages in the terminal window from which you invoked it. Appendix A lists the error messages generated by the `cluster_map_create` utility.

If the cluster configuration map is not created, see Section 11.2.1 for troubleshooting information.

For more information on this utility, see `cluster_map_create`(8). The syntax of the `/etc/CCM` file is described in `CCM`(5).

11.1.2 Starting the Cluster Monitor

Before starting the Cluster Monitor from a remote client display, make sure you have set up security to allow the member that will be running the Cluster Monitor to access the client display.

Note

> You must have root privilege to configure clients and servers using this application.

To start the Cluster Monitor from the Common Desktop Environment (CDE), follow these steps:

1. Click on the Application Manager icon on the CDE front panel.
2. Double click on the System Admin application group icon.
3. Double click on the TruCluster Tools application group icon.
4. Double click on the Cluster Monitor application icon.

To start the Cluster Monitor from the command line, follow these steps:

1. Log in to the member system as root.
2. If you logged in to the member system from a remote workstation, set the DISPLAY variable to reference the remote workstation.
3. Run the cmon program by entering the following command:

   ```
   # nohup /usr/bin/X11/cmon &
   ```

   The nohup utility allows you to log out of the member system without exiting the Cluster Monitor. After you log out, the Cluster Monitor continues to be displayed on the remote workstation. To use the high availability feature of the Cluster Monitor, you must set up a shared Network File System (NFS) area, as described in Section 11.3, and invoke the cmon command with the –ha flag.

See cmon(8) for information about invoking the Cluster Monitor from the command line, including its command-line options. See X(1X) for details on the X toolkit command-line options that you can use with the Cluster Monitor. For information about using the Cluster Monitor, see the online help.

11.2 Troubleshooting the Cluster Monitor

This section provides troubleshooting information for the following situations:

- Cluster configuration map not created (Section 11.2.1)
- Cluster monitor does not start (Section 11.2.2)
- Cluster monitor does not show a member (Section 11.2.3)
- Errors reported by the cluster_map_create utility (Section 11.2.4)
- Errors reported by the cmon utility (Section 11.2.5)
- Other problems (Section 11.2.6)

11.2.1 Cluster Configuration Map Not Created

If the cluster configuration map cannot be created, the following are likely reasons:

- An incorrect cluster_map_create command was issued.
- The cluster_map_create utility does not recognize a cluster member or is experiencing difficulty when communicating with a cluster member.
- There is an inconsistent SCSI bus configuration within the cluster.

If the cluster configuration map is not created when you run the `cluster_map_create` utility, check the following:

- Verify that all shared SCSI buses and devices on the shared buses are symmetrically configured on all members to which they are connected (see the TruCluster Software Products *Software Installation* manual).
- Use the `ping` command to determine whether all ASE members are reachable from one another over the MEMORY CHANNEL interconnect.
- Ensure that each member system's `/.rhosts` file contains an entry for each member system (using the MEMORY CHANNEL IP name).
- Run the `cnxshow` utility and ensure that each member system is recognized and running in the cluster. If a member system does not display in the `cnxshow` utility:

    1. Ensure that you have correctly configured the MEMORY CHANNEL interconnect. See the TruCluster Software Products *Hardware Configuration* manual and the MEMORY CHANNEL *User's Guide* for configuration information.
    2. Enter the `sysconfig -q rm` command and view the MEMORY CHANNEL kernel attributes (see Appendix B).

- Ensure that all member systems are recognized and running by using the `asemgr` utility. If a member system does not display in the `asemgr` utility, then add it.

See Appendix A for specific errors generated by the `cluster_map_create` utility.

11.2.2 Cluster Monitor Does Not Start

If you installed the Cluster Monitor subset, and the cluster configuration map has been successfully created, use the `ps` command with the `ag` options to verify that the `tractd` and `submon` daemons are running on each cluster member.

If the daemons are not running, check that the value of the `CMS_CONF` variable in the `/etc/rc.config` file is set to `on`. Then, manually start the daemons on all systems in the cluster. Start the `tractd` daemon first, then start the `submon` daemon. The `tractd` daemon must perform some extra initialization tasks when started for the first time in a cluster. To allow this initialization to be completed, wait a few seconds before starting the `submon` daemon on the first system.

If the `tractd` and `submon` daemons are running, inspect the `daemon.log` file and analyze messages from the `cmon` utility to determine the cause of the problem.

If the system warns you that it is unable to open the display, use the `setenv DISPLAY` command to set the display to your workstation before entering the `cmon` command.

11.2.3 Cluster Monitor Does Not Show a Member

If the Cluster Monitor does not show a cluster member system in its display, run the `cnxshow` utility on the system that is not shown as a member, and compare the output with that produced by running the `cnxshow` utility on another cluster member. If there is an inconsistency in the output, reboot the members that are inconsistent.

If the output from the `cnxshow` utility is consistent on all cluster members, re-create the cluster configuration map (using the `-full` option) and restart the Cluster Monitor. For information on re-creating the cluster configuration map, see Section 11.1.1.

11.2.4 Errors Reported by the cluster_map_create Utility

The following errors are reported by `cluster_map_create` utility:

No previous cluster map.

You entered the `cluster_map_create` command with the `-append` option, but no cluster configuration map file (`/etc/CCM`) exists to which to append. Reenter the `cluster_map_create` command with the `-full` option (without the `-append` option) to create a cluster configuration map file.

No members found!

The `cluster_map_create` utility has detected that there are no other member systems in the cluster. Member systems are either incorrectly configured for the Production Server or the cluster communications subsystem is not operating properly. To correct this problem, follow these steps:

1. Run the `cnxshow` utility and ensure that each member system is recognized and running in the cluster. If a member system is not displayed by the `cnxshow` utility:

    a. Ensure that you have correctly configured the MEMORY CHANNEL interconnect. See the TruCluster Software Products *Hardware Configuration* manual and the MEMORY CHANNEL *User's Guide* for configuration information.

    b. Enter the `sysconfig -q rm` command and view the MEMORY CHANNEL kernel attributes (see Appendix B).

2. Run the asemgr utility and ensure that the member system is recognized and running. If a member system is not recognized by the asemgr utility, add it.
3. Reenter the cluster_map_create command.

Cluster or ASE member *hostname* is either unreachable or improperly configured.

The cluster_map_create utility could not contact a member system. The member system is either incorrectly configured for the Production Server or the cluster communication subsystem is not operating properly. To correct this problem, follow these steps:

1. Ensure that each member system's /.rhosts file contains an entry for each member system (using the MEMORY CHANNEL Internet Protocol (IP) name).
2. Run the cnxshow utility and ensure that each member system is recognized and running in the cluster. If a member system is not displayed by the cnxshow utility:

   a. Ensure that you have correctly configured the MEMORY CHANNEL interconnect. See the TruCluster Software Products *Hardware Configuration* manual and the MEMORY CHANNEL *User's Guide* for configuration information.

   b. Enter the sysconfig -q rm command and view the MEMORY CHANNEL kernel attributes (see Appendix B).
3. Run the asemgr utility and ensure that the member system is recognized and running. If a member system is not recognized by the asemgr utility, add it.
4. Reenter the cluster_map_create command.

Error: Asymmetric shared SCSIs in ASE *ASE_ID* between members *hostname* and *hostname*

The cluster_map_create utility has found an inconsistency in the number of SCSI buses between member systems. To correct this problem, follow these steps:

1. Run the scu utility and ensure that the same number of buses are configured between member systems (see the TruCluster Software Products *Software Installation* manual).
2. Reenter the cluster_map_create command.

Badly formatted *filename* file.

You entered the cluster_map_create command with the -append option, but the cluster_map_create command cannot append to the original cluster configuration map file (/etc/CCM). Reenter the cluster_map_create command without the -append option.

Error: failed to make the cluster map.

The cluster_map_create utility could not read the contents of an input file it created to generate the cluster configuration map. To correct this problem, follow these steps:

1. Ensure that each member system's /.rhosts file contains an entry for each member system (using the MEMORY CHANNEL IP name).
2. Run the scu utility to ensure that the number of SCSI bus and shared SCI device identifiers are consistent between member systems (see the TruCluster Software Products *Software Installation* manual).
3. Reenter the cluster_map_create command.

Error: No saved map input file *filename* to append to.

You entered the cluster_map_create command with the -append option, but no cluster configuration map file exists to which to append. Reenter the cluster_map_create command without the -append command to create the cluster configuration map file (/etc/CCM).

Member *hostname* is unreachable. Failure to distribute new cluster map.

A cluster configuration map file (/etc/CCM) has been created for the cluster; however, the cluster_map_create utility cannot distribute it to the member system identified in the message. The member system is either incorrectly configured for Production Server or the cluster communication subsystem is not operating properly. To correct this problem, follow these steps:

1. Ensure that each member system's /.rhosts file contains an entry for each member system (using the MEMORY CHANNEL IP name).
2. Run the cnxshow utility and ensure that each member system is recognized and running in the cluster. If a member system is not displayed by the cnxshow utility:

    a. Ensure that you have correctly configured the MEMORY CHANNEL interconnect in a Production Server cluster. See the TruCluster Software Products *Hardware Configuration* manual and the MEMORY CHANNEL *User's Guide* for configuration information.

   b. Enter the `sysconfig -q rm` command and view the MEMORY CHANNEL kernel attributes (see Appendix B).

3. Run the `asemgr` utility and ensure that the member system is recognized and running. If a member system is not recognized by the `asemgr` utility, add it.
4. Reenter the `cluster_map_create` command.

Failure to load the new cluster map on member *hostname*.

The member system identified in the message did not receive a copy of the cluster configuration map. The member system is either incorrectly configured, or the cluster communication subsystem is not operating properly. To correct this problem, follow these steps:

1. Ensure that each member system's `/.rhosts` file contains an entry for each member system.
2. In a Production Server cluster, run the `cnxshow` utility and ensure that each member system is recognized and running in the cluster. If a member system is not displayed by the `cnxshow` utility:

   a. Ensure that you have correctly configured the MEMORY CHANNEL interconnect. See the TruCluster Software Products *Hardware Configuration* manual and the MEMORY CHANNEL *User's Guide* for configuration information.

   b. Enter the `sysconfig -q rm` command and view the MEMORY CHANNEL kernel attributes (see Appendix B).

3. Run the `asemgr` utility and ensure that the member system is recognized and running. If a member system is not recognized by the `asemgr` utility, add it.
4. Reenter the `cluster_map_create` command.

Failure to load new map onto all cluster members.

The `cluster_map_create` utility could not start the `tractd` daemon and `submon` process on cluster members; therefore, it could not distribute the cluster configuration map. Member systems are either incorrectly configured for Production Server or the cluster communication subsystem is not operating properly. To correct this problem, follow these steps:

1. Ensure that each member system's `/.rhosts` file contains an entry for each member system (using the MEMORY CHANNEL IP name).

2. In a Production Server cluster, run the `cnxshow` utility and ensure that each member system is recognized and running in the cluster. If a member system is not displayed by the `cnxshow` utility:

   a. In a Production Server cluster, ensure that you have correctly configured the MEMORY CHANNEL interconnect. See the TruCluster Software Products *Hardware Configuration* manual and the MEMORY CHANNEL *User's Guide* for configuration information.

   b. Enter the `sysconfig -q rm` command and view the MEMORY CHANNEL kernel attributes (see Appendix B).

3. Run the `asemgr` utility and ensure that the member system is recognized and running. If a member system is not recognized by the `asemgr` utility, add it.

4. Reenter the `cluster_map_create` command.

11.2.5 Error Reported by the cmon Utility

Man Page could not be formatted. The requested Man Page is either not present, or corrupt.

One of the base operating system reference pages subsets required by the Cluster Monitor online help is not installed. Check the dependencies listed in the TruCluster Software Products *Software Installation* manual.

11.2.6 Other Problems

If the X color map is fully allocated or becomes fully allocated when you invoke the Cluster Monitor, the Cluster Monitor may use other colors in place of those it cannot allocate. Typically this happens when you are using other color-intensive or graphic-intensive applications at the same time as you are using the Cluster Monitor. Many of these applications have mechanisms that allow you to run them with a reduced color map. This may allow you to run the Cluster Monitor with its full color palette.

If you are in the middle of an available server environment (ASE) operation, such as a service relocation, the Cluster Monitor may temporarily display an erroneous message. If there is no problem in the ASE, the message will not appear in the display for the next reporting cycle (approximately 20 seconds).

11.3 Setting Up a Highly Available Cluster Monitor Service

You can set up a Network File System (NFS) service to make the Cluster Monitor graphical user interface highly available. Invoke the Cluster Monitor by using the `cmon –ha` command, which makes the Cluster Monitor highly available. You must also run the Cluster Monitor on a member system with its display set to a workstation or another member system.

If a member system running a highly available Cluster Monitor fails, the Cluster Monitor is restarted on the member system currently running the NFS service. The Cluster Monitor will display on the same workstation as before the failure.

If the member system running the NFS service fails, the Cluster Monitor will also restart, but there will be a short delay while the service relocates to another member. After the service relocates, it restarts the Cluster Monitor on the local system.

After the service relocates, there is a short delay while the NFS locking mechanism restarts. The Cluster Monitor acquires a lock on the run list file, adds an entry so that it is registered in case of another failure, and restarts its display. To minimize the effect of the NFS locking delay, you can set up the Automatic Service Placement (ASP) policy for the NFS service so that the service runs only on members that rarely fail.

To set up a highly available Cluster Monitor, follow these steps:

1. Set up the Cluster Monitor as described in the previous sections.
2. Determine the name of the NFS service, for example `cmonha`, and assign an Internet Protocol (IP) address to it, specifying the service name and address in the `/etc/hosts` file on all member systems. See Chapter 4 for detailed information about preparing disks for a service.
3. Use the `asemgr` utility to set up the NFS service. When you set up the NFS service, you must specify:

   - The name of the service, which is also an IP host name
   - The shared disk specification for the `/var/cmon` directory
   - The `/var/cmon` directory mount point
   - Root write permission for the `/var/cmon` mount point
   - The service's ASP (`b` option for Balance Services)

   After you add the service, it is started on a member system.
4. Using the `asemgr` utility, modify the ASE exports file for the Cluster Monitor service so that the phrase `-root=0` is at the end of the exports line. This preserves root ID mapping for all clients of the service.

5. Because the member systems will act as both clients and servers of the NFS service, you must run nfssetup on each member system and NFS-mount /var/cmon from the NFS service using its IP host name.

6. On one member system, create a /var/cmon/run directory. NFS exports this directory on the service's shared disk to each member system.

7. Use the asemgr utility to modify the NFS service and set up user-defined start and stop action scripts. See Chapter 4 for information about action scripts. See Chapter 10 for information about modifying services. Add the following commands at the bottom of the user-defined start action script, before the exit command:

   ```
   if [ ! -f /usr/sbin/cmonsvc ]

   then

           exit 2

   fi

   /usr/sbin/cmonsvc
   ```

   The cmonsvc daemon uses the host status monitor (HSM) daemon to monitor the member systems in the available server environment (ASE). If a member fails, the cmonsvc daemon checks the /var/cmon/run/runList file to determine if the member had been running the cmon –ha command. If so, the Cluster Monitor is restarted on the member that is running the NFS service. Add the following commands at the bottom of the user-defined stop action script, before the exit command:

   ```
   if [ ! -f /usr/sbin/sendTrig ]

   then

           exit 2

   fi

   /usr/sbin/sendTrig -e cmonsvcStop
   ```

   The sendTrig program sends a cmonsvcStop trigger message to the cmonsvc daemon on the member system running the service. This stops the service.

8. Invoke the Cluster Monitor from a remote system as described in Section 11.1.2, but use the following cmon command:

   ```
   # nohup /usr/bin/X11/cmon -ha -d hostname:0.0 &
   ```

The -d option is not needed if the DISPLAY environment variable is set on the member system. When the Cluster Monitor is invoked with the previous command line, it writes a line to the /var/cmon/run/runList file. The line contains the name of the member system, the cmon command's process ID, and the name of the system that is displaying the Cluster Monitor. That line is then used to restart the Cluster Monitor, if necessary.

12

Troubleshooting

12.1 Using ASE Event Logging

The available server environment (ASE) logger daemon (aselogger) tracks the ASE messages generated by all the member systems. A logger daemon can be run on one or more member systems. If you have more than one member system running a logger daemon, you will have virtually duplicate logs on the member systems. Messages appear in the log files in the order that they were logged, not necessarily in the order that they occurred.

During software installation, the TruCluster software installation procedure prompted you to determine if you want to run the ASE logger daemon each time the system is booted. If you chose not to run the logger daemon when you installed the TruCluster software, you can invoke the following command and then reboot the system to start the logger daemon each time the system is booted:

```
# rcmgr set ASELOGGER 1
```

Note

A temporary stop in a network or a high network load may cause the aselogger daemon to overflow its message queue, resulting in the loss of some log messages on the system running the daemon. To avoid losing messages, run the aselogger daemon on each member system.

The ASE logger daemon logs messages generated by the asemgr utility, the director daemon, the agent daemon, and the logger daemon. Messages generated by the host status monitor (HSM) daemon and the availability manager (AM) driver are logged to the local system. In addition, if the ASE logger daemon stops, all daemon messages are logged only to the system on which they occurred. Note that when the TruCluster software first starts, the initial messages that are generated may be logged only to the local system.

The logger daemon uses the DIGITAL UNIX event logging facility, syslog, to collect messages that are logged by the various kernel, command, utility, and application processes. Messages are either logged to a local file or forwarded to a remote system, as specified in the /etc/syslog.conf file on each member system running the logger daemon.

The /etc/syslog.conf event logging configuration file specifies how a member system logs messages. If you use the default logging configuration, all asemgr utility and ASE daemon messages are logged to the /var/adm/syslog.dated/*date* /daemon.log file. The AM driver messages are logged to the /kern.log file in the same directory.

In addition, you can set the severity level of ASE error logging by using the asemgr utility. This allows you to limit the ASE messages that are logged. See Section 12.1.3 for more information.

To examine the ASE messages generated by the asemgr utility and the logger, director, and agent daemons, check the event logging files of any member system that is running a logger daemon. To examine the ASE messages generated by the HSM daemon and the AM driver on a particular member system, check that system's event logging files. Appendix A contains a partial list of important event messages and their descriptions.

The following example shows a remote message and a local message on member system gideontc, which is running a logger daemon:

```
Jan 27 11:22:27  gideontc  ASE: pigeon Agent Error: HSM reported state

         change of zen.tst.com, a non-member host


Jan 27 11:22:29  gideontc  ASE: local AseLogger Notice: connected to Agent

    1           2      3      4      5      6       7
```

The ASE messages are logged in a specific format and include the following information:

1 Date and timestamp.

2 Local system name.

3 Identifier (not used in messages from the AM driver).

4 System that generated the message—Note that local is specified if the message was logged locally or was not logged using the logger daemon. This information is not specified in messages from the AM driver.

5 Source of the message—The following components can generate an ASE message:

AseMgr	The `asemgr` utility
Director	The ASE director daemon
Agent	The ASE agent daemon
HSM	The HSM daemon
AseLogger	The logger daemon
AM	The AM driver
vmunix	The kernel
AseUtility	A process or daemon unrelated to ASE

6 Severity of the message—The severity level is not included in messages from the AM driver. Messages can have the following severity levels:

info	A low-level informational message
notice	A high-level informational message about significant activity in the ASE
warning	A message about activity in the ASE that may indicate an error condition
error	A message about an error that was detected
alert	A message about a critical condition that requires immediate attention

7 Message text.

The ASE action scripts capture any output from the commands that they execute. If the action script fails, the command output is logged as errors and the source of the message is specified in the log files as `AseUtility`.

12.1.1 Configure Mail for the ASE Logger Daemon

By default, the ASE logger daemon logs alert messages in the `daemon.log` file in the `/var/adm/syslog.dated/`*`date`* directory and notification is sent to root on the local system. You can use the `mailsetup` program to configure mail so that the superuser can receive error alert messages from the ASE logger daemon. (You can use the Mail option on the `setup` utility menu to run this program.) See `mailsetup`(8) for more information. For information on setting up mail to fail over, see Section 5.5.

12.1.2 Displaying the Members Running a Logger Daemon

To determine which member systems are running a logger daemon, choose the "Obtaining ASE Status" item from the ASE Main Menu and then choose the "Display the location(s) of the logger" item.

12.1.3 Setting and Displaying the Message Logging Severity Level

ASE message logging uses the DIGITAL UNIX `syslog` function and `syslogd` daemon. However, you can use the `asemgr` utility to specify the severity level of the messages that you want the ASE logger daemon to log, which restricts the severity level of the messages that are logged.

There are five possible severity levels that can be logged, as described in Section 12.1.1. The following table describes the types of messages associated with the possible severity levels:

Message Type	Description
Informational	Logs messages of all severity levels. This is the default.
Notice, warning, and error logging	Logs messages with the `notice`, `warning`, `error`, and `alert` severity levels.
Warning and error logging	Logs messages with the `warning`, `error`, and `alert` severity levels.
Error logging only	Logs messages with the `error` and `alert` severity levels.

To set the severity level for message logging, choose the "Set the logging level" item from the Managing the ASE menu. Example 12–1 shows how to set the severity level for message logging.

Example 12–1: Setting the Logging Severity Level

```
Enter the logging level for the ASE:


    i)  Informational (log everything)

    n)  Notice, warning, and error logging

    w)  Warning and error logging

    e)  Error logging only


    x)  Exit to Managing the ASE
Enter your choice [i]: n
```

You can display the severity level of the messages being logged by choosing the "Display the level of logging" item from the Obtaining ASE Status menu.

12.1.4 Disabling ASE Event Logging

To disable ASE event logging on a member system, you must stop ASE services, which stops the enabled logger daemon, reset the `ASELOGGER` parameter to zero, and restart ASE services for the change to take effect.

Enter the following commands to disable ASE envent logging on a member system:

```
# /sbin/init.d/asemember stop

# rcmgr set ASELOGGER 0

# /sbin/init.d/asemember start
```

12.1.5 Editing and Testing the Error Alert Script

TruCluster software provides you with a script that executes a specified task when an error with the alert severity level occurs. You use the `asemgr` utility to edit and test the script.

The default error alert script sends mail to users that you specify in the script. You can edit the script to specify which users will receive the mail, and you can specify some other action to take when a severe error occurs.

To edit the error alert script, choose the "Edit the error alert script" item from the Managing the ASE menu. The `asemgr` utility invokes the `vi` editor or the editor defined by the `EDITOR` environment variable, and you can specify the users to which you want mail sent or you can make other changes to the script.

Example 12–2 shows how to edit the error alert script.

Example 12–2: Editing the Error Alert Script

```
#  Define ADMIN on next line to get mail for critical ASE errors


ADMIN=root


PATH=/sbin:/usr/sbin:/usr/bin

export PATH



ERR_FILE=/var/ase/tmp/alertMsg

TIME=‘date +"%D %T"‘

HSM_STATUS=‘awk -F: ’{print $2}’ ${ERR_FILE} | sed ’s/ //g’‘
```

Example 12–2: Editing the Error Alert Script (cont.)

```
case    "${HSM_STATUS}" in
                HSM_NI_STATUS)

                        awk -f /var/ase/lib/ni_status_awk ${ERR_FILE}
                        ;;
                HSM_PATH_STATUS)
                        awk -f /var/ase/lib/path_status_awk ${ERR_FILE}
                        ;;
esac

if [ -n "${ADMIN}" ]; then
        if [ ! -f "${ERR_FILE}" ]; then
                echo "Critical ASE error detected on ‘date‘" >
         ${ERR_FILE}
        fi

        mailx -s "***Critical ASE error - ${TIME}" ${ADMIN} < ${ERR_FILE}
fi

rm -f ${ERR_FILE}
:wq
```

To test the error alert script, choose the "Test the error alert script" item from the Managing the ASE menu. ASE sends a test alert message to the Logger daemon and invokes the error alert script.

12.2 Resetting the ASE Daemons

You can reset the available server environment (ASE) daemons on a member system if problems occur in the ASE. Resetting the ASE daemons stops the ASE director, logger, and host status monitor (HSM) daemons and initializes the ASE agent daemons on a system. The agent daemons then restart all the daemons to make the ASE fully operational. If resetting the ASE daemons does not fix the problem, you can initialize or reboot the member system.

To reset the ASE daemons on a member system, use the following command:

```
/sbin/init.d/asemember restart
```

12.3 Controlling the Priority of the ASE Daemons

You must ensure that the available server environment (ASE) daemons do not time out, because other system processes have a higher scheduling priority. The ASE daemons must have a scheduling priority that is higher than normal system processes; they must be able to respond to administrative commands and other events in the ASE. The daemons' high priority enables the ASE to operate even when the member systems are busy. See the DIGITAL UNIX *System Administration* manual for information about scheduling processes.

If there are processes other than those generated by the ASE with a scheduling priority that is higher than the priority of the ASE daemons, the daemons could time out while waiting to run. If this occurs, messages such as the following are written to the log file, indicating that operations are timing out:

```
Mar 8 13:09:28 surry ASE: surry AseMgr Error: ASE timeout -

                Unable to stop service.
```

The ASE agent daemon (aseagent) and logger daemon (aselogger) are started in the /sbin/init.d/asemember script with a "nice" value of -5, which raises the priority of the daemons. The processes that descend from the ASE daemons inherit the raised scheduling priority. For example, the director daemon (asedirector) and any programs or scripts started by the ASE daemons have the same raised priority as the agent and logger daemons.

You can raise the scheduling priority of the ASE daemons by changing the "nice" value specified in the lines in the /sbin/init.d/asemember file that start the aseagent and aselogger daemons. See nice(1) for more information about scheduling priorities.

Note that ASE daemons started with a "nice" priority will not always stay at that priority. Over time, if the member systems do not reboot, the daemons' priority may return to the average run priority. When the member systems reboot, the daemons' priority is raised again according to the "nice" value in the /sbin/init.d/asemember script.

Therefore, the default /sbin/init.d/asemember script contains the following command, which supersedes the “nice” value for the asehsm daemon and runs the daemon with a fixed high priority that does not degrade over time:

```
aseagent −p hsm
```

If you do not want the fixed high priority for the asehsm daemon, remove this command from the /sbin/init.d/asemember script.

You can also raise and fix the priority of the aseagent, asedirector, and asehsm daemons by including the following command in the /sbin/init.d/asemember script:

```
aseagent −p all
```

12.4 Connection Manager Removed System from the Cluster (PS)

The connection manager’s monitor daemon, cnxmond is started on all cluster members by the /sbin/init.d/clumember script. The cnxmond daemon has the following two options that, when multiplied, specify the longest duration that communications can be inoperative between a system and the connection manager monitor daemon:

- The -p option specifies a ping interval; that is, the interval during which at least one ping must be received from the cnxpingd daemon on a member system. (Normally, the cnxpingd daemon sends two pings during this interval.)
- The -D option is a multiplier that determines a timeout interval, which is based on the ping interval.

For example:

```
cnxmond -p 10 -D 6
```

This command results in a 60–second timeout interval.

When started by the clumember script, the cnxmond daemon searches the /etc/rc.config file to determine the values for the -p and -D options. The value for the -p option is obtained from the CNX_INTERVAL variable; the value for the -D option is obtained from the CNX_WAVES variable.

When the cnxmond daemon detects an interruption in communications with a system (that is, no ping is received during the timeout interval), the connection manager removes the system from the cluster. Investigate the source of the communications problem and, if necessary, use the rcmgr set command to increase the value of the CNX_WAVES variable. For example, to change the value of CNX_WAVES to 10, enter the following command:

```
# rcmgr set CNX_WAVES 10
```

A

Error Messages

This appendix contains a partial list of important messages generated by the TruCluster software. These messages have an Alert severity level and are included in the `daemon.log` file unless otherwise noted. A message with an Alert severity level indicates that a critical condition exists and needs the immediate attention of a system manager.

Log file entries specify the following information:

- Time of event
- Name of the local system
- Component identifier
- Member on which the event was generated
- Daemon that generated the event
- Event severity level
- Message text

If the daemon that generated an event is disconnected from the available server environment (ASE) logger daemon, and the message arrived after the disconnect, the ASE logger daemon may not be able to identify the daemon that sent the message. In this case, the source of the event is specified as “unknown client.” For example:

```
Aug 31 11:34:35 staff1 DECsafe: unknown client Info: ASE_INQ_SERVICES

Reply from Director seq: 12 ch: 3  ASE_OK
```

Messages that specify `AseUtility` as the daemon that generated the message were produced by a command or daemon unrelated to the TruCluster software. For example, the following commands were produced by the Logical Storage Manager (LSM) software:

```
AseUtility Error: voldisk: Volume daemon is not accessible


AseUtility Error: voldisk define of rz19 failed


AseUtility Error: voldisk: Device rz19: define failed: Device path invalid
```

The ASE action scripts capture output from the commands that they execute. This output is sent to the logger daemon. If the action script fails, the command output is logged as errors. See the appropriate software documentation for information on errors not related to the TruCluster software.

The following sections describe some of the Alert messages generated by the TruCluster software.

A.1 ASE Agent Daemon Alert Messages

This section describes some Alert messages generated by the available server environment (ASE) agent daemon.

```
Can’t stop service <service> for failed device. rebooting!
```

A device has failed and the agent cannot stop the service associated with the failed device. If a stop fails, `umount` may have failed, because a file is open locally on the NFS file system. If the service is relocated to another member and later relocated back to the original member and the member’s cache for the file system did not get flushed because of the failed `umount`, the cache could get flushed when the service restarts on the original member and could cause file corruption. To prevent this, the ASE agent daemon reboots the local node.

```
Member <member> cut off from net
```

The member is disconnected from the network.

```
Member <member> is not available
```

The member that was running the director is not answering pings over the network or over the SCSI bus; therefore, it is considered unavailable.

```
device access failure on <device> from <host>
```

The specified device cannot be reached from the specified host.

```
AM can’t access <device> on <host> on reservation reset
```

The reservation for the specified device on the specified host has been lost. This could happen if a SCSI reset occurred. Usually if this occurs, the device can be rereserved. However, in this case, the ASE agent daemon cannot open the device special file for the specified device, so the reservation cannot be redeemed.

```
AM failed to rereserve <device> on <host>
```

The disk reservation was lost because of a SCSI reset, and the ASE agent daemon was unable to rereserve the device.

```
AM reports a lost reservation for <device> on <host>
```

The reservation for the specified device was lost on the specified host. The reservation may have been taken when the ASE director daemon started the service on a different host.

```
Can’t fetch new configuration data base!
```

The ASE agent daemon stored a new configuration database, but cannot fetch the new database. The ASE agent daemon exits, and the system manager must resolve the problem.

```
Network is partitioned between local host and <remote_host>
```

The ASE agent daemon has discovered, through the host status monitor (HSM) daemon, that the local member is separated from the specified remote host because of a network partition. The system manager may have to resolve this condition, which could be caused by bad cable routing. The ASE agent daemon logs an Alert message for each member that is cut off from the partitioned member.

```
Cut off from net and can’t stop services. reboot!
```

The ASE agent daemon has been cut off from the network; therefore, it is stopping all of the services currently running on the member so they can be started on another member. If a stop fails, `umount` may have failed because a file is open locally on the NFS file system. If the service is relocated to another member and later relocated back to the original member and the member’s cache for the file system did not get flushed because of the failed `umount`, the cache could get flushed when the service restarts on the original member and could cause file corruption. To prevent this, the ASE agent daemon reboots the local node.

```
Possible security breach attempt: connect tried from unknown <remote_host>
```

```
Possible security breach attempt: connect request from non-member <remote_host>
```

A process on a nonmember system tried to connect to the ASE agent daemon. For security purposes, the ASE agent daemon’s connection maintenance code refuses connection requests from systems that are not in the ASE agent daemon’s current member list. One of the previous Alert messages is logged if a connection request is received from a nonmember system.

```
main: fatal error...
```

The ASE agent daemon encountered an error from which it could not recover and exited. This Alert message is logged, in addition to more detailed Alert messages that describe the reason that the ASE agent daemon exited.

```
possible device failure: <device>
```

The ASE agent daemon tried to start a service but discovered that the devices used by that service are unreachable.

A.2 ASE Director Daemon Alert Messages

This section describes some Alert messages generated by the available server environment (ASE) director daemon.

```
Lost connection to the HSM... exiting
```

The ASE director daemon exited because it lost its connection to the ASE host status monitor (HSM) daemon.

```
Possible security breach attempt: connect tried from unknown <remote_host>
```

```
Possible security breach attempt: connect request from nonmember <remote_host>
```

A process on a nonmember system tried to connect to the ASE director daemon. For security purposes, the ASE director daemon’s connection maintenance code refuses connection requests from systems that are not in the ASE director daemon’s current member list. One of the previous Alert messages is logged if a connection request is received from a nonmember system.

```
Unable to start service <service>
```

The ASE director daemon cannot start the specified service. If a service is restricted to run on a subset of the members, this message indicates that it cannot run on any of those members. Check the appropriate `daemon.log` event logging file for more information.

```
Unable to stop service <service> due to a timeout.  The service is in an
unknown state.
```

The ASE director daemon timed out waiting for the ASE agent daemon to reply to a stop service request.

```
Cannot contact local agent... exiting
```

The ASE director daemon exited because it could not contact its local agent.

```
Network connection down... exiting
```

The ASE director daemon exited because its network connection was not available.

```
Received message from agent which is not in the config database
```

The ASE director daemon received a message from a nonmember agent.

```
Can’t ping my agent, exiting...
```

The ASE agent daemon on the member that is running the ASE director daemon is not registered with the `portmap` daemon.

```
Unable to start service <service> on <host>.
```

A service relocation failed.

```
Cannot start service <service>.
```

After a device failure, a service cannot be started on any potential member.

```
Member <member> is not available.
```

A member is not answering pings over the network or the SCSI bus; therefore, it is considered unavailable.

```
Service <service> cannot be run on any available members.
```

The ASE director daemon cannot start the specified service. If a service is restricted to run on a subset of the members, this message indicates that it cannot run on any of those members. Check the appropriate `daemon.log` event logging file for more information.

```
Unable to stop service <service> due to a timeout.
The service is in an unknown state.
```

One of the stop scripts did not return an exit value within its timeout period, and the stop action may not have completed. It is important to ensure that the service is completely stopped before continuing.

```
A member has an invalid IP address.
```

```
ASE members are on different subnets.
```

The Internet Protocol (IP) address must be a valid address on the same subnet as the other ASE members.

```
Can’t ping agent on <member>
```

The ASE agent daemon on the specified member is not registered with the `portmap` daemon.

```
Can’t open channel to agent on <member>
```

The ASE director daemon cannot establish a connection with the agent on the specified member.

A.3 ASE Host Status Monitor Daemon Alert Messages

This section describes some Alert messages generated by the available server environment (ASE) host status monitor (HSM) daemon.

```
Network ping to host <host> is working but SCSI ping is not
```

A problem exists in all of the SCSI bus paths between the host specified in the message and the member that reported the message. Check the cabling between systems and disks on the shared buses.

```
Network ping to host <host> is working and now SCSI ping is also working
```

The condition described in the first ASE HSM daemon message has been cleared. SCSI pings can now be sent between the hosts on at least one of the shared buses.

A.4 The asemgr Utility Alert Messages

This section describes some Alert messages generated by the `asemgr` utility.

```
Test of alert script
```

This message is generated when you chose the "Test the error alert script" item from the `asemgr` utility's Managing the ASE menu.

```
Bad return code from ****
```

This message is generated by a return code from a routine that was not expected and indicates a bug in the TruCluster software. If it occurs, contact your field service representative.

```
Net partition - cannot find a director.
```

The `asemgr` utility cannot find the ASE director daemon because of a network partition.

```
Unable to translate host <host> to an IP address
```

A routine cannot map a member host name to an Internet Protocol (IP) address. There could be a problem with the `/etc/hosts` file or with Berkeley Internet Name Domain (BIND).

```
Could not allocate database
```

```
Could not malloc
```

These messages occur if a `malloc` operation fails. They indicate that the system is running out of memory or swap space.

```
Configuration database is corrupted (Invalid length of ASE version)
```

```
BUG NOTICE: Exit before finishing unmarshal_tree
```

Something is wrong with the ASE database (for example, it has been corrupted).

A.5 MEMORY CHANNEL Alert Messages

This section describes some Alert messages generated by the MEMORY CHANNEL subsystem.

```
memory channel - alternate on-line
```

In a redundant MEMORY CHANNEL configuration, the alternate MEMORY CHANNEL interconnect has come on line. This message is printed only when the alternate comes on line after MEMORY CHANNEL software initialization.

```
switching from mc<number> to mc<number>
```

The cluster is failing over from the primary MEMORY CHANNEL interconnect to the secondary MEMORY CHANNEL interconnect.

```
rm_sw_init: can’t fail over from mc<number> to mc<number>
```

The cluster cannot fail over to the secondary MEMORY CHANNEL interconnect due to hardware problems with the secondary MEMORY CHANNEL interconnect.

```
requesting memory channel failover, node <node>
```

A member system is requesting other member systems to fail over to the secondary MEMORY CHANNEL interconnect.

```
memory channel - checking cables
```

The MEMORY CHANNEL subsystem is checking that the primary MEMORY CHANNEL interconnect is plugged into the same hub on all member systems.

```
memory channel failover request from node <node>
```

A MEMORY CHANNEL failover request has been received from the specified member system.

```
rm_boot_request_init: didn’t switch
```

The cluster cannot fail over to the secondary MEMORY CHANNEL interconnect due to hardware problems with the secondary MEMORY CHANNEL interconnect.

```
memory channel node <node> already cluster member,crashing
node <node>
```

A node that has been identified as a cluster member is requesting cluster membership. The MEMORY CHANNEL subsystem will shut it down to restore consistency.

```
memory channel - failed initialization
```

A hardware problem has prevented MEMORY CHANNEL subsystem initialization.

```
received a request from node <node> to failover
```

The specified node has requested a failover to the secondary MEMORY CHANNEL interconnect.

```
rm_failover_rmerror_request: can’t fail over from mc<number> to mc<number>
```

Failover to the secondary MEMORY CHANNEL interconnect is not possible, probably due to a member system’s not being able to access the secondary MEMORY CHANNEL interconnect.

```
rmerror_get_errcnt_kl:crashing node <node>
```

The specified MEMORY CHANNEL node is unresponsive and is being shut down.

```
rmerror_free_errcnt_lk: Too many retries, node <node>
must be down
```

```
rmerror_init:Error_count = <number> unit = <number>
Err_reg = <value> Node = <node>
```

A MEMORY CHANNEL error interrupt has been received and error recovery is in progress.

```
rmerror_init:crashing node <node>
```

The specified node is unresponsive and is being shut down.

```
rmerror_state_change:
         unit = <number>  Err_reg = <value> node = <node>
```

A state change has been received, indicating that another member system has joined or left the cluster.

```
rmerror_state_change: failed to failover
```

The cluster made an unsuccessful attempt to fail over from the primary MEMORY CHANNEL interconnect to the secondary MEMORY CHANNEL interconnect. It is likely that a member system cannot access the secondary MEMORY CHANNEL interconnect.

```
rmerror_railover:
         Node = <node>  Flag = <value>  Action = <value>
```

The MEMORY CHANNEL subsystem has requested a failover to the secondary MEMORY CHANNEL interconnect.

`rmerror_failover: no alternate mc to fail over to`

No functional secondary MEMORY CHANNEL interconnect is available for failover.

`rmerror_failover: negative error count`

Failover has been simultaneously initiated on multiple member systems. This is an informational message.

`rmerror_failover_1:crashing node <node>`

The specified MEMORY CHANNEL node is unresponsive and is being shut down.

`rmerror_failover: not every node can failover`

The cluster aborted a failover to the secondary MEMORY CHANNEL interconnect, because not all member systems could fail over to it.

`rmerror_failover_2: crashing node <node>`

The specified MEMORY CHANNEL node is unresponsive and is being shut down.

`checking for existing memory channel nodes`

The MEMORY CHANNEL subsystem is looking for other nodes connected to the MEMORY CHANNEL interconnect that may be either running or in the process of booting.

`unresponsive mc nodes - waiting for node mask`

A node connected to the MEMORY CHANNEL interconnect is not responding to boot requests. The MEMORY CHANNEL subsystem is waiting for the node to boot.

`crashing unresponsive node <node>`

The node indicated in the message did not respond to repeated boot requests. It may be hung, so the MEMORY CHANNEL software attempts to crash it to allow cluster formation to progress. This `crashing ...` message is usually preceded by several `unresponsive mc nodes ...` messages.

`booting as primary memory channel node`

This MEMORY CHANNEL node is the first node to boot and initialize its MEMORY CHANNEL subsystem.

`memory channel software inited - node <node>`

Initialization of low-level MEMORY CHANNEL software is complete.

```
requesting memory channel interrupt, node <node>
```

This MEMORY CHANNEL node has requested an interrupt from another node, which has already initialized its low-level MEMORY CHANNEL software. This is the first step a node takes to initialize its MEMORY CHANNEL software when another MEMORY CHANNEL node is already initialized.

```
requesting memory channel update interrupt, node <node>
```

This MEMORY CHANNEL node has requested an update interrupt from another node, which has already initialized its low-level MEMORY CHANNEL software. This is the second step a node takes to initialize its MEMORY CHANNEL software when another MEMORY CHANNEL node is already initialized.

```
memory channel status request from node <node>
```

A MEMORY CHANNEL node is looking for other existing MEMORY CHANNEL nodes.

```
memory channel request from node <node>
```

This MEMORY CHANNEL node is responding to an interrupt from another node, which is attempting to initialize its low-level MEMORY CHANNEL software. This is the first step a node takes to initialize its MEMORY CHANNEL software when another MEMORY CHANNEL node is already initialized.

```
memory channel update request from node <node>
```

This MEMORY CHANNEL node is responding to an update interrupt from another node, which is attempting to initialize its low-level MEMORY CHANNEL software. This is the second step a node takes to initialize its MEMORY CHANNEL software when another MEMORY CHANNEL node is already initialized.

```
memory channel - adding node <node>
```

Low-level MEMORY CHANNEL software is adding another node.

```
memory channel - removing node <node>
```

Low-level MEMORY CHANNEL software is removing a node.

```
memory channel node <node> timed out, hardware does not see it
```

A node is not responding. It will be removed.

```
memory channel thread init
```

The general-purpose MEMORY CHANNEL thread has completed initialization.

A.6 DLM Alert Messages

This section describes some Alert messages generated by the distributed lock manager (DLM) subsystem. These messages are logged to the console and kernel log in the `/usr/adm/syslog.dated` file. Some, as noted, are also logged to the user's terminal.

```
dlm_subsys_configure: can't init lkid table
```

Either the system has an insufficient amount of memory or the number of locks allocated at boot time (indicated by the `dlm_locks_cfg` kernel attribute) is too large. To fix this problem, either decrease the value of the `dlm_locks_cfg` kernel attribute in the `/etc/sysconfigtab` file and reboot, or add more memory to the system.

```
dlm_subsys_configure: can't init rsb table
```

Either the system has an insufficient amount of memory or the size of the DLM resource hash table (indicated by the `rhash_size` kernel attribute) is too large. To fix this problem, either decrease the value of the `rhash_size` kernel attribute in the `/etc/sysconfigtab` file and reboot, or add more memory to the system.

```
dlm_subsys_configure: can't init pdb table
```

Either the system has an insufficient amount of memory or the size of the process descriptor block hash table (indicated by the `pdb_hash_size` kernel attribute) is too large. To fix this problem, either decrease the value of the `pdb_hash_size` kernel attribute in the `/etc/sysconfigtab` file and reboot, or add more memory to the system.

```
dlm_subsys_configure: can't start timeoutq
```

The DLM cannot start the DLM timeout queue thread. To fix this problem, reboot the member system. If the problem recurs, contact your DIGITAL support representative.

```
dlm_subsys_configure: dlm_hab_configure failed
```

The DLM cannot configure its habitat. To fix this problem, reboot the member system. If the problem recurs, contact your DIGITAL support representative.

```
dlm: configured
```

The DLM subsystem has been configured successfully at boot time.

```
dlm_subsys_configure: configure failed
```

DLM configuration has failed on this member system.

```
dlm_create_lock: pid <value> copyout err of lkid <value>
```

The `dlm_lock` or `dlm_quelock` function cannot return a lock ID to the buffer specified in the function call. This message is sent to the console, system log, and the user's terminal. To fix this problem, check the application program that called the function and ensure that it passes a valid buffer address. See the TruCluster Production Server *Application Programming Interfaces* manual for more information about DLM functions.

```
dlm_create_lock
    COMP_LOCK: pid <value> IVLOCKID lkid <value> uaddr <value>
```

The `dlm_lock` function has attempted to use an invalid lock ID. This can occur when the function is interrupted by a signal and, before it resumed, the application that called the function dequeued the lock or corrupted the lock ID in its signal handler. This message is sent to the console, system log, and the user's terminal. See the TruCluster Production Server *Application Programming Interfaces* manual for more information about DLM functions.

```
dlm_create_lock:

    pid <value> err while copying out valblk for lkid <value>

dlm_convert_lock:

    valblk copyout fault: lkid <value> kvalbp <value>

                   uvalb_p <value>
```

The `dlm_lock`, `dlm_quelock`, `dlm_cvt`, or `dlm_quecvt` function cannot return the resource's value block to the buffer specified in the function call. This message is sent to the console, system log, and the user's terminal. To fix this problem, check the application program that called the function and ensure that it passes a valid buffer address. See the TruCluster Production Server *Application Programming Interfaces* manual for more information about DLM functions.

```
dlm_collect: pid <value> DLM_EFAULT notf_entry

dlm_collect: pid <value> DLM_EFAULT ngot
```

The `dlm_notify` function cannot return the blocking notification routine parameter or hint to the buffer specified in the function call. This message is sent to the console, system log, and the user's terminal. To fix this problem, check the application program that called the function and ensure that it passes a valid buffer address. See the TruCluster Production Server *Application Programming Interfaces* manual for more information about DLM functions.

A.7 DRD Alert Messages

This section describes some Alert messages generated by the distributed raw disk (DRD) subsystem. DRD messages are logged to the `kern.log` file unless otherwise noted.

```
drd_configure_subsys: failed in drd_driver_configure
```

One of the subcomponents of DRD was unable to initialize. There will usually be an accompanying error message providing more detail on the cause of the initialization error. Verify that all hardware components are operational.

```
drd_configure: drd-maphash-size, invalid size.
```

```
drd_bp_pool_configure: bogus tunables, using default.
```

An invalid value was specified for a tunable parameter. See `drd` (7) for a description of the tunable parameters.

```
drd_configure: subsystem unconfiguration not yet supported.
```

The DRD subsystem cannot be dynamically unconfigured.

```
drd_map_delete: can’t delete, drain failed.
```

The underlying physical device driver failed to complete outstanding I/O operations. Check the system error logs for driver-specific errors.

```
drd_map_add: LMF PAK not registered.
```

The required product license has not been registered. Use the `lmf` command to register the appropriate Production Authorization Key (PAK).

```
drd_map_add: rejecting map on validation errors.
```

A corrupt or invalid map entry has been received. The underlying device driver type or device type may not be supported or operational.

```
drd_resolve_map: Can’t find server for DRD drd3.
```

The DRD subsystem has repeatedly tried to determine which node within the cluster is the server of the specified disk. The retry timeout limit has been reached and the DRD subsystem is returning an error to the calling application. Use the `asemgr` utility to verify that the service is operational. This error could indicate that stale DRD device special files are being accessed.

```
drd_map_rpc_add: attempt to replace local map with remote.
```

```
drd_map_rpc_add: rejecting new remote map.
```

The server of a DRD disk has received a new map entry, indicating that another node also believes that it is the disk's server. At any given time, there should be only one server. Check for the occurrence of any errors from the available server environment (ASE) subsystem that may determine the cause of this problem.

```
drd_map_rpc_add: drd_map_add() failed with 23.
```

The error number specifies an error return status when attempting to add a new DRD map entry. One of the validation checks has failed, or a DRD server is unable to open the underlying device.

```
bss_open: device type not supported with drd.
```

The underlying driver type or device type does not meet DRD's validation requirements.

```
bss_rm_init: register_RM_member_callback failed
```

```
bss_rm_init: get_RM_information failed
```

```
bss_rm_init_sync: RM_GET_CHARACTERISTICS failed
```

```
bsc_rm_init: get_RM_information failed
```

```
bsc_rm_init: register_RM_member_callback failed
```

The underlying MEMORY CHANNEL subsystem is returning an error status to the DRD. Check for related error messages and verify that the MEMORY CHANNEL hardware is operational.

```
bss_subsys_configure: invalid bssd count
```

This indicates that the argument passed to the bssd command is not within the acceptable range. See bssd(8) for details.

```
bss_subsys_configure: bssd not restartable, reboot needed.
```

The bssd daemon cannot be killed and restarted. See bssd(8) for details. This message indicates that the bssd daemon has been killed off and an attempt has been made to restart it. To restart the bssd daemon, reboot the system.

```
bss_subsys_configure: bssd already running.
```

An attempt was made to start a second bssd daemon while one bssd daemon is already running. Only one instance of the bssd daemon is allowed to run at a time.

```
Another BSSD is already running, exiting.
```

An attempt was made to start a second bssd daemon while one bssd daemon was already running. Only one instance of the bssd daemon is allowed to run at a time. This message appears in the daemon.log file.

```
Daemon not restartable.  Reboot required.
```

The bssd daemon has been killed off and an attempt is being made to restart it. The bssd daemon cannot be restarted. You must reboot the system reboot to make DRD operational again. This message appears in the daemon.log file.

```
Can’t register with portmap.
```

The portmapper daemon may have been killed off, which prevents the bssd daemon from establishing its network connection. This message appears in the daemon.log file.

```
Returned from kernel call, exiting.
```

The bssd daemon is exiting. This informational message appears in the daemon.log file.

```
DRD is NOT loaded and configured.
```

```
Unable to open DRD control device
```

The DRD subsystem has not been initialized or has failed to initialize. Check the kern.log file for related error messages. This message appears in the daemon.log file.

```
Failed ioctl to set socket descriptor.
```

The kernel portion of DRD has rejected the bssd daemon’s attempts to connect. Check the kern.log file for related error messages. This message appears in the daemon.log file.

B

Kernel Attributes (PS)

This appendix contains a partial list of kernel attributes provided by each Production Server Software subystem.

Use the following command to display the current settings of these attributes for a given subystem:

```
# sysconfig -q <subsystem-name> <attribute-list>
```

Table B–1 shows the subsystem name associated with each Production Server Software component.

Table B–1: Configurable Production Server Subsystems

Subsystem Name	Component	Attributes
`clubase`	Cluster base component	See Table B–2.
`cnxagent`	Connection manager agent daemon	See Table B–3.
`drd`	Distributed raw disk (DRD)	See the `drd`(7) reference page.
`dlm`	Distributed lock manager (DLM)	See Table B–4.
`dlmsl`	DLM layer	
`mcnet`	Internet Protocol (IP) over MEMORY CHANNEL	See Table B–5.
`mcs`	MEMORY CHANNEL application programming interface (API)	See Table B–7.
`rm`[a]	MEMORY CHANNEL subsystem	See Table B–6.

[a] The `rm` facility name represents remote memory, a synonym for MEMORY CHANNEL.

You tune the performance of a kernel subsystem by using one of the following methods to set one or more attributes in the `/etc/sysconfigtab` file:

- Add or edit a `<subsystem-name>` stanza entry in the `/etc/sysconfigtab` file to change an attribute's value and have the new value take effect at the next system boot.
- Use the following command to change the value of an attribute that can be reset so that its new value takes effect immediately at run time:

```
# sysconfig -r <subsystem-name> <attribute-list>
```

To allow the change to be preserved over the next system boot, you must also edit the /etc/sysconfigtab file. For example, to change the value of the drd-print-info attribute to 1, enter the following command:

```
# sysconfig -r drd drd-print-info=1

drd-print-info: reconfigured
```

You can also use the configuration manager framework, as described in the DIGITAL UNIX *System Administration* manual, to change attributes and otherwise administer a cluster kernel subsystem on another host. To do this, set up the host names in the /etc/cfgmgr.auth file on the remote client system and then specify the –h flag to the /sbin/sysconfig command, as in the following example:

```
# sysconfig -h fcbra13 -r drd drd-do-local-io=0

drd-do-local-io: reconfigured
```

Note

The default settings of configurable cluster subsystem kernel attributes should be enough for most purposes. Although most kernel attributes are described in this section, many are reserved for debugging, development, and testing purposes.

Table B–2 lists the cluster base kernel attributes.

Table B–2: Cluster Base Kernel Attributes

Attribute	Type	Description
cluster_disable	Query; set at boot time	When set to 1, disables all cluster components. Typically, you disable the cluster software when you are upgrading the DIGITAL UNIX operating system underlying the running version of the TruCluster Production Server software. See the TruCluster Software Products *Software Installation* manual for more information.
cluster_maj_version	Query	Major number of the TruCluster software product. For example, the major version of Version 1.5 is 1.

Table B–2: Cluster Base Kernel Attributes (cont.)

Attribute	Type	Description
cluster_min_version	Query	Minor number of the TruCluster software product. For example, the minor version of Version 1.5 is 5.
cluster_version	Query	String representing the installed version of the TruCluster Software product. For example, "Digital UNIX TruCluster T1.5-3 (Rev. 181.4); 08/04/97 09:27."

Table B–3 lists the cluster agent daemon kernel attributes.

Table B–3: Cluster Agent Daemon Kernel Attributes

Attribute	Type	Description
cnx_debug_msg_level	Query; set at boot time or run time	Increases or decreases the amount of debugging messages issued by the connection manager agent daemon. The default value is 1.

Table B–4 lists the distributed lock manager (DLM) kernel attributes (DLM statistics).

Table B–4: Distributed Lock Manager Kernel Attributes (DLM Statistics)

Attribute	Type	Description
dlm_name	Query	Name and version number of the DLM.
dlm_lkbs_allocated	Query	Number of active locks on this member system. The value of this attribute accounts for all copies of locks on this node.
dlm_rsbs_allocated	Query	Number of resources mastered on this member system. Each resource name is represented by a resource block. When a process on the local system has a lock on a given resource, a resource block exists on both the system that is the resource's master and on the local system. The locks on a given resource are represented by a chain of lock blocks connected to the resource block.

Table B–4: Distributed Lock Manager Kernel Attributes (DLM Statistics) (cont.)

Attribute	Type	Description
dlm_tot_lkids	Query	Total number of lock IDs on this cluster member. Each member system has its own lock ID table, which describes each lock block that is in use on that system. Each allocated lock block is identified by a lock ID handle. The DLM uses this table to quickly access lock blocks.
dlm_lkids_inuse	Query	Number of lock IDs actually in use on this member system. This value matches the number of active locks (dlm_lkbs_allocated).
dlm_ddlckq_len	Query	Number of locks currently on the deadlock queue.
dlm_timeoutq_len	Query	Number of locks currently on the timeout queue.
dlm_lock_in	Query	Number of new lock requests from other cluster members to this member.
dlm_lock_loc	Query	Number of new lock requests that are mastered on this member system.
dlm_lock_out	Query	Number of new lock requests from this member to locks mastered on other member systems.
dlm_mng_local	Query	Number of new lock requests from this member to either the director or other member systems thought to be the lock master.
dlm_resend	Query	Number of new lock requests sent from this member to other member systems that resulted in a resend response. A cluster member returns a resend response when it is not, or is no longer, the master for the requested lock. For example, this can happen when the lock's master is transitioning from one system to another. A high number of resend responses indicates that the lock master is moving from member to member, which affects performance.

Table B–4: Distributed Lock Manager Kernel Attributes (DLM Statistics) (cont.)

Attribute	Type	Description
`dlm_retry`	Query	Number of new lock requests sent from this member to other member systems that resulted in a retry response. For example, a cluster member returns a retry response, for instance, when it has insufficient resources to satisfy the request, or if a lock rebuild started when the request was being sent.
`dlm_cvt_in`	Query	Number of lock conversion requests from other member systems to this member as the resource master.
`dlm_cvt_loc`	Query	Number of lock conversion requests for resources mastered locally.
`dlm_cvt_out`	Query	Number of lock conversion requests from this member to locks mastered on other member systems.
`dlm_unlock_in`	Query	Number of unlock requests from other cluster members to this member as the resource master.
`dlm_unlock_loc`	Query	Number of unlock requests for resources mastered locally.
`dlm_unlock_out`	Query	Number of unlock requests from this member to locks mastered on other member systems.
`dlm_blk_in`	Query	Number of blocking notifications sent from member systems mastering resources to processes on the local system.
`dlm_blk_loc`	Query	Number of blocking notifications sent to processes on the local system for resources mastered on the local system.
`dlm_blk_out`	Query	Number of blocking notifications sent to processes on other systems for resources mastered on the local system.

Table B–4: Distributed Lock Manager Kernel Attributes (DLM Statistics) (cont.)

Attribute	Type	Description
`dlm_dir_in`	Query	Number of times the local node received new lock requests as the directory node. This usually results in a resend request or a message to the requestor to master the lock locally.
`dlm_dir_out`	Query	Number of requests the local node has sent to other directory nodes.
`dlm_timeo_lks`	Query	Number of locks that have timed out (that is, have not been granted within the timeout interval specified on the initial request).
`dlm_ddlck_in`	Query	Number of deadlock search requests involving locks or processes on the local system.
`dlm_ddlck_out`	Query	Number of deadlock search requests from this member to other member systems.
`dlm_ddlck_srch`	Query	Number of deadlock searches that have occurred on the local system.
`dlm_ddlck_cvt_fnd`	Query	Number of conversion deadlocks detected on the local system. Conversion deadlocks are described in the TruCluster Production Server Software *Application Programming Interfaces* manual.
`dlm_ddlck_res_fnd`	Query	Number of multiple resource deadlocks found on the local system. Multiple resource deadlocks are described in the TruCluster Production Server Software *Application Programming Interfaces* manual.
`dlm_notqd`	Query	Number of lock requests that were not queued (that is, those requests that either were granted or will be granted).
`dlm_wait`	Query	Number of new lock requests that were placed on the waiting queue of the resource.

Table B–4: Distributed Lock Manager Kernel Attributes (DLM Statistics) (cont.)

Attribute	Type	Description
dlm_cvt_wait	Query	Number of conversion lock requests that were placed on the conversion queue of the resource.
dlm_collect_zero	Query	Number of times a process attempted to collect a blocking notification after the blocking notification had been cancelled or delivered.
dlm_no_msg_mem	Query	Number of times the DLM has failed in an attempt to allocated memory for a message.
dlm_add_lkid_seg	Query	Indicates the number of times the DLM had to allocate memory for another block (segment) of lock ID handles.
dlm_expand_lkid_tbl	Query	Indicates the number of times the DLM had to expand the lock ID segment table (an array of pointers to lock ID segments).
dlm_pdbs_allocated	Query	Number of processes running on this cluster member that are using the DLM.
dlm_kpdbs_allocated	Query	This is equivalent to the dlm_pdbs_allocated attribute, except it relates to kernel threads using kernel DLM interfaces instead of user processes.
dlm_gdbs_allocated	Query	Number of group lock containers on this cluster member.
dlm_perm_gdbs	Query	Number of permanent group lock containers on this member system.
dlm_txids_allocated	Query	Number of transactions active on this cluster member.
dlm_long_rhash_chain	Query	Length of the longest linked list in the resource hash table.
dlm_empty_rhash_chain	Query	Number of empty linked lists (chains) in the resource hash table.

Table B–4: Distributed Lock Manager Kernel Attributes (DLM Statistics) (cont.)

Attribute	Type	Description
`dlm_txid_wrap`	Query	Set when the number of DLM transactions done on a single lock ID handle has overflowed the largest value that can fit in an unsigned long datum.
`rhash_size`	Query; set at boot time	Number of entries in the resource hash table on this member system. Increasing the size of the resource hash table can shorten the lengths of the chains of which the table is comprised. By reducing the time it takes for the DLM to search these chains, this can improve DLM performance. The default is 8192 entries.
`pdb_hash_size`	Query; set at boot time	Number of entries in the process hash table on this member system. Set this attribute to the approximate number of processes that will use the DLM on this system. The default is 64 entries.
`kpdb_hash_size`	Query; set at boot time	This is equivalent to the `pdb_hash_size` attribute, except it relates to kernel threads using kernel DLM interfaces instead of user processes.
`gdb_hash_size`	Query; set at boot time	Number of entries in the group lock container hash table on this member system. Set this attribute to the approximate number of group lock containers that will be on this member system. The default is 64 entries.
`txid_hash_size`	Query; set at boot time	Number of entries in the transaction hash table on this member system. Set this attribute to the approximate number of transactions that will be active in the cluster. The default is 64 entries.
`dlm_deadlock_wait`	Query; set at boot time	Length of time (in seconds) a lock waits on this member system to be included in a deadlock search. The default is 10 seconds.

Table B–4: Distributed Lock Manager Kernel Attributes (DLM Statistics) (cont.)

Attribute	Type	Description
`dlm_dirwt`	Query; set at boot time	Directory weight value. Setting the directory weight value increases the likelihood that this node will be the director of a resource. Its value can range from 0 to 4, where 4 provides the highest likelihood and 0 is the default.
`dlm_disable_grptx`	Query; set at boot time	When set, indicates that the DLM group container and transaction ID features have not been not enabled. Its value should always be false (0).
`dlm_disable_rd`	Query; set at boot time	When set, indicates that the DLM recovery domain domain (persistent resource) support has not been not enabled. This should be a temporary condition during rolling upgrade. Once the upgrade is complete this attribute should always be false (0).
`dlm_kernel_interfaces_enabled`	Query; set at boot time	When set, indicates that the DLM subsystem has been initialized such that the DLM kernel programming interfaces are usable. If not set, kernel calls into the DLM kernel API causes the DLM_ENOSYS status to be returned. This attribute is the equivalent of the `dlm_enabled` attribute but applies to the DLM kernel API.
`dlm_validate_rsb`	Query; set at boot time or run time	Validate resource block. [Tuning not supported.]
`dlm_locks_cfg`	Query; set at boot time	Number of locks configured on this member system at boot time. To improve lock startup performance, set this attribute to twice the number of expected locks.
`dlm_deadlock_scan`	Query; set at boot time or run time	Number of seconds that determines how often the DLM scans the deadlock queue on this member system. The default is once per second.

Table B–4: Distributed Lock Manager Kernel Attributes (DLM Statistics) (cont.)

Attribute	Type	Description
`dlm_zero_stats`	Query; set at boot time or run time	Reset all DLM statistics to zero (0).
`dlm_rd_count`	Query	Number of recovery domains in the cluster.
`dlm_rd_in`	Query	Number of recovery domain messages sent by this node.
`dlm_rd_out`	Query	Number of recovery domain messages received by this node.

Table B–5 lists the kernel attributes that configure the MEMORY CHANNEL network driver.

Table B–5: MEMORY CHANNEL Network Driver Kernel Attributes

Attribute	Type	Description
`dochecksum`	Query set at boot time and at run time	When set to 1 (the default), the node checksums all incoming and outgoing messages through Transmission Control Protocol/Internet Protocol (TCP/IP). When set to 0, checksums are not performed. Turning checksums off results in a slight increase in MEMORY CHANNEL performance.

Table B–5: MEMORY CHANNEL Network Driver Kernel Attributes (cont.)

Attribute	Type	Description
		This attribute must be set to the same value on all cluster members. Once you have set this attribute on all member systems, you must halt the entire cluster and reboot each member to effect the change.
`rx_mapping_enabled`	Query set at boot time and at run time	By default, this attribute is set to zero (0), which enables receive mapping in the cluster communications subsystem. Disabling receive mapping by setting this attribute to 1 results in a slight performance gain. This attribute must be set to the same value on all cluster members. Once you have set this attribute on all member systems, you must halt the entire cluster and reboot each member to effect the change.

Table B–6 lists the MEMORY CHANNEL kernel attributes.

Table B–6: MEMORY CHANNEL Kernel Attributes

Attribute	Type	Description
`rm_char_connected`	Query	Bitmap representing the nodes connected to a MEMORY CHANNEL hub.
`rm_char_member`	Query	Bitmap representing the nodes participating in the cluster.

Table B–6: MEMORY CHANNEL Kernel Attributes (cont.)

Attribute	Type	Description
`rm_char_flags`	Query	Flags. The flags attribute can have the following settings: `RM_CHAR_FLAG_DUAL`: Member participates in a no-single-point-of-failure, redundant MEMORY CHANNEL configuration. `RM_CHAR_FLAG_PRIM_HUBLESS`: MEMORY CHANNEL interconnect uses a MEMORY CHANNEL configuration in virtual hub mode or, in a redundant MEMORY CHANNEL configuration, the primary MEMORY CHANNEL interconnect uses virtual hub node. `RM_CHAR_FLAG_ALT_HUBLESS`: In a redundant MEMORY CHANNEL configuration, the secondary MEMORY CHANNEL interconnect uses a virtual hub mode.
`rm_char_alloc_count`	Query	Number of pages of MEMORY CHANNEL address space that have been allocated.
`rm_char_prim_alt`	Query	Zero (0) if the member is using the primary MEMORY CHANNEL interconnect in a redundant MEMORY CHANNEL configuration, 1 if it is using the secondary MEMORY CHANNEL interconnect.
`rm_no_inheritance`	Query; set at boot time	When 1, causes the child process not to inherit certain MEMORY CHANNEL objects when a process forks. Because inheritance of these objects is not supported by internal accounting mechanisms, you must leave this attribute set to 1.
`rm_rail_style`	Query set at boot time	Sets multirail reliability styles, as follows: 0 for `RM_STYLE_SINGLE` and 1 for `RM_STYLE_FOP` (failover pair). This attribute must be set to the same value on all cluster members. Once you have set this attribute on all member systems, you must halt the entire cluster and reboot each member to effect the change.[a]

Table B–6: MEMORY CHANNEL Kernel Attributes (cont.)

Attribute	Type	Description
rm_errors	Query; set at boot time	Indicates the number of errors at boot time.
rm_error_interval	Query; set at boot time	Specifies the interval during which the number of errors indicated in rm_error_threshold are monitored. The default interval is 60 seconds.
rm_error_threshold	Query; set at boot time	Specifies the number of errors allowed by a rail during a given interval before the rail is considered faulty, and a failover occurs to another rail (if another rail has been specified as the failover rail).

[a] RM_STYLE_SINGLE (rm_rail_style=0) treats each single physical MEMORY CHANNEL interconnect (or rail) in a system as a logical connection. The MEMORY CHANNEL application programming interface (API) library (shipped with the TruCluster Production Server Software and TruCluster MEMORY CHANNEL Software products) can use multiple logical rails simultaneously at their aggregate bandwidth and memory capacity. Failover between rails is not supported for this multirail reliability style.

RM_STYLE_FOP (rm_rail_style=1) treats each pair of MEMORY CHANNEL adapters on a system as a single logical rail. There is no gain in bandwidth or memory capacity beyond that of a single physical connection; but if one physical connection fails, the other provides failover capability. RM_STYLE_FOP is the default rm_rail_style value.

For more information, see the TruCluster Production Server Software *MEMORY CHANNEL Application Programming Interfaces* manual.

Table B–7 lists the MEMORY CHANNEL API kernel attributes.

Table B–7: MEMORY CHANNEL API Kernel Attributes

Attribute	Type	Description
mcs_dbg1	Query; set at boot time	Tuning is not supported.
mcs_num_events	Query; set at boot time	Tuning is not supported.
mcs_ignore tags	Query; set at boot time	Tuning is not supported.

Table B–7: MEMORY CHANNEL API Kernel Attributes (cont.)

Attribute	Type	Description
mcs_mcm_eps	Query; set at boot time	Tuning is not supported.
mcs_rts	Query; set at boot time	Tuning is not supported.

C

Configuration Variables

Table C–1 contains a partial list of configuration variables provided by the TruCluster software products.

Table C–1: Cluster Configuration Variables

Variable	Description
`ASE`	When this variable is set to `on`, the available server environment (ASE) availability services are enabled on the member system.
`ASE_ID`	Specifies a value from 0 to 63 that the cluster software uses to uniquely identify the ASE in which a system resides in a Production Server cluster. Each ASE has a unique ASE identifier; all systems in the same ASE share the same ASE identifier. A member of an ASE in an Available Server configuration always has an ASE_ID of zero (0). See the TruCluster Software Products *Software Installation* manual for a discussion of ASE identifiers.
`ASELOGGER`	When 1, specifies that the ASE logger daemon (`aselogger`) is started on the system at boot time. See the TruCluster Software Products *Software Installation* manual for a discussion of the ASE logger daemon.
`ASE_PARTIAL_MIRRORING`	When this variable is set to `on`, prevents a service that uses Logical Storage Manager (LSM) mirroring from starting if only one plex of the mirrored data is available.
`ASEROUTING`	Enables host-based routes from a server that has multiple network interfaces. This can make for faster connections to a service by avoiding routers and making use of the multiple interfaces.
`CLUSTER_NET`	Specifies the Internet Protocol (IP) name of the MEMORY CHANNEL interconnect for TruCluster Production Server clusters and TruCluster MEMORY CHANNEL Software configurations, or the hostname in TruCluster Available Server configurations. See the TruCluster Software Products *Software Installation* manual for more information.

Table C–1: Cluster Configuration Variables (cont.)

Variable	Description
CMS_CONF	When set to on, indicates that the Cluster Monitor has been configured on the member system, and that submon and tractd are running.
CNX_DISK	Specifies the names of one or more tie-breaker disks used to avoid partitioning in a two-member cluster using a MEMORY CHANNEL connection in virtual hub mode. See Section 2.5 and cnxset(8) for more information.
CNX_INTERVAL	Specifies the timeout interval (also known as the maximum ping interval allowed), in seconds, for each member system. A member system times out if it is unable to send at least one ping in this interval. An effective ping interval is derived from this timeout value and sent to the ping daemon (cnxpingd) on each node. The /sbin/init.d/clumember script supplies a default value for this interval when starting the cnxpingd daemon.
CNX_VERBOSE	When this variable is set, the connection manager monitor daemon (cnxmond) runs in verbose mode, providing detailed diagnostic messages during its operation. The /sbin/init.d/clumember script supplies a default value for this parameter when starting the cnxmond daemon.
CNX_WAVES	Specifies the number of ping intervals (as specified by the −p flag) the connection manager's node monitor daemon should wait before removing from the cluster a member that cannot communicate. The /sbin/init.d/clumember script uses a default value for this parameter.
IMC_AUTO_INIT	When this variable is set to 1, the MEMORY CHANNEL API library is automatically initialized at boot time. This initialization involves reserving 0.5 MB for the MEMORY CHANNEL memory the application programming interface (API) library requires. The default value of this attribute is 1.
IMC_MAX_ALLOC	Determines the maximum aggregate amount of MEMORY CHANNEL address space the MEMORY CHANNEL API library can allocate for its use across the cluster. If the value of this variable differs among cluster members, the largest value specified on any individual member determines the value set for the cluster. The default amount of address space is 10 MB.

Table C–1: Cluster Configuration Variables (cont.)

Variable	Description
IMC_MAX_RECV	Determines the maximum amount of physical memory the MEMORY CHANNEL API library can map for reading MEMORY CHANNEL address space. This limit is node-specific and can vary from member to member. The default amount of address space is 10 MB.
NUM_BIOD	Indicates the number of asynchronous I/O servers for distributed raw disk (DRD) services (bsc_biod daemons) that are automatically started at boot time on this member system. See bsc_biod(8) for more information.
TCR_INSTALL	Indicates a successful installation when equal to MCS, ASE, or TCR. Indicates an unsuccessful installation when equal to BAD.
TCR_PACKAGE	Indicates a successful installation when equal to MCA, ASE, or TCR.

Glossary

The terms in this glossary are commonly used in a TruCluster software environment.

action script

Scripts that are used to make an application or data highly available by configuring an application or data on a member system. Action scripts break down a procedure (for example, starting an application or exporting data) into a series of steps, which are performed in order when executing that procedure. There are five types of action scripts: add, delete, start, stop, and check action scripts, and there are two versions of each type: internal action scripts, which cannot be modified manually, and user-defined action scripts, which allow you to customize the behavior of the service.

adapter

A device that converts the protocol and hardware interface of one bus type into that of another bus.

address switches

Electrical switches on the side or rear of some disk drives that determine the SCSI address setting for the drive.

advanced RISC computing

External interface to console firmware for operating systems that expect firmware compliance with the Advanced RISC Computing Standard Specification.

ARC

See advanced RISC computing.

available server environment

A set of systems, disks, shared SCSI buses, and software that allows you to configure applications and disks so that they are highly available to client systems.

ASE

See available server environment.

ASE_ID

A number from 0 to 63 that identifies an ASE within a cluster and allows the `asemgr` utility to generate unique clusterwide names for distributed raw disk (DRD) special files. Each ASE in a cluster has its own distinct ASE ID. All cluster members in the same ASE use the same ASE ID.

ASE service

A service that an administrator sets up in an ASE by using the `asemgr` utility. TruCluster software uses a service to maintain the availability of applications or data. A service consists of a unique name, an automatic service placement (ASP) policy, an application or disk specification, and action scripts that contain the commands to start and stop the application or to fail over the disk data. The action scripts implement the status changes for the service by performing necessary configuration changes and starting and stopping processes.

A member system in an ASE runs a service until a hardware or software failure or an explicit action by an administrator causes the service to run on another member system in the ASE.

Automatic Service Placement policy

Enables you to control which member systems are allowed to run a service. You must specify an ASP policy when you add a service. For example, you can allow any member system to run a service, or you can restrict a service to a specific member system or systems.

ASP policy

See automatic service placement policy.

availability

The amount of time that hardware or software is available during the time it is scheduled to be available. For the TruCluster software, the ability to function despite a specific hardware or software failure. See also **highly available.** To make an ASE service available despite a particular failure, it is necessary to make the hardware and software it depends on capable of operating despite that failure. For example, a distributed raw disk (DRD) service can be made available despite an MEMORY CHANNEL interconnect failure by configuring a redundant MEMORY CHANNEL interconnect so that if the primary MEMORY CHANNEL interconnect fails, the DRD service will use the other MEMORY CHANNEL interconnect.

bus

Flat or twisted-wire cable or a backplane composed of individual identical circuits. A bus interconnects computer system components to provide communications paths for addresses, data, and control information.

client

A computer system that uses resources provided by another computer, called a **server**.

cluster

A loosely coupled collection of servers that share storage and other resources that make applications and data highly available. A cluster consists of communications media, member systems, peripheral devices, and applications. The systems communicate over a high-performance interconnect.

cluster configuration map

A file (`/etc/CCM`) that statically records the hardware configuration of a cluster for display by the Cluster Monitor utility. You use the `cluster_map_create` utility to generate a cluster configuration map when you first configure a cluster and, subsequently, each time you add or remove hardware.

cluster interconnect

Private physical bus employed by cluster members for intracluster communications.

Cluster Monitor

Cluster software component that provides a graphical view of the cluster configuration. You can use the Cluster Monitor utility to monitor the availability of services and the connectivity among member systems in the cluster. You can also use it to manage services and to start disk management applications.

cold swap

The ability to turn off power to a device, replace it, and then turn on power to the device.

connection manager

Cluster software component that coordinates participation of systems in the cluster, and maintains cluster integrity when computers join or leave the cluster.

differential SCSI bus

A SCSI bus where the signal's level is determined by the potential difference between two wires.

distributed lock manager

Cluster software component that synchronizes access to shared resources among cooperating processes throughout the cluster.

DLM

See distributed lock manager.

distributed raw disk

A storage technology that uses an ASE service to provide clusterwide access to a disk. The service exports a raw disk to all member systems. The raw disk must be on a shared SCSI bus. If the member system running the DRD service fails, the service can fail over to another member system on the same shared SCSI bus.

DRD

See distributed raw disk.

failover

A transfer of the responsibility to provide an ASE service. A failover occurs when a hardware or software failure causes a service to restart on a viable member system.

fast SCSI

An optional mode of SCSI-2 that allows transmission rates of up to 10 MB per second.

fast bus speed

A bus speed that uses the fast synchronous transfer option, enabling I/O devices to attain high peak-rate transfers (10 MB per second) in synchronous mode.

firmware

Software code stored in hardware.

highly available

In the TruCluster software, the ability to survive any single hardware or software failure.

A cluster can be considered highly available if the hardware and software provides protection against any single failure, such as a system or disk failure or a SCSI cable disconnection.

An ASE service can be considered highly available if the hardware it depends on provides protection against any single failure, and the service is configured to fail over in case of a failure.

hot swap

The ability to replace a device on a shared bus while the bus is active.

hot standby

A member system that is available to run an ASE service if the primary member system running the service fails.

local bus

See private SCSI bus.

lock file

A file that indicates that operations on one or more other files are restricted or prohibited. The presence of the lock file can be used as the indication, or the lock file can contain information describing the nature of the restrictions.

Logical Storage Manager

A disk storage management tool that protects against data loss, improves disk I/O performance, and customizes the disk configuration.

System administrators use LSM to perform disk management functions without disrupting users or applications accessing data on those disks.

In an ASE, you can use LSM to mirror disks across shared SCSI buses. This results in greater data reliability and integrity. You can use a DRD service to make an LSM volume accessible clusterwide.

LSM

See Logical Storage Manager.

LSM disk group

A group of Logical Storage Manager (LSM) disks that share a common configuration. The configuration information for an LSM disk group consists of a set of records describing objects including LSM disks, LSM volumes, LSM plexes, and LSM subdisks that are associated with the LSM disk group. Each LSM disk group has an administrator-assigned name that can be used to reference that LSM disk group.

LSM volume

A Logical Storage Manager (LSM) volume is a DIGITAL UNIX special device that contains data used by a UNIX file system, a database, or other applications. LSM transparently places an LSM volume between applications and a physical disk. Applications then operate on the LSM volume rather than on the physical disk. For example, a file system is created on an LSM volume rather than on a physical disk.

An LSM volume presents block and raw interfaces that are compatible in their use with disk partition special devices. Because an LSM volume is a virtual device, it can be mirrored, spanned across disk drives, moved to use different storage, and striped using administrative commands. The configuration of an LSM volume can be changed using LSM utilities without disrupting applications or file systems that are using the LSM volume.

LSM plex

A Logical Storage Manager (LSM) plex is a copy of an LSM volume's logical data address space, sometimes known as a mirror. An LSM volume can have up to eight LSM plexes associated with it. A read can be satisfied from any LSM plex, while a write is directed to all LSM plexes.

logical unit number

A physical or virtual peripheral device addressable through a target. LUNs use their target's bus connection to communicate on a SCSI bus.

LUN

See logical unit number.

member system

The basic computing resource in a cluster. A member system must be physically connected to a cluster interconnect and at least one shared SCSI bus. The connection manager dynamically determines cluster membership based on communications among the cluster members.

MEMORY CHANNEL

A peripheral component interconnect (PCI)-based cluster interconnect that promotes fast and reliable communications between cluster members.

MEMORY CHANNEL interconnect

MEMORY CHANNEL interconnect. A type of cluster interconnect that consists of a MEMORY CHANNEL adapter installed in a PCI slot in each member system, one or more MEMORY CHANNEL link cables to connect the adapters, and an optional MEMORY CHANNEL hub.

mount point

A directory file that is the name of a mounted file system.

network

Two or more computing systems that are linked for the purpose of exchanging information and sharing resources.

network interface

The network adapter and the software that allows a system to communicate over a network.

partition

An abnormal condition in which nodes in an existing TruCluster software configuration divide into two independent clusters.

peripheral component interconnect

An industry-standard expansion I/O bus that is a synchronous, asymmetrical I/O channel.

PCI

See peripheral component interconnect.

private SCSI bus

A SCSI bus that connects private storage to the local system.

private storage

A storage device on a private SCSI bus. Storage devices include hard disks, floppy disks, and compact disk drives, tape drives, and other devices.

redundant array of inexpensive disks

A technique that organizes disk data to improve performance and reliability. RAID has three attributes:

- It is a set of physical disks viewed by the user as a single logical device or multiple logical devices.
- Disk data is distributed across the physical set of drives in a defined manner.
- Redundant disk capacity is added so data can be recovered if a drive fails.

RAID

See redundant array of inexpensive disks.

redundant

Describes duplicate hardware that provides spare capacity that can be used when a component fails.

relocate a service

To stop an ASE service on one member system and restart it on another member system.

relocation policy

See ASP policy.

script

A program to be interpreted and executed by the shell.

SCSI

See Small Computer System Interface.

SCSI-2

An extension to the original SCSI standard featuring multiple systems on the same bus and hot swap. Hot swap is the ability to replace a device on a shared bus while the bus is active. The SCSI-2 standard is ANSI standard X3.T9.2/86-109.

SCSI adapter

A storage adapter that provides a connection between an I/O bus and a SCSI bus.

SCSI bus

A bus that supports the transmission and signalling requirements of a SCSI protocol. See shared SCSI bus and private SCSI bus.

SCSI bus speed

The data transfer speed for a SCSI bus. SCSI bus speed can be either slow, up to 5 million bytes per second, or fast, up to 10 million bytes per second.

SCSI controller

An adapter or module that is installed in a member system's I/O bus slot that provides a connection to a shared SCSI bus.

SCSI device

A SCSI controller, peripheral controller, or intelligent peripheral that can be attached to a SCSI bus.

SCSI ID

Unique address that identifies a device on a SCSI bus.

server

A computing system that provides a specific set of applications or data to clients. For a service in an ASE, the server is the member system that is currently running the service.

service

See ASE service.

shared SCSI bus

A SCSI bus that is connected to more than one member system and, optionally, one or more storage devices.

shared storage

Disks that are connected to a shared SCSI bus.

signal converter

Converts signals between a single-ended SCSI bus and a differential SCSI bus.

single-ended SCSI bus

A signal path in which one data lead and one ground lead are utilized to make a device connection. This transmission method is economical, but is more susceptible to noise than a differential SCSI bus.

Small Computer System Interface

An American National Standards Institute (ANSI) standard interface for connecting disks and other peripheral devices to a computer system. SCSI-based devices can be configured in a series, with multiple devices on the same bus. In this manual, SCSI refers to SCSI-2. SCSI is pronounced *skuh-zee*.

SRM

External interface to console firmware for operating systems that expect firmware compliance with the Alpha System Reference Manual (SRM).

standard mode

A MEMORY CHANNEL interconnect configuration that uses a MEMORY CHANNEL hub to connect MEMORY CHANNEL adapters. To set up a MEMORY CHANNEL interconnect in standard mode, use a link cable to connect each MEMORY CHANNEL adapter to a linecard installed in a MEMORY CHANNEL hub.

storage availability domain

A collection of nodes that can access commonly shared storage devices in an available server environment (ASE).

StorageWorks

The DIGITAL modular storage subsystem (MSS), which consists of a family of mass storage products that can be configured to meet current and future storage needs.

subset

A software module that can be installed, which is compatible with the DIGITAL UNIX `setld` software installation utility.

system bus

The private (nonshared) interconnect used on the CPU subsystem. This bus connects the processor module, the memory module, and the I/O module.

target

A device that can be addressed by a SCSI ID on a SCSI bus.

terminator

Resistor array device used for terminating a SCSI bus. A SCSI bus must be terminated at its two physical ends.

tie-breaker disk

One to three disks used by the connection manager to prevent cluster partitions in a two-member cluster that does not use a hub.

trilink connector

A connector that joins two cables to a single device.

virtual hub mode

A MEMORY CHANNEL interconnect configuration that does not use a MEMORY CHANNEL hub to connect MEMORY CHANNEL adapters. Virtual hub mode is supported only for clusters that have two member systems. To set up a MEMORY CHANNEL interconnect in virtual hub mode, use a MEMORY

CHANNEL link cable to connect the MEMORY CHANNEL adapter in one member system to the corresponding MEMORY CHANNEL adapter in the other member system.

warm swap

To replace a device on a shared bus while the bus is not active.

Y cable

A cable that joins two cables to a single device.

Index

How to Order Additional Documentation

Technical Support

If you need help deciding which documentation best meets your needs, call 800-DIGITAL (800-344-4825) before placing your electronic, telephone, or direct mail order.

Electronic Orders

To place an order at the Electronic Store, dial 800-234-1998 using a modem from anywhere in the USA, Canada, or Puerto Rico. If you need assistance using the Electronic Store, call 800-DIGITAL (800-344-4825).

Telephone and Direct Mail Orders

Your Location	Call	Contact
Continental USA, Alaska, or Hawaii	800-DIGITAL	Digital Equipment Corporation P.O. Box CS2008 Nashua, New Hampshire 03061
Puerto Rico	809-754-7575	Local Digital subsidiary
Canada	800-267-6215	Digital Equipment of Canada Attn: DECdirect Operations KAO2/2 P.O. Box 13000 100 Herzberg Road Kanata, Ontario, Canada K2K 2A6
International	—	Local Digital subsidiary or approved distributor
Internal (submit an Internal Software Order Form, EN-01740-07)	—	SSB Order Processing – NQO/V19 *or* U.S. Software Supply Business Digital Equipment Corporation 10 Cotton Road Nashua, NH 03063-1260

Reader’s Comments

TruCluster Software Products
Administration
AA-R88JA-TE

Digital welcomes your comments and suggestions on this manual. Your input will help us to write documentation that meets your needs. Please send your suggestions using one of the following methods:

- This postage-paid form
- Internet electronic mail: `readers_comment@zk3.dec.com`
- Fax: (603) 884-0120, Attn: UBPG Publications, ZKO3-3/Y32

If you are not using this form, please be sure you include the name of the document, the page number, and the product name and version.

Please rate this manual:

	Excellent	Good	Fair	Poor
Accuracy (software works as manual says)	□	□	□	□
Clarity (easy to understand)	□	□	□	□
Organization (structure of subject matter)	□	□	□	□
Figures (useful)	□	□	□	□
Examples (useful)	□	□	□	□
Index (ability to find topic)	□	□	□	□
Usability (ability to access information quickly)	□	□	□	□

Please list errors you have found in this manual:

Page ________ Description ____

________ ____

________ ____

________ ____

________ ____

Additional comments or suggestions to improve this manual:

____

____

____

____

What version of the software described by this manual are you using? ____________________

Name, title, department ____
Mailing address ____
Electronic mail ____
Telephone ____
Date ____

Do Not Cut or Tear − Fold Here and Tape

digital™

NO POSTAGE
NECESSARY IF
MAILED IN THE
UNITED STATES

BUSINESS REPLY MAIL
FIRST−CLASS MAIL PERMIT NO. 33 MAYNARD MA

POSTAGE WILL BE PAID BY ADDRESSEE

DIGITAL EQUIPMENT CORPORATION
UEG PUBLICATIONS MANAGER
ZKO3−3/Y32
110 SPIT BROOK RD
NASHUA NH 03062−9987

Do Not Cut or Tear − Fold Here

Cut on
Dotted
Line