Sun Cluster Overview for Solaris OS

Sun Microsystems, Inc.
4150 Network Circle
Santa Clara, CA 95054
U.S.A.

Part No: 821–0518–10
November 2009, Revision A

091123@23031

Contents

Preface

Sun™ Cluster Overview for Solaris OS introduces the Sun Cluster product by explaining the purpose of the product and how Sun Cluster achieves this purpose. This book also explains key concepts for Sun Cluster. The information in this document enables you to become familiar with Sun Cluster features and functionality.

Related Documentation

Information about related Sun Cluster topics is available in the documentation that is listed in the following table. All Sun Cluster documentation is available at http://docs.sun.com.

Topic	Documentation
Overview	*Sun Cluster Overview for Solaris OS* *Sun Cluster 3.2 11/09 Documentation Center*
Concepts	*Sun Cluster Concepts Guide for Solaris OS*
Hardware installation and administration	*Sun Cluster 3.1 - 3.2 Hardware Administration Manual for Solaris OS* Individual hardware administration guides
Software installation	*Sun Cluster Software Installation Guide for Solaris OS* *Sun Cluster Quick Start Guide for Solaris OS*
Data service installation and administration	*Sun Cluster Data Services Planning and Administration Guide for Solaris OS* Individual data service guides
Data service development	*Sun Cluster Data Services Developer's Guide for Solaris OS*
System administration	*Sun Cluster System Administration Guide for Solaris OS* *Sun Cluster Quick Reference*
Software upgrade	*Sun Cluster Upgrade Guide for Solaris OS*
Error messages	*Sun Cluster Error Messages Guide for Solaris OS*

Topic	Documentation
Command and function references	*Sun Cluster Reference Manual for Solaris OS* *Sun Cluster Data Services Reference Manual for Solaris OS* *Sun Cluster Quorum Server Reference Manual for Solaris OS*

For a complete list of Sun Cluster documentation, see the release notes for your release of Sun Cluster software at http://wikis.sun.com/display/SunCluster/Home/.

Documentation, Support, and Training

The Sun web site provides information about the following additional resources:

- Documentation (`http://www.sun.com/documentation/`)
- Support (`http://www.sun.com/support/`)
- Training (`http://www.sun.com/training/`)

Sun Welcomes Your Comments

Sun is interested in improving its documentation and welcomes your comments and suggestions. To share your comments, go to `http://docs.sun.com` and click Feedback.

Getting Help

If you have problems installing or using the Sun Cluster system, contact your service provider and provide the following information:

- Your name and email address (if available)
- Your company name, address, and phone number
- The model and serial numbers of your systems
- The release number of the operating environment (for example, Solaris 9)
- The release number of the Sun Cluster software (for example, 3.2 11/09)

Use the following commands to gather information about each Solaris host on your system for your service provider.

Command	Function
`prtconf -v`	Displays the size of the system memory and reports information about peripheral devices
`psrinfo -v`	Displays information about processors
`showrev -p`	Reports which patches are installed
`prtdiag -v`	Displays system diagnostic information
`scinstall -pv`	Displays Sun Cluster software release and package version information
`scstat`	Provides a snapshot of the cluster status
`scconf -p`	Lists cluster configuration information
`scrgadm -p`	Displays information about installed resources, resource groups, and resource types

Also have available the contents of the `/var/adm/messages` file.

Typographic Conventions

The following table describes the typographic conventions that are used in this book.

TABLE P–1 Typographic Conventions

Typeface	Meaning	Example
`AaBbCc123`	The names of commands, files, and directories, and onscreen computer output	Edit your `.login` file. Use `ls -a` to list all files. `machine_name% you have mail.`

`AaBbCc123`	What you type, contrasted with onscreen computer output	`machine_name%` **`su`**
		`Password:`

aabbcc123	Placeholder: replace with a real name or value	The command to remove a file is `rm` *filename*.
AaBbCc123	Book titles, new terms, and terms to be emphasized	Read Chapter 6 in the *User's Guide*. A *cache* is a copy that is stored locally. Do *not* save the file. **Note:** Some emphasized items appear bold online.

Shell Prompts in Command Examples

The following table shows the default UNIX® system prompt and superuser prompt for the C shell, Bourne shell, and Korn shell.

TABLE P–2 Shell Prompts

Shell	Prompt
C shell	`machine_name%`
C shell for superuser	`machine_name#`
Bourne shell and Korn shell	`$`
Bourne shell and Korn shell for superuser	`#`

◆ ◆ ◆ CHAPTER 1

Introduction to Sun Cluster

A Sun Cluster configuration is an integrated hardware and Sun Cluster software solution that is used to create highly available and scalable services. This chapter provides a high-level overview of Sun Cluster features.

This chapter contains the following sections:

- "Making Applications Highly Available With Sun Cluster" on page 9
- "Monitoring Failure" on page 13
- "Administration and Configuration Tools" on page 14

Making Applications Highly Available With Sun Cluster

A cluster is a collection of loosely coupled computing nodes that provides a single client view of network services or applications, including databases, web services, and file services.

In a clustered environment, the nodes are connected by an interconnect and work together as a single entity to provide increased availability and performance.

Highly available clusters provide nearly continuous access to data and applications by keeping the cluster running through failures that would normally bring down a single server system. No single failure, hardware, software, or network, can cause a cluster to fail. By contrast, fault-tolerant hardware systems provide constant access to data and applications, but at a higher cost because of specialized hardware. Fault-tolerant systems usually have no provision for software failures.

Each Sun Cluster system is a collection of tightly coupled nodes that provide a single administration view of network services and applications. The Sun Cluster system achieves high availability through a combination of the following hardware and software:

- Redundant disk systems provide storage. These disk systems are generally mirrored to permit uninterrupted operation if a disk or subsystem fails. Redundant connections to the disk systems ensures that data is not isolated if a server, controller, or cable fails. A high-speed interconnect among Solaris hosts provides access to resources. All hosts in the cluster are also connected to a public network, enabling clients on multiple networks to access the cluster.
- Redundant hot-swappable components, such as power supplies and cooling systems, improve availability by enabling systems to continue operation after a hardware failure. Hot-swappable components provide the ability to add or remove hardware components in a functioning system without bringing it down.
- Sun Cluster software's high-availability framework detects a node failure quickly and migrates the application or service to another node that runs in an identical environment. At no time are all applications unavailable. Applications unaffected by a down node are fully available during recovery. Furthermore, applications of the failed node become available as soon as they are recovered. A recovered application does not have to wait for all other applications to complete their recovery.

Availability Management

An application is highly available if it survives any single software or hardware failure in the system. Failures that are caused by bugs or data corruption within the application itself are excluded. The following apply to highly available applications:

- Recovery is transparent from the applications that use a resource.
- Resource access is fully preserved across node failure.
- Applications cannot detect that the hosting node has been moved to another node.
- Failure of a single node is completely transparent to programs on remaining nodes that use the files, devices, and disk volumes that are attached to this node.

Failover and Scalable Services and Parallel Applications

Failover and scalable services and parallel applications enable you to make your applications highly available and to improve an application's performance on a cluster.

A failover service provides high availability through redundancy. When a failure occurs, you can configure an application that is running to either restart on the same node, or be moved to another node in the cluster, without user intervention.

To increase performance, a scalable service leverages the multiple nodes in a cluster to concurrently run an application. In a scalable configuration, each node in the cluster can provide data and process client requests.

Parallel databases enable multiple instances of the database server to do the following:

- Participate in the cluster
- Handle different queries on the same database simultaneously
- Provide parallel query capability on large queries

For more information about failover and scalable services and parallel applications, see "Data Service Types" on page 28.

IP Network Multipathing

Clients make data requests to the cluster through the public network. Each Solaris host is connected to at least one public network through one or multiple public network adapters.

IP network multipathing enables a server to have multiple network ports connected to the same subnet. First, IP network multipathing software provides resilience from network adapter failure by detecting the failure or repair of a network adapter. The software then simultaneously switches the network address to and from the alternative adapter. When more than one network adapter is functional, IP network multipathing increases data throughput by spreading outbound packets across adapters.

Storage Management

Multihost storage makes disks highly available by connecting the disks to multiple Solaris hosts. Multiple hosts enable multiple paths to access the data. If one path fails, another one is available to take its place.

Multihost disks enable the following cluster processes:

- Tolerating single-host failures.
- Centralizing application data, application binaries, and configuration files.
- Protecting against host failures. If client requests are accessing the data through a host that fails, the requests are switched over to use another host that has a direct connection to the same disks.
- Providing access either globally through a primary host that "masters" the disks, or by direct concurrent access through local paths.

Volume Management Support

A volume manager enables you to manage large numbers of disks and the data on those disks. Volume managers can increase storage capacity and data availability by offering the following features:

- Disk-drive striping and concatenation
- Disk-mirroring
- Disk-drive hot spares
- Disk-failure handling and disk replacements

Sun Cluster systems support the following volume managers:

- Solaris Volume Manager
- Multi-owner Solaris Volume Manager for Sun Cluster
- Veritas Volume Manager

Solaris I/O Multipathing (MPxIO)

Solaris I/O multipathing (MPxIO), which was formerly named Sun StorEdge Traffic Manager, is fully integrated in the Solaris Operating System I/O framework. Solaris I/O multipathing enables you to represent and manage devices that are accessible through multiple I/O controller interfaces within a single instance of the Solaris operating system.

The Solaris I/O multipathing architecture provides the following features:

- Protection against I/O outages due to I/O controller failures
- Automatic switches to an alternate controller upon an I/O controller failure
- Increased I/O performance by load balancing across multiple I/O channels

Hardware Redundant Array of Independent Disks Support

Sun Cluster systems support the use of hardware Redundant Array of Independent Disks (RAID) and host-based software RAID. Hardware RAID uses the storage array's or storage system's hardware redundancy to ensure that independent hardware failures do not impact data availability. If you mirror across separate storage arrays, host-based software RAID ensures that independent hardware failures do not impact data availability when an entire storage array is offline. Although you can use hardware RAID and host-based software RAID concurrently, you need only one RAID solution to maintain a high degree of data availability.

Cluster File System Support

Because one of the inherent properties of clustered systems is shared resources, a cluster requires a file system that addresses the need for files to be shared coherently. In a Sun Cluster file system, a *cluster file system* enables users or applications to access any file on any node of the cluster by using remote or local standard UNIX APIs.

Sun Cluster systems support the following cluster file systems:

- UNIX® File System (UFS) – Uses Sun Cluster Proxy System (PxFS)
- Veritas File System (VxFS) – Uses PxFS

Sun Cluster software supports the following as highly available failover local file systems:

- UFS
- Solaris ZFS™
- Sun QFS
- VxFS

If an application is moved from one node to another node, no change is required for the application to access the same files. No changes need to be made to existing applications to fully utilize the cluster file system.

Campus Clusters

Standard Sun Cluster systems provide high availability and reliability from a single location. If your application must remain available after unpredictable disasters such as an earthquake, flood, or power outage, you can configure your cluster as a campus cluster.

Campus clusters enable you to locate cluster components, such as Solaris hosts and shared storage, in separate rooms that are several kilometers apart. You can separate your hosts and shared storage and locate them in different facilities around your corporate campus or elsewhere within several kilometers. When a disaster strikes one location, the surviving hosts can take over service for the failed host. This enables applications and data to remain available for your users. For additional information about campus cluster configurations, see the *Sun Cluster 3.1 - 3.2 Hardware Administration Manual for Solaris OS*.

Monitoring Failure

The Sun Cluster system makes the path between users and data highly available by using multihost disks, multipathing, and a cluster file system. The Sun Cluster system monitors failures for the following:

- Applications – Most of the Sun Cluster data services supply a fault monitor that periodically probes the data service to determine its health. A fault monitor verifies that the application daemon or daemons are running and that clients are being served. Based on the information that is returned by probes, a predefined action such as restarting daemons or causing a failover can be initiated.
- Disk Paths – Sun Cluster software supports disk-path monitoring (DPM). DPM improves the overall reliability of failover and switchover by reporting the failure of a secondary disk path.

- Internet Protocol (IP) Multipath – Solaris IP network multipathing software on Sun Cluster systems provide the basic mechanism for monitoring public network adapters. IP multipathing also enables failover of IP addresses from one adapter to another adapter when a fault is detected.
- Quorum Devices - Sun Cluster software supports quorum device monitoring by periodically testing that quorum works on quorum devices. When Sun Cluster software detects a failure, the Sun Cluster system reports the failure and marks the quorum device that is not working correctly. When the Sun Cluster system detects that a previously failed quorum device now operates correctly, the system automatically brings the quorum device back into service. Bringing the quorum device back into service includes placing the correct quorum reservation information on the device. The Sun Cluster system automatically monitors any configured quorum device that is not in maintenance mode, regardless of type.

Administration and Configuration Tools

You can install, configure, and administer the Sun Cluster system either though the Sun Cluster Manager GUI or through the command-line interface (CLI).

The Sun Cluster system also has a module that runs as part of Sun Management Center software that provides a GUI to certain cluster tasks.

Sun Cluster Manager

Sun Cluster Manager is a browser-based tool for administering Sun Cluster systems. The Sun Cluster Manager software enables administrators to perform system management and monitoring, software installation, and system configuration.

The Sun Cluster Manager software includes the following features:

- Built-in security and authorization mechanisms
- Secure Sockets Layer (SSL) support
- Role-based access control (RBAC)
- Pluggable Authentication Module (PAM)
- NAFO and IP network multipathing group administration facilities
- Quorum devices, transports, shared storage device, and resource group administration
- Sophisticated error checking and autodetection of private interconnects

Command-Line Interface

The Sun Cluster command-line interface (CLI) is a set of utilities you can use to install and administer Sun Cluster systems, and administer the volume manager portion of Sun Cluster software.

You can perform the following Sun Cluster administration tasks through the Sun Cluster CLI:

- Validating a Sun Cluster configuration
- Installing and configuring Sun Cluster software
- Updating a Sun Cluster configuration
- Managing the registration of resource types, the creation of resource groups, and the activation of resources within a resource group
- Changing node mastery and states for resource groups and device groups
- Controlling access with role-based access control (RBAC)
- Shutting down the entire cluster

Sun Management Center

The Sun Cluster system also has a module that runs as part of Sun Management Center software. Sun Management Center software serves as the cluster's base for administrative and monitoring operations and enables system administrators to perform the following tasks through a GUI or CLI:

- Configuring a remote system
- Monitoring performance
- Detecting and isolating hardware and software faults

Sun Management Center software can also be used as the interface to manage dynamic reconfiguration within Sun Cluster servers. Dynamic reconfiguration includes domain creation, dynamic board attach, and dynamic detach.

Role-Based Access Control

In conventional UNIX systems, the root user, also referred to as superuser, is omnipotent, with the ability to read and write to any file, run all programs, and send kill signals to any process. Solaris role-based access control (RBAC) is an alternative to the all-or-nothing superuser model. RBAC uses the security principle of least privilege, which is that no user should be given more privilege than necessary for performing his or her job.

RBAC enables an organization to separate superuser capabilities and package them into special user accounts or roles for assignment to specific individuals. This separation and packaging enables a variety of security policies. Accounts can be set up for special-purpose administrators in such areas as security, networking, firewall, backups, and system operation.

CHAPTER 2

Key Concepts for Sun Cluster

This chapter explains the key concepts related to the hardware and software components of the Sun Cluster system that you need to understand before working with Sun Cluster systems.

This chapter contains the following sections:

Clusters, Nodes, and Hosts

A cluster is a collection of one or more nodes that belong exclusively to that collection. In a cluster that runs on the Solaris 10 OS, a *global cluster* and a *zone cluster* are types of clusters. In a cluster that runs on any version of the Solaris OS that was released before the Solaris 10 OS, a node is a *physical machine* that contributes to cluster membership and is not a quorum device. In a cluster that runs on the Solaris 10 OS, the concept of a node changes. A node is a Solaris *zone* that is associated with a cluster. In this environment, a *Solaris host*, or simply *host*, is one of the following hardware or software configurations that runs the Solaris OS and its own processes:

- A “bare metal” physical machine that is not configured with a virtual machine or as a hardware domain
- A Sun Logical Domains (LDoms) guest domain
- A Sun Logical Domains (LDoms) I/O domain
- A hardware domain

In a Solaris 10 environment, a *voting node* is a zone that contributes votes to the total number of quorum votes, that is, membership votes in a cluster. This total determines whether the cluster has sufficient votes to continue operating. A *non-voting node* is a zone that does *not* contribute to the total number of quorum votes, that is, membership votes in a cluster.

In a clustered environment, the nodes are connected by an interconnect and work together as a single entity to provide increased availability and performance.

In a Solaris 10 environment, a global cluster is a type of cluster that is composed only of one or more global-cluster voting nodes and optionally, zero or more global-cluster non-voting nodes.

Note – A global cluster can optionally also include `solaris8`, `solaris9`, `lx` (Linux), or `native` brand, non-global zones that are not nodes, but high availability containers (as resources).

A global-cluster voting node is a `native` brand, global zone in a global cluster that contributes votes to the total number of quorum votes, that is, membership votes in the cluster. This total determines whether the cluster has sufficient votes to continue operating. A global-cluster non-voting node is a `native` brand, non-global zone in a global cluster that does *not* contribute votes to the total number of quorum votes, that is, membership votes in the cluster.

In a Solaris 10 environment, a zone cluster is a type of cluster that is composed only of one or more `cluster` brand, voting nodes. A zone cluster depends on, and therefore requires, a global cluster. A global cluster does not contain a zone cluster. You cannot configure a zone cluster without a global cluster. A zone cluster has, at most, one zone cluster node on a machine.

Note – A zone-cluster node continues to operate only as long as the global-cluster voting node on the same machine continues to operate. If a global-cluster voting node on a machine fails, all zone-cluster nodes on that machine fail as well.

The Sun Cluster software enables you to have one to sixteen Solaris hosts in a cluster, depending on the hardware configuration. Contact your Sun representative for information about the number of Solaris hosts that are supported on your particular hardware configuration.

Solaris hosts in a cluster are generally attached to one or more disks. Solaris hosts that are not attached to disks use the cluster file system to access the multihost disks. Solaris hosts in parallel database configurations share concurrent access to some or all disks.

Every node in the cluster is aware when another node joins or leaves the cluster. Also, every node in the cluster is aware of the resources that are running locally as well as the resources that are running on the other cluster nodes.

Solaris hosts in the same cluster should have similar processing, memory, and I/O capability to enable failover to occur without significant degradation in performance. Because of the possibility of failover, each host should have sufficient capacity to meet service level agreements if a node fails.

Zone Cluster

This section describes the primary features and benefits of a zone cluster.

Features and Benefits of a Zone Cluster

A zone cluster provides the following features and benefits.

- Application fault isolation – A failure of applications on one zone cluster does not affect applications on other zone clusters. For example, if a zone cluster node starts, halts, or reboots, nodes in other zone clusters are not affected.
- Security – Applications that are, or a person who is, logged into a zone cluster node cannot see or modify elements in the global cluster or in other zone clusters. A zone cluster only contains those elements, such as file systems, ZFS datasets, or network resources that are explicitly configured as part of that zone cluster. A failover application in a zone cluster can fail over or switch over only from one node in a zone cluster to another node in the same zone cluster. All instances of a scalable application run only in the same zone cluster. The zone cluster is a security container that applications cannot escape.
- Resource management – You can apply the full range of Solaris resource management controls to a zone cluster. Consequently, you can control all applications on a node in a zone cluster at the zone level. This control enables you to better manage the resources that are available to a zone cluster node. For example, this control enables you to place an application in a zone cluster and reduce the number of CPUs. You can thus reduce your per-CPU license fee.
- Delegated administration – You can delegate the ability to manage applications in a zone cluster to an administrator who is operating in that zone cluster. A zone cluster functions independently of the global cluster and other zone clusters. As the global zone administrator, you can set up cross-cluster dependencies and affinities and administer applications in the zone cluster.

- Simplified cluster – All you need to do in a zone cluster is administer the applications and resources that are used by those applications. As the global-zone administrator, you can create, manage, and remove a zone cluster at any time by issuing a command both inside and outside that zone cluster. You can do so without affecting the global cluster or other zone clusters.

Cluster Interconnect

The cluster interconnect is the physical configuration of devices that are used to transfer cluster-private communications and data service communications between Solaris hosts in the cluster.

Redundant interconnects enable operations to continue over the surviving interconnects while system administrators isolate failures and repair communication. The Sun Cluster software detects, repairs, and automatically re-initiates communication over a repaired interconnect.

For more information, see “Cluster-Interconnect Components” on page 39.

Cluster Membership

The Cluster Membership Monitor (CMM) is a distributed set of agents that exchange messages over the cluster interconnect to complete the following tasks:

- Enforcing a consistent membership view on all nodes (quorum)
- Driving synchronized reconfiguration in response to membership changes
- Handling cluster partitioning
- Ensuring full connectivity among all cluster members by leaving unhealthy nodes out of the cluster until they are repaired

The main function of the CMM is to establish cluster membership, which requires a cluster-wide agreement on the set of nodes that participate in the cluster at any time. The CMM detects major cluster status changes on each node, such as loss of communication between one or more nodes. The CMM relies on the transport kernel module to generate heartbeats across the transport medium to other nodes in the cluster. When the CMM does not detect a heartbeat from a node within a defined timeout period, the CMM considers the node to have failed and the CMM initiates a cluster reconfiguration to renegotiate cluster membership.

To determine cluster membership and to ensure data integrity, the CMM performs the following tasks:

- Accounting for a change in cluster membership, such as a node's joining or leaving the cluster
- Ensuring that an unhealthy node leaves the cluster

- Ensuring that an unhealthy node remains inactive until it is repaired
- Preventing the cluster from partitioning itself into subsets of nodes

See "Split Brain and Amnesia" on page 23 for more information about how the cluster protects itself from partitioning into multiple separate clusters.

Cluster Configuration Repository

The Cluster Configuration Repository (CCR) is a private, cluster-wide, distributed database for storing information that pertains to the configuration and state of the cluster. To avoid corrupting configuration data, each node must be aware of the current state of the cluster resources. The CCR ensures that all nodes have a consistent view of the cluster. The CCR is updated when error or recovery situations occur or when the general status of the cluster changes.

The CCR structures contain the following types of information:

- Cluster and node names
- Cluster transport configuration
- The names of Solaris Volume Manager disk sets or Veritas disk groups
- A list of nodes that can master each disk group
- Operational parameter values for data services
- Paths to data service callback methods
- DID device configuration
- Current cluster status

Quorum Devices

A quorum device is a shared storage device or quorum server that is shared by two or more nodes and that contributes votes that are used to establish a quorum. The cluster can operate only when a quorum of votes is available. The quorum device is used when a cluster becomes partitioned into separate sets of nodes to establish which set of nodes constitutes the new cluster.

Both cluster nodes and quorum devices vote to form quorum. By default, cluster nodes acquire a quorum vote count of one when they boot and become cluster members. Nodes can have a vote count of zero when the node is being installed, or when an administrator has placed a node into the maintenance state.

Quorum devices acquire quorum vote counts that are based on the number of node connections to the device. When you set up a quorum device, it acquires a maximum vote count of N-1 where N is the number of connected votes to the quorum device. For example, a quorum device that is connected to two nodes with nonzero vote counts has a quorum count of one (two minus one).

Fault Monitors

Sun Cluster system makes all components on the ”path” between users and data highly available by monitoring the applications themselves, the file system, and network interfaces.

The Sun Cluster software detects a node failure quickly and creates an equivalent server for the resources on the failed node. The Sun Cluster software ensures that resources unaffected by the failed node are constantly available during the recovery and that resources of the failed node become available as soon as they are recovered.

Data Services Monitoring

Each Sun Cluster data service supplies a fault monitor that periodically probes the data service to determine its health. A fault monitor verifies that the application daemon or daemons are running and that clients are being served. Based on the information returned by probes, predefined actions such as restarting daemons or causing a failover, can be initiated.

Disk-Path Monitoring

Sun Cluster software supports disk-path monitoring (DPM). DPM improves the overall reliability of failover and switchover by reporting the failure of a secondary disk-path. You can use one of two methods for monitoring disk paths. The first method is provided by the `cldevice` command. This command enables you to monitor, unmonitor, or display the status of disk paths in your cluster. See the `cldevice`(1CL) man page for more information about command-line options.

The second method for monitoring disk paths in your cluster is provided by the Sun Cluster Manager graphical user interface (GUI). Sun Cluster Manager provides a topological view of the monitored disk paths. The view is updated every 10 minutes to provide information about the number of failed pings.

IP Multipath Monitoring

Each Solaris host in a cluster has its own IP network multipathing configuration, which can differ from the configuration on other hosts in the cluster. IP network multipathing monitors the following network communication failures:

- The transmit and receive path of the network adapter has stopped transmitting packets.
- The attachment of the network adapter to the link is down.
- The port on the switch does not transmit or receive packets.
- The physical interface in a group is not present at system boot.

Quorum Device Monitoring

Sun Cluster software supports the monitoring of quorum devices. Periodically, each node in the cluster tests the ability of the local node to work correctly with each configured quorum device that has a configured path to the local node and is not in maintenance mode. This test consists of an attempt to read the quorum keys on the quorum device.

When the Sun Cluster system discovers that a formerly healthy quorum device has failed, the system automatically marks the quorum device as unhealthy. When the Sun Cluster system discovers that a formerly unhealthy quorum device is now healthy, the system marks the quorum device as healthy and places the appropriate quorum information on the quorum device.

The Sun Cluster system generates reports when the health status of a quorum device changes. When nodes reconfigure, an unhealthy quorum device cannot contribute votes to membership. Consequently, the cluster might not continue to operate.

Data Integrity

The Sun Cluster system attempts to prevent data corruption and ensure data integrity. Because cluster nodes share data and resources, a cluster must never split into separate partitions that are active at the same time. The CMM guarantees that only one cluster is operational at any time.

Split Brain and Amnesia

Two types of problems can arise from cluster partitions: *split brain* and *amnesia*. Split brain occurs when the cluster interconnect between Solaris hosts is lost and the cluster becomes partitioned into subclusters, and each subcluster believes that it is the only partition. A subcluster that is not aware of the other subclusters could cause a conflict in shared resources, such as duplicate network addresses and data corruption.

Amnesia occurs if all the nodes leave the cluster in staggered groups. An example is a two-node cluster with nodes A and B. If node A goes down, the configuration data in the CCR is updated on node B only, and not node A. If node B goes down at a later time, and if node A is rebooted, node A will be running with old contents of the CCR. This state is called amnesia and might lead to running a cluster with stale configuration information.

You can avoid split brain and amnesia by giving each node one vote and mandating a majority of votes for an operational cluster. A partition with the majority of votes has a quorum and is enabled to operate. This majority vote mechanism works well if more than two nodes are in the cluster. In a two-node cluster, a majority is two. If such a cluster becomes partitioned, an external vote enables a partition to gain quorum. This external vote is provided by a quorum device. A quorum device can be any disk that is shared between the two nodes.

Table 2–1 describes how Sun Cluster software uses quorum to avoid split brain and amnesia.

TABLE 2–1 Cluster Quorum, and Split Brain and Amnesia Problems

Partition Type	Quorum Solution
Split brain	Enables only the partition (subcluster) with a majority of votes to run as the cluster (only one partition can exist with such a majority). After a node loses the race for quorum, that node panics.
Amnesia	Guarantees that when a cluster is booted, it has at least one node that was a member of the most recent cluster membership (and thus has the latest configuration data).

Fencing

When split brain occurs, not all nodes can communicate, so individual nodes or subsets of nodes might try to form individual or subset clusters. Each subset or partition might “believe” it has sole access and ownership to the multihost disks. Attempts by multiple nodes to write to the disks can result in data corruption.

If a node loses connectivity with other nodes, the node attempts to form a cluster with the nodes with which communication is possible. If that set of nodes does not form a quorum, Sun Cluster software halts the node and “fences” the node from the disks. Thus, Sun Cluster software prevents the node from accessing the disks. Only current member nodes have access to the disks, ensuring data integrity.

You can turn off fencing for selected disks or for all disks.

Caution – If you turn off fencing under the wrong circumstances, your data can be vulnerable to corruption during application failover. Examine this data corruption possibility carefully when you are considering turning off fencing. If your shared storage device does not support the SCSI protocol, such as a Serial Advanced Technology Attachment (SATA) disk, or if you want to allow access to the cluster's storage from hosts outside the cluster, turn off fencing.

Failfast

The purpose of *failfast* is to halt a component that is not healthy enough to continue correct operations. Sun Cluster software includes many failfast mechanisms that detect different unhealthy conditions.

If the Sun Cluster system detects a critical failure on the global-cluster voting node, the system forcibly shuts down the Solaris host.

When the Sun Cluster system detects a critical failure on any other type of node, for example, a global-cluster non-voting node or zone-cluster node, the system reboots that node.

Sun Cluster software monitors the nodes that belong to the cluster. Communication or node failures can change the number of nodes in a cluster. If the cluster does not maintain sufficient votes, Sun Cluster software halts that set of nodes.

Sun Cluster software maintains a number of critical cluster-specific daemons. Some daemons support the global-cluster voting node, while others support other types of nodes. A daemon is critical to the node that the daemon supports, which might differ from the node on which the daemon runs. For example, some daemons in the global zone support a non-global zone. For this reason, these daemons are critical to the health of the non-global zone rather than the global zone.

Shared Devices, Local Devices, and Device Groups

The cluster file system makes all files across a cluster equally accessible and visible to all nodes. Similarly, Sun Cluster software makes all devices on a cluster accessible and visible throughout the cluster. That is, the I/O subsystem enables access to any device in the cluster, from any node, without regard to where the device is physically attached. This access is referred to as shared device access.

Shared Devices

Sun Cluster systems use shared devices to provide cluster-wide, highly available access to any device in a cluster, from any node.

How Sun Cluster Uses Shared Devices

Generally, if a node fails while providing access to a shared device, the Sun Cluster software switches over to another path to the device and redirects the access to that path. This redirection is easy with shared devices because the same name is used for the device regardless of the path. Access to a remote device is performed in the same way as on a local device that uses the same name. Also, the API to access a shared device on a cluster is the same as the API that is used to access a device locally.

Sun Cluster shared devices include disks, CD-ROMs, and tapes. However, disks are the only multiported shared devices that are supported. This limited support means that CD-ROM and tape devices are not currently highly available devices. The local disks on each server are also not multiported, and thus are not highly available devices.

The DID framework assigns a common (normalized) logical name to each disk, CD-ROM, and tape device in the cluster. This assignment enables consistent access to each device from any node in the cluster.

Device ID

The Sun Cluster software manages shared devices through a construct that is known as the device ID (DID) driver. This driver is used to automatically assign unique IDs to every device in the cluster, including multihost disks, tape drives, and CD-ROMs.

The DID driver is an integral part of the shared device access feature of the cluster. The DID driver probes all nodes of the cluster and builds a list of unique disk devices. The DID driver also assigns each device a unique major and minor number that is consistent on all nodes of the cluster. Access to the shared devices is performed by using the normalized ID logical name, instead of the traditional Solaris logical name.

This approach ensures that any application accessing disks, such as Solaris Volume Manager, uses a consistent path across the cluster. This consistency is especially important for multihost disks, because the local major and minor numbers for each device can vary from node to node. These numbers can change the Solaris device naming conventions as well.

Local Devices

The Sun Cluster software also manages local devices. These devices are accessible only on a Solaris host that is running a service and has a physical connection to the cluster. Local devices can have a performance benefit over shared devices because local devices do not have to replicate state information on multiple hosts simultaneously. The failure of the domain of the device removes access to the device unless the device can be shared by multiple hosts.

Device Groups

Device groups enable volume manager disk groups to become “shared” because they provide multipath and multihost support to the underlying disks. Each cluster Solaris host that is physically attached to the multihost disks provides a path to the device group.

In the Sun Cluster system, you can control multihost disks that are using Sun Cluster software by registering the multihost disks as device groups. This registration provides the Sun Cluster system with information about which nodes have a path to which volume manager disk groups. The Sun Cluster software creates a raw device group for each disk and tape device in the cluster. These cluster device groups remain in an offline state until you access them as shared devices either by mounting a cluster file system or by accessing a raw database file.

Data Services

A data service is the combination of software and configuration files that enables an application to run without modification in a Sun Cluster configuration. When running in a Sun Cluster configuration, an application runs as a resource under the control of the Resource Group Manager (RGM). A data service enables you to configure an application such as Sun Java System Web Server or Oracle database to run on a cluster instead of on a single server.

The software of a data service provides implementations of Sun Cluster management methods that perform the following operations on the application:

- Starting the application
- Stopping the application
- Monitoring faults in the application and recovering from these faults

The configuration files of a data service define the properties of the resource that represents the application to the RGM.

The RGM controls the disposition of the failover and scalable data services in the cluster. The RGM is responsible for starting and stopping the data services on selected nodes of the cluster in response to cluster membership changes. The RGM enables data service applications to utilize the cluster framework.

The RGM controls data services as resources. These implementations are either supplied by Sun or created by a developer who uses a generic data service template, the Data Service Development Library API (DSDL API), or the Resource Management API (RMAPI). The cluster administrator creates and manages resources in containers that are called resource groups. RGM and administrator actions cause resources and resource groups to move between online and offline states.

Description of a Resource Type

A resource type is a collection of properties that describe an application to the cluster. This collection includes information about how the application is to be started, stopped, and monitored on nodes of the cluster. A resource type also includes application-specific properties that need to be defined in order to use the application in the cluster. Sun Cluster data services has several predefined resource types. For example, Sun Cluster HA for Oracle is the resource type `SUNW.oracle-server` and Sun Cluster HA for Apache is the resource type `SUNW.apache`.

Description of a Resource

A resource is an instance of a resource type that is defined cluster wide. The resource type enables multiple instances of an application to be installed on the cluster. When you initialize a resource, the RGM assigns values to application-specific properties and the resource inherits any properties on the resource type level.

Data services utilize several types of resources. Applications such as Apache Web Server or Sun Java System Web Server utilize network addresses (logical hostnames and shared addresses) on which the applications depend. Application and network resources form a basic unit that is managed by the RGM.

Description of a Resource Group

Resources that are managed by the RGM are placed into resource groups so that they can be managed as a unit. A resource group is a set of related or interdependent resources. For example, a resource derived from a `SUNW.LogicalHostname` resource type might be placed in the same resource group as a resource derived from an Oracle database resource type. A resource group migrates as a unit if a failover or switchover is initiated on the resource group.

Data Service Types

Data services enable applications to become highly available and scalable services help prevent significant application interruption after any single failure within the cluster.

When you configure a data service, you must configure the data service as one of the following data service types:

- Failover data service
- Scalable data service
- Parallel data service

Description of a Failover Data Service

Failover is the process by which the cluster automatically relocates an application from a failed primary node to a designated redundant secondary node. Failover applications have the following characteristics:

- Capable of running on only one node of the cluster
- Not cluster-aware
- Dependent on the cluster framework for high availability

If the fault monitor detects an error, it either attempts to restart the instance on the same node, or to start the instance on another node (failover), depending on how the data service has been configured. Failover services use a failover resource group, which is a container for application instance resources and network resources (logical hostnames). Logical hostnames are IP addresses that can be configured up on one node, and later, automatically configured down on the original node and configured up on another node.

Clients might have a brief interruption in service and might need to reconnect after the failover has finished. However, clients are not aware of the change in the physical server that is providing the service.

Description of a Scalable Data Service

The scalable data service enables application instances to run on multiple nodes simultaneously. Scalable services use two resource groups. The scalable resource group contains the application resources and the failover resource group contains the network resources (shared addresses) on which the scalable service depends. The scalable resource group can be online on multiple nodes, so multiple instances of the service can be running simultaneously. The failover resource group that hosts the shared address is online on only one node at a time. All nodes that host a scalable service use the same shared address to host the service.

The cluster receives service requests through a single network interface (the global interface). These requests are distributed to the nodes, based on one of several predefined algorithms that are set by the load-balancing policy. The cluster can use the load-balancing policy to balance the service load between several nodes.

Description of a Parallel Application

Sun Cluster systems provide an environment that shares parallel execution of applications across all the nodes of the cluster by using parallel databases. Sun Cluster Support for Oracle Real Application Clusters is a set of packages that, when installed, enables Oracle Real Application Clusters to run on Sun Cluster nodes. This data service also enables Sun Cluster Support for Oracle Real Application Clusters to be managed by using Sun Cluster commands.

A parallel application has been instrumented to run in a cluster environment so that the application can be mastered by two or more nodes simultaneously. In an Oracle Real Application Clusters environment, multiple Oracle instances cooperate to provide access to the same shared database. The Oracle clients can use any of the instances to access the database. Thus, if one or more instances have failed, clients can connect to a surviving instance and continue to access the database.

System Resource Usage

System resources concern aspects of CPU usage, memory usage, swap usage, and disk and network throughput.

Sun Cluster software enables you to monitor how much of a specific system resource is being used by an *object type* such as a node, disk, network interface, Sun Cluster resource group, or Solaris zone. Monitor system resource usage can be part of your resource management policy. Sun Cluster also enables you to control the CPU assigned to a resource group and to control the size of the processor set a resource group runs in.

System Resource Monitoring

By monitoring system resource usage through Sun Cluster software, you can collect data that reflects how a service using specific system resources is performing and you can discover resource bottlenecks or overload and so preempt problems and more efficiently manage workloads. Data about system resource usage can help you determine what hardware resources are under utilized and what applications are using a lot of resources. Based on this data you can assign applications to nodes that have the necessary resources and choose the node to which to fail over. This consolidation can help you optimize the way that you use your hardware and software resources.

If you consider a particular data value to be critical for a system resource, you can set a *threshold* for this value. When setting a threshold, you also choose how critical this threshold is by assigning it a severity level. If the threshold is crossed, Sun Cluster changes the severity level of the threshold to the severity level you choose. For more information about configuring data collection and threshold, see Chapter 10, "Configuring Control of CPU Usage," in *Sun Cluster System Administration Guide for Solaris OS*.

CPU Control

Each application and service running on a cluster has specific CPU needs. Table 2–2 lists the CPU control activities available on different versions of the Solaris Operating System.

TABLE 2–2 CPU Control

Solaris Version	Zone	Control
Solaris 9 Operating System	N/A	Assign CPU shares
Solaris 10 Operating System	Global	Assign CPU shares
Solaris 10 Operating System	Non-global	Assign CPU shares Assign number of CPU Create dedicated processor sets

Note – The Fair Share Scheduler must be the default scheduler on the cluster if you want to apply CPU shares.

Controlling the CPU assigned to a resource group in a dedicated processor set in a non-global zone offers you the strictest level of control of CPU because if you reserve CPU for a resource group, this CPU is not available to other resource groups. For information about configuring CPU control, see Chapter 10, "Configuring Control of CPU Usage," in *Sun Cluster System Administration Guide for Solaris OS*.

Visualization of System Resource Usage

You can visualize system resource data and the CPU attribution in two ways, by using the command line or through the Sun Cluster Manager graphical user interface. The output from the command is a tabular representation of the monitoring data you request. Through the Sun Cluster Manager, you can visualize data in graphical form. The system resources that you choose to monitor determine the data you can visualize.

CHAPTER 3

Sun Cluster Architecture

Sun Cluster architecture permits a group of systems to be deployed, managed, and viewed as a single, large system.

This chapter contains the following sections:

Sun Cluster Hardware Environment

The following hardware components make up a cluster:

- Solaris hosts that are connected to local disks (unshared) that provide the main computing platform of the cluster.
- Multihost storage provides disks that are shared between Solaris hosts.
- Removable media are configured as shared devices, such as tapes and CD-ROM.
- Cluster interconnect provides a channel for internode communication.
- Public network interfaces enable the network interfaces that are used by client systems to access data services on the cluster.

Figure 3–1 illustrates how the hardware components work with each other.

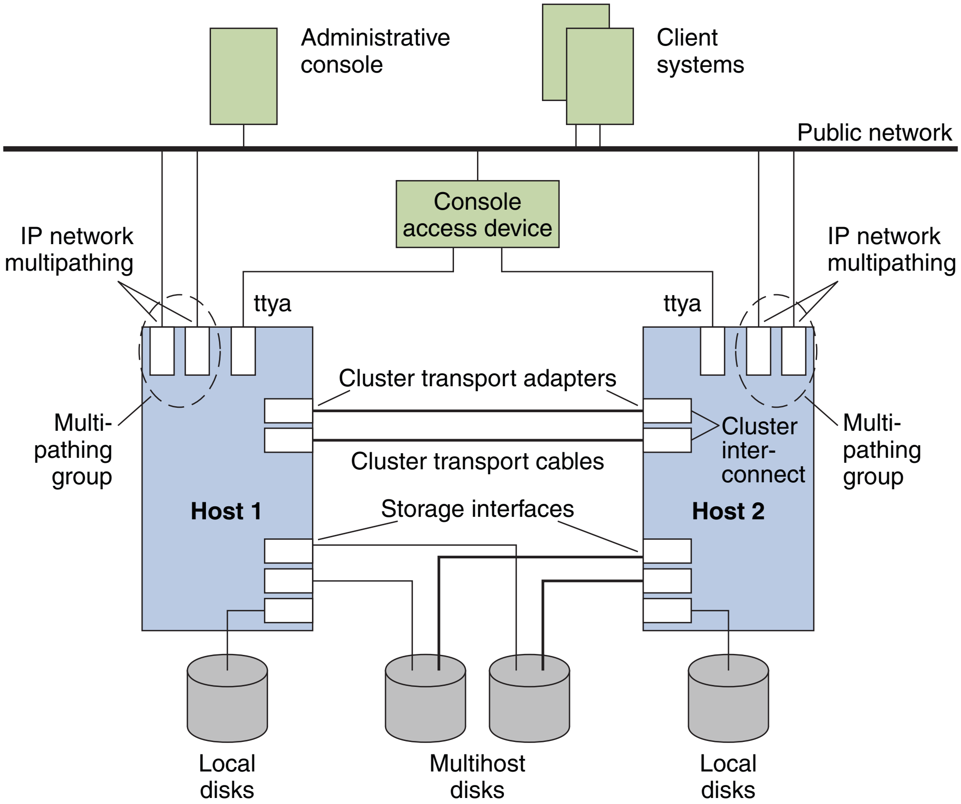

FIGURE 3–1 Sun Cluster Hardware Components

Sun Cluster Software Environment

To function as a cluster member, a Solaris host must have the following software installed:

- Solaris software
- Sun Cluster software
- Data service application
- Volume management (Solaris^{TM} Volume Manager or Veritas Volume Manager)

  An exception is a configuration that uses volume management on the box. This configuration might not require a software volume manager.

Figure 3–2 shows a high-level view of the software components that work together to create the Sun Cluster software environment.

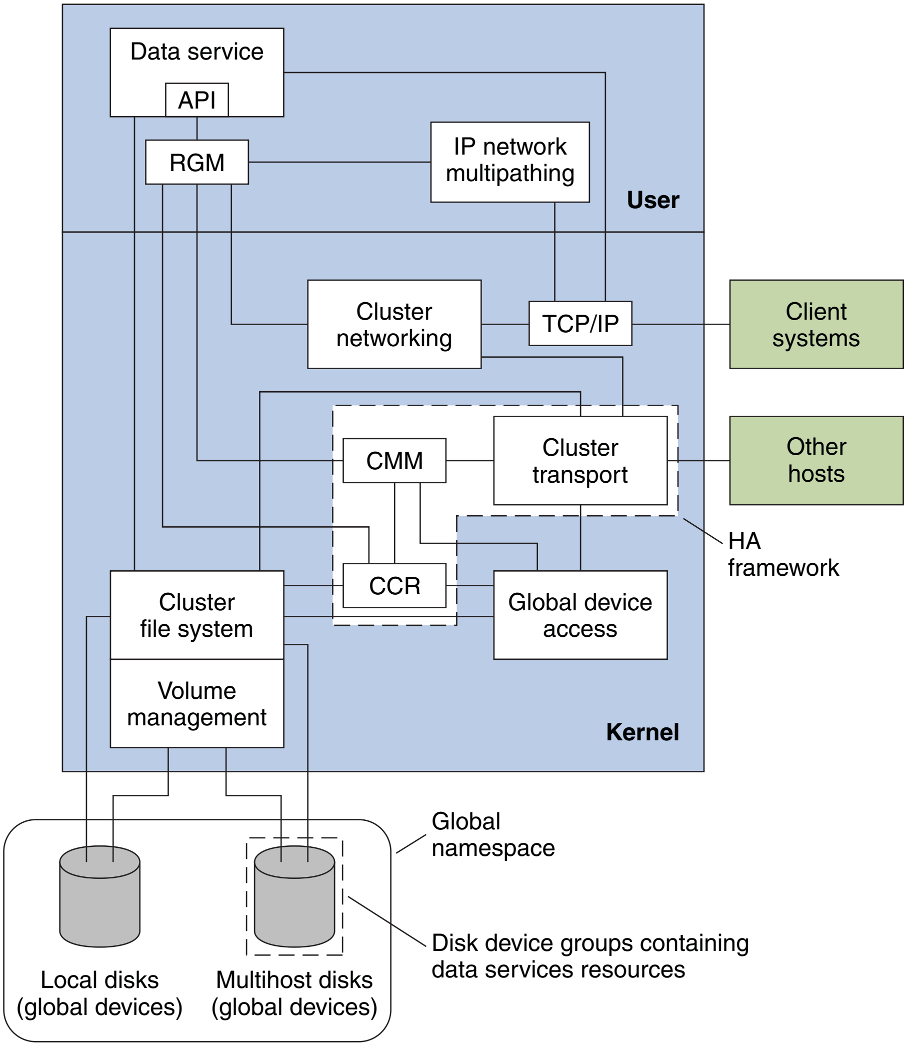

FIGURE 3–2 Sun Cluster Software Architecture

Cluster Membership Monitor

To ensure that data is safe from corruption, all nodes must reach a consistent agreement on the cluster membership. When necessary, the CMM coordinates a cluster reconfiguration of cluster services in response to a failure.

The CMM receives information about connectivity to other nodes from the cluster transport layer. The CMM uses the cluster interconnect to exchange state information during a reconfiguration.

After detecting a change in cluster membership, the CMM performs a synchronized configuration of the cluster. In this configuration, cluster resources might be redistributed, based on the new membership of the cluster.

The CMM runs entirely in the kernel.

Cluster Configuration Repository (CCR)

The CCR relies on the CMM to guarantee that a cluster is running only when quorum is established. The CCR is responsible for verifying data consistency across the cluster, performing recovery as necessary, and facilitating updates to the data.

Cluster File Systems

A cluster file system is a proxy between the following:

- The kernel on one Solaris host and the underlying file system
- The volume manager that is running on a Solaris host that has a physical connection to the disk or disks

Cluster file systems are dependent on shared devices (disks, tapes, CD-ROMs). The shared devices can be accessed from any Solaris host in the cluster through the same file name (for example, `/dev/global/`). That host does not need a physical connection to the storage device. You can use a shared device the same as a regular device, that is, you can create a file system on a shared device by using `newfs` or `mkfs`.

The cluster file system has the following features:

- File access locations are transparent. A process can open a file that is located anywhere in the system. Also, processes on all hosts can use the same path name to locate a file.

  Note – When the cluster file system reads files, it does not update the access time on those files.

- Coherency protocols are used to preserve the UNIX file access semantics even if the file is accessed concurrently from multiple hosts.
- Extensive caching is used with zero-copy bulk I/O movement to move file data efficiently.
- The cluster file system provides highly available advisory file-locking functionality by using the `fcntl(2)` interfaces. Applications that run on multiple cluster hosts can synchronize access to data by using advisory file locking on a cluster file system file. File locks are recovered immediately from nodes that leave the cluster, and from applications that fail while holding locks.

- Continuous access to data is ensured, even when failures occur. Applications are not affected by failures if a path to disks is still operational. This guarantee is maintained for raw disk access and all file system operations.
- Cluster file systems are independent from the underlying file system and volume management software. Cluster file systems make any supported on-disk file system global.

Scalable Data Services

The primary goal of cluster networking is to provide scalability for data services. Scalability means that as the load offered to a service increases, a data service can maintain a constant response time to this increased workload as new nodes are added to the cluster and new server instances are run. A good example of a scalable data service is a web service. Typically, a scalable data service is composed of several instances, each of which runs on different nodes of the cluster. Together, these instances behave as a single service for a remote client of that service and implement the functionality of the service. A scalable web service with several `httpd` daemons that run on different nodes can have any daemon serve a client request. The daemon that serves the request depends on a *load-balancing policy*. The reply to the client appears to come from the service, not the particular daemon that serviced the request, thus preserving the single-service appearance.

The following figure depicts the scalable service architecture.

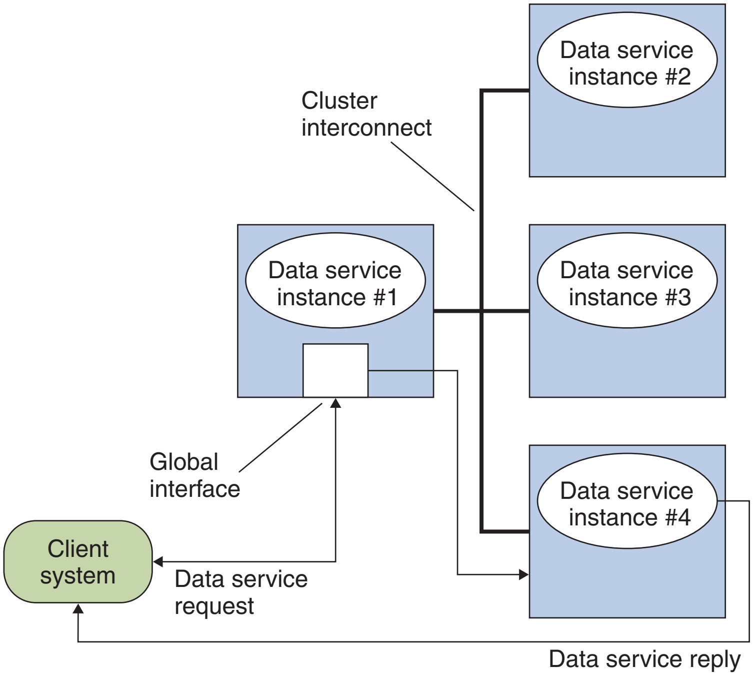

FIGURE 3–3 Scalable Data Service Architecture

The nodes that are not hosting the global interface (proxy nodes) have the shared address hosted on their loopback interfaces. Packets that are coming into the global interface are distributed to other cluster nodes, based on configurable load-balancing policies. The possible load-balancing policies are described next.

Load-Balancing Policies

Load balancing improves performance of the scalable service, both in response time and in throughput.

Two classes of scalable data services exist: *pure* and *sticky*. A pure service is one where any instance can respond to client requests. A sticky service has the cluster balancing the load for requests to the node. Those requests are not redirected to other instances.

A pure service uses a weighted load-balancing policy. Under this load-balancing policy, client requests are by default uniformly distributed over the server instances in the cluster. For example, in a three-node cluster where each node has the weight of 1, each node services one-third of the requests from any client on behalf of that service. Weights can be changed at any time through the `clresource(1cl)` command interface or through the Sun Cluster Manager GUI.

A sticky service has two types: *ordinary sticky* and *wildcard sticky*. Sticky services allow concurrent application-level sessions over multiple TCP connections to share in-state memory (application session state).

Ordinary sticky services permit a client to share state between multiple concurrent TCP connections. The client is said to be “sticky” toward the server instance listening on a single port. The client is guaranteed that all requests go to the same server instance, if that instance remains up and accessible and the load balancing policy is not changed while the service is online.

Wildcard sticky services use dynamically assigned port numbers, but still expect client requests to go to the same node. The client is “sticky wildcard” over ports toward the same IP address.

Multihost Disk Storage

Sun Cluster software makes disks highly available by utilizing multihost disk storage, which can be connected to more than one node at a time. Volume management software can be used to arrange these disks into shared storage that is mastered by a cluster node. The disks are then configured to move to another node if a failure occurs. The use of multi-hosted disks in Sun Cluster systems provides a variety of benefits, including the following:

- Global access to file systems
- Multiple access paths to file systems and data
- Tolerance for single-node failures

Cluster-Interconnect Components

You can set up from one to six cluster interconnects in a cluster. While a single cluster interconnect reduces the number of adapter ports that is used for the private interconnect, it provides no redundancy and less availability. If a single interconnect fails, moreover, the cluster spends more time performing automatic recovery. Two or more cluster interconnects provide redundancy and scalability, and therefore higher availability, by avoiding a single point of failure.

The Sun Cluster interconnect uses Fast Ethernet, Gigabit-Ethernet, InfiniBand, or the Scalable Coherent Interface (SCI, IEEE 1596-1992), enabling high-performance cluster-private communications.

In clustered environments, high-speed, low-latency interconnects and protocols for internode communications are essential. The SCI interconnect in Sun Cluster systems offers improved performance over typical network interface cards (NICs).

The RSM Reliable Datagram Transport (RSMRDT) driver consists of a driver that is built on top of the RSM API and a library that exports the RSMRDT-API interface. The driver provides enhanced Oracle Real Application Clusters performance. The driver also enhances load-balancing and high-availability (HA) functions by providing them directly inside the driver, making them available to the clients transparently.

The cluster interconnect consists of the following hardware components:

- *Adapters* – The network interface cards that reside in each cluster host. A network adapter with multiple interfaces could become a single point of failure if the entire adapter fails.
- *Switches* – The switches, also called junctions, that reside outside of the cluster hosts. Switches perform pass-through and switching functions to enable you to connect more than two hosts. In a two-host cluster, unless the adapter hardware requires switches, you do not need switches because the hosts can be directly connected to each other through redundant physical cables. Those redundant cables are connected to redundant adapters on each host. Configurations with three or more hosts require switches.
- *Cables* – The physical connections that are placed between either two network adapters or an adapter and a switch.

Figure 3–4 shows how the three components are connected.

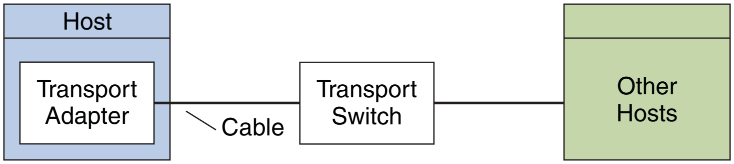

FIGURE 3–4 Cluster Interconnect

IP Network Multipathing Groups

Public network adapters are organized into IP multipathing groups (multipathing groups). Each multipathing group has one or more public network adapters. Each adapter in a multipathing group can be active, or you can configure standby interfaces that are inactive unless a failover occurs.

Multipathing groups provide the foundation for logical hostname and shared address resources. The same multipathing group on a node can host any number of logical hostname or shared address resources. To monitor public network connectivity of cluster nodes, you can create multipathing.

For more information about logical hostname and shared address resources, see the *Sun Cluster Data Services Planning and Administration Guide for Solaris OS*.

Public Network Interfaces

Clients connect to the cluster through the public network interfaces. Each network adapter card can connect to one or more public networks, depending on whether the card has multiple hardware interfaces. You can set up hosts to include multiple public network interface cards that are configured so that multiple cards are active, and serve as failover backups for one another. If one of the adapters fails, the Solaris Internet Protocol (IP) network multipathing software on Sun Cluster is called to fail over the defective interface to another adapter in the group.

Index