Sun™ Cluster 3 Architecture

A Technical Overview

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303
1 (800) 786.7638
1.512.434.1511

Please
Recycle

Contents

Introduction 1
Clustering Overview 2
- High Availability 2
- Failover and Scalability 3
- Types of Clusters 4
- Complexity of Solutions for Availability 5
- Causes of System Downtime 5
- Costs of System Downtime 6
- Reducing Risk 7
- Predictable Service Levels 7
- Lowering Probability of Failure and Impact of Failure 8

Sun Cluster 3 Features and Benefits 9
- Global Devices 9
- Global File Service 9
- Global Network Service 10
- Scalable Services 10
- Failover Services 10
- Faster Failover 10
- Diskless Failover 11
- Simplified System Administration 11
- Sun Cluster 3 API 11
- Cluster Framework Integration into the Solaris Operating Environment 11

Sun Cluster 3 Overview 13
Hardware Overview 13
Sun Cluster 3 Software Overview 15
Deployment Examples 20
Unmodified Applications 20
Failover Services 21
Scalable Services 21
Tightly Integrated Applications 22
Sun Cluster 3 Implementation Details 23
Global Devices 23
Volume Managers 24
Global File Service 24
Cluster Membership 26
Continuous Service Availability 27
Cluster Networking 30
Creating Cluster-Aware Applications 33
Resources 33
Resource Group Manager 34
Creating Agents 34
Cluster Management 35
Sun Management Center Integration 35
Service Offerings 38
Sun Cluster Application Readiness Service 39
Sun Cluster Data Center Readiness Service 39
SunReady Sustaining Services 39
Sun Remote Services 40
Sun Educational Services 40
Additional Information 41

CHAPTER 1

Introduction

This white paper is written for IT Architects and technical customers who require a thorough understanding of the Sun™ Cluster 3 architecture. The document covers the features and benefits of the Sun Cluster 3 service delivery platform, its integration with the Solaris™ Operating Environment, and its highly-available and scalable services. First, the effects of cluster technologies are discussed — their benefits to business, how they reduce risk to the organization, and how Sun Cluster 3 improves service deployment and manageability. Next is an overview of the Sun Cluster 3 product and specific implementation details. And finally, service offerings from Sun Support Services are presented.

CHAPTER 2

Clustering Overview

Sun Cluster 3 is Sun's next-generation clustering technology, focusing on integrated availability, scalability, and manageability of the service delivery platform. It provides a set of continuously available operating system services that enable high availability of all application services. Sun Cluster 3 provides a single management view for all these services, enabling the cluster to be administered as a whole. The cluster can scale in capacity, while incremental management costs are minimized. Sun Cluster 3 extends the Solaris 8 platform, including devices, file systems, and networks across cluster nodes, while maintaining full Solaris software compatibility for existing applications. System administration tasks such as data backup and restore, installation of patches, software upgrades, and adding new hardware can be done without interrupting service delivery.

Sun Cluster 3 — also known by its internal development code name *Full Moon* — enables applications to scale by using larger servers and multiples of them concurrently. To users, the cluster appears as a single system, capable of providing near-continuous access to application services, including databases, Web services, and file services. Individual cluster nodes still run independent copies of the Solaris Operating Environment, providing fault isolation so that no single failure in either hardware or software will interrupt the delivery of the application services.

High Availability

Sun Cluster 3 delivers high availability through a combination of well-managed hardware and software components. Availability is maximized by configuring the cluster without any single points of failure, including public networks, cluster interconnect, storage, and software service components. Data integrity is maintained by the cluster, ensuring that only healthy nodes are allowed to participate in the delivery of services. The health of each individual node, along with its hardware and software components, is constantly monitored. Failing or failed nodes are prevented from delivering services and accessing data. Failed software components can be restarted within the cluster, and failed nodes may return to the cluster following repair.

The Full Moon framework is designed to provide a continuously available platform, enabling highly available or continuously available services. This can be contrasted to fault-tolerant hardware systems, which provide constant access to data and applications, but cost more because of specialized hardware. And these fault-tolerant systems typically do not account for software failures.

Failover and Scalability

Sun Cluster 3 enables applications to be implemented as a failover or scalable service. A failover application provides high availability; a scalable application provides both high availability as well as increased performance. Both failover and scalable applications can run on the same cluster concurrently.

Failover

Failover is the process by which the cluster automatically relocates an application from a failed node to a healthy one. When a failover occurs, clients may see a brief interruption of service, but they are not aware that the application has been rehosted on a different physical cluster node. Many types of failure do not have to result in the application failing over to another node. For example, since the file system, by itself, is a continuously available service, any single failure related to the file service will go unnoticed to the application, thereby minimizing the chance of having to failover in the first place.

Scalability

Scalability enables a service to meet increasing load requirements, while delivering the same quality of service. A scalable application service leverages the multiple nodes in a cluster by running multiple instances of the same application service concurrently on multiple nodes.

At the service level, scalable performance is provided through parallelism. Cluster services typically use inter-job parallelism, where individual transactions can execute in parallel on multiple cluster nodes as long as the application data set is sufficiently partitionable. A good example is a Web service, where individual instances can process client requests independently of each other. In cases where application data sets are not partitionable, the application could still be made into a scalable service, but it would be up to the application to maintain data synchronization.

Types of Clusters

There are many types of clusters available today that can be roughly categorized by their intended purpose:

Clustering for Scaling and Resilience

These types of clusters are loosely coupled, containing multiple systems strung together through a workload-balancing paradigm such as a load-balancing switch. They are horizontally scaling clusters. The systems are not aware of other systems within the cluster. The individual application instances typically manage only autonomous, independent transactions or sessions with independent data sets, which are either replicated to the individual nodes or available over the network such as a NFS server. Cluster nodes maintain session state only; they do not maintain transactional state. Nodes are administered individually, aided by home-grown procedures and practices that ease the management of large numbers of these systems. A good example of a horizontally scaling cluster is a typical Web farm, where data integrity is not usually a problem because the data set is mostly read-only, and dynamic transactions such as order processing are handled on another tier in the architecture.

Clustering for Performance

These types of clusters are typical for high performance computing (HPC) workloads that focus on performance and scalability by applying as many CPUs to a problem as possible. Most scientific clusters use some form of batch-processing or work-sharing software. There are usually no resilience features; in case of application or system failure, a checkpoint/resume mechanism might exist to provide the restarting of failed batch loads. Scientific clusters are used for a wide range of disciplines, including biology (genome mapping, protein folding), engineering (turbo-fan design, automobile design), high-energy physics (nuclear-weapons simulation), astrophysics (galaxy simulation), and meteorology (climate simulation, earth/ocean modeling).

Clustering for High Availability

High availability (HA) clusters add availability features on top of operating system infrastructures. Availability is achieved by using scripts that monitor application service health on individual cluster nodes. In case of service failure (due to the failing of disks, networks, or the application service itself), the application will be restarted on another node. Individual cluster nodes are primarily administered independently of each other.

General-Purpose Clusters

Sun Cluster 3 is a general-purpose cluster that focuses on several key attributes for the data center: high availability, scalability, and manageability. It is a tightly-coupled cluster where nodes have a notion of cluster membership, which is necessary to maintain data integrity. It enables multiple instances of the Solaris platform to be viewed and managed as a single environment. Resources are shared across all nodes in the cluster, enabling applications to run anywhere in the cluster with a consistent view of these resources, even across failures. Applications can be made scalable through replication, thereby increasing performance, throughput, and failure resiliency.

Complexity of Solutions for Availability

Due to the explosive growth of the Internet — impacting IT infrastructure in the amount of data, number of users, and applications — data center infrastructures are becoming more complex and difficult to manage. To cope with exponential business growth, IT departments have scaled their infrastructures by adding more servers to increase performance and availability. This server sprawl is now creating another problem: how to manage all of these servers.

A major concern of IT departments is whether changes to the infrastructure or application can be made quickly or incrementally to the environment, without negatively affecting the entire infrastructure. Another concern is whether system maintenance can be performed without affecting service levels. The process of getting to where IT is today has been to keep the application up and running. Functionality is the initial market differentiator. Availability and scalability are after-thoughts, and have come into focus because of rapid growth in the Internet market. Because these critical infrastructure decisions were not considered during the development and design process, changing these applications and infrastructures presents costly and risky challenges.

Causes of System Downtime

- Planned downtime — usually the largest cause of downtime, includes such activities as adding hardware to the system, upgrading operating systems or applications, proactive rebooting to cleanup logs or directories
- Hardware — disks, system cooling, power supplies, memory, CPUs
- Software failures/bugs — client, server, network
- People — make mistakes or do not completely understand how to manage or operate a system

With Sun Cluster 3, most hardware and software maintenance can be performed without interrupting service delivery. For the people and process factors, Sun Cluster 3 provides a single management view that enables an administrator to perform many maintenance tasks on the cluster as if it were a single system. For example, a disk can be added to one node and automatically be made available to all other nodes. This eases administration and lowers the chance of human error. Application service delivery levels are maximized by enabling applications to run on multiple nodes concurrently. If one node is interrupted, the others can continue to provide service, avoiding any point in time where the service is unavailable.

Costs of System Downtime

The difference between Internet site downtime from internal infrastructure downtime is that customers are instantly impacted rather than being shielded from downtime by company employees. Downtime must be viewed from the user point of view, and includes inability to access or purchase products or services from the site (for any reason), as well as performance issues. Performance is an attribute of availability, and when response time takes longer than a user will tolerate (typically somewhere from 6 to 10 seconds), they will consider the service unavailable and leave it.

In the Net economy, the cost of downtime is usually much greater. Losses from downtime include direct revenue (e.g., lost sales during the outage period that are not made up at a later time), revenues from current and prospective customers due to negative publicity, and harm to the enterprise's reputation. The complexity of most Web application infrastructures also makes it difficult to attain consistent 100 percent uptime. Moreover, the cost of achieving higher levels of availability increases exponentially. To substantiate investments in availability, enterprises should compute a return on investment analysis to assess that the benefit (reduced cost of downtime) outweighs the cost of the investment.

According to a 1999 Standish Group[1] report on downtime and high-end enterprise systems, Sun Cluster had the lowest operational cost of the major enterprise server vendors as well as the lowest operational cost for running mission-critical applications because of its simplified implementation. In addition, because Sun Cluster requires fewer people to manage it, operational costs are also lower.

The Sun Cluster 3 architecture provides further simplicity for managing multiple systems, further reducing the operational costs associated with managing clusters. Fast failover times, centralized management, global networking and storage, and continuous application availability all reduce the total cost of clusters.

1. The Standish Group, “TCO in the Trenches,” A Standish Group Research Note, November 1999

Reducing Risk

Complexity hampers the ability to make change rapidly to the IT infrastructure and increases the risk of such change to the infrastructure's ability to support business growth. This risk represents potential financial losses because of downtime, dissatisfied customers, and missed market opportunities.

Reducing complexity not only reduces the risk to the organization, it also reduces costs associated with increasing service levels. Responding quickly to competitive threats in the marketplace is critical to business success. If an organization takes longer to deploy new services than a competitor, that competitor will capture the market. At the same time, if operational or deployment costs are higher than the competition, this will also affect the bottom line. Reducing complexity requires fewer people to manage and deploy systems, and less risk of errors. Senior-level staff can focus on the most challenging concerns of the organization, while junior staff carry out simplified system tasks.

Predictable Service Levels

Reducing complexity usually goes hand in hand with reducing risk as well as cost. This enables businesses to deliver constant service levels under ever-increasing demand, creating a competitive advantage. Predictable service requires the following elements:

- Ability to scale the service without downtime
- Continuous availability with quick response time
- Disciplined processes and trained personnel
- High resource utilization rates

Sun Cluster is a core component of Sun's Genesys architectural platform and the overall datacenter.com Initiative, which proposes the following key market objectives:

- Increase Service Levels

  Service levels encompasses the availability and predictability of service. Predictability includes performance, measured by throughput or response times, as well as data integrity, security, and user interface consistency.

- Contain Service-Level Costs

  Service-level costs refer to the costs of operating the IT infrastructure, which is a subset of the total cost of ownership.

- Achieve High Operational Simplicity

  Complex environments are difficult and costly to manage. Complexity also reduces the predictability and reliability of applications and systems.

- Lower Risk to the Customer

  Cross-generational continuity and investment protection is key to lower customer risks.

Lowering Probability of Failure and Impact of Failure

Sun Cluster 3 significantly reduces the effect and impact of several common failures relating to data, network, and server. For all types of single failures, Sun Cluster 3 provides alternate paths to access the data or network. With Scalable Services, other application instances continue to provide service during the failure of a single system. Data integrity is maximized through a continuously-available Global File Service with built-in transactional semantics, which allows file system operations to be consistent even in case of failure.

CHAPTER 3

Sun Cluster 3 Features and Benefits

Sun Cluster 3 enables multiple Solaris system nodes to be clustered together to form a single-managed, continuously-available, complete platform environment for deploying applications. It enables a number of cluster-wide Solaris software resource building blocks, such as a Global Devices, Global File Service, and Global Network Services, that are always available to the services that depend on them, even in case of failures. In many cases, hardware or software failures are completely transparent to the applications.

Key features and benefits of Sun Cluster 3 are:

Global Devices

Sun Cluster 3 automatically detects all storage devices on system boot up.

Global Devices provide improved ease of use, seamless storage scalability, and simplified cluster administration. Devices can be accessed from anywhere in the cluster, and if the device allows redundant hardware connectivity, physical path failures are shielded from the user.

Global File Service

Data no longer needs to be physically attached to the server that hosts the service. The Global File Service provides continuous data availability and reduced complexity of system administration. Application data is always available, anywhere in the cluster.

Global Network Service

The ability of Sun Cluster 3 to abstract a service from a network interface means that an IP service can reside anywhere in a cluster environment. In Sun Cluster 3, there is a *many-to-many* relationship between IP services and IP addresses. Multiple IP addresses can be used for a single service, or multiple services can use a single IP address. IP interfaces can be started or stopped dynamically and migrated from one Solaris domain to another without bringing down the service.

In addition, Sun Cluster 3 utilizes the network adapter failover (NAFO) mechanism to provide failover capabilities within the same Solaris domain. These facilities are seamless to the application, and result in improved service availability. Failures of network interface components do not result in service interruption.

Scalable Services

Scalable Services enable a single application or a series of applications to run across multiple systems in the cluster. This ensures that service is maintained during hardware and software outages. Scalable Services enable IT organizations to maintain service levels for critical applications and services.

Failover Services

Sun Cluster 3 delivers inherent failover services. It enables IT organizations to maintain service levels on critical applications and services. Multiple services can share the cluster, and multiple service instances ensure that the service is maintained in the event of a failure.

Faster Failover

Sun Cluster 3 provides fast error detection, fast software failover, and parallelized application restarts. When a failover occurs, users may see a brief interruption in service. Only applications which are directly affected by a failure might experience this interruption. All others will continue to provide service uninterrupted.

Diskless Failover

With the foundation of these key abstracted capabilities — Global Devices, Global File Service, and Scalable Services — application failover does not require storage device or file system failover. Hence, diskless failover provides flexibility and reduces failover time.

Simplified System Administration

The Sun Cluster 3 framework simplifies cluster administration. All cluster resources can be managed and administered as if they were on a single system. This provides tremendous cost savings for organizations by reducing risk of errors and lowering the number of administrators required to administer the cluster.

Sun Cluster 3 API

The Sun Cluster 3 API enables off-the-shelf applications to be integrated into the cluster framework. An SDK and agent builder wizard tool are available to ease the development effort for this.

Cluster Framework Integration into the Solaris Operating Environment

Sun Cluster technology is tightly integrated into the Solaris Operating Environment. It extends the Solaris platform, allowing applications to provide continuous and scalable service levels. The Sun Cluster 3 framework enables system resources such as devices, file systems, and networks to be shared in the cluster. It provides a single, integrated hardware and software platform for applications that share the available cluster resources. More resources can be added to the platform by adding more resources to the cluster, either in the form of additional cluster nodes or by adding resources to existing nodes.

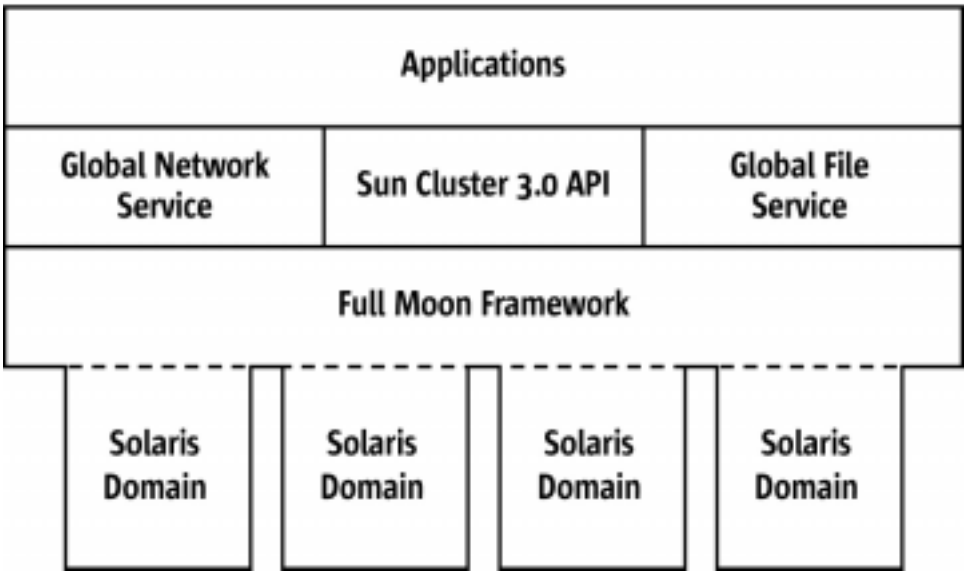

FIGURE 1 Sun's Clustering Framework

The integration of Sun Cluster 3 with Solaris technology allows clusters to be managed as a whole by the administrator, with a single management view. Existing Solaris software management interfaces are extended to operate on a cluster-wide basis. This greatly reduces overall cost of ownership; the cost of additional cluster nodes does not result in linear management cost increases.

CHAPTER 4

Sun Cluster 3 Overview

Hardware Overview

The Sun Cluster 3 architecture generally depends on off-the-shelf hardware components. Cluster nodes are individual Solaris system domains that are connected by a high-speed cluster interconnect. Access to common data is guaranteed by connecting multiported storage devices to two or more (if supported by the storage devices) cluster nodes. This way, if one cluster node fails, the data will still be available through an alternate hardware path.

Cluster Nodes

Each cluster node is characterized by the fact that it runs an independent copy of the Solaris Operating Environment. This could be, for example, an independent physical server such as the Sun Enterprise™ 6500, or a single domain within a Sun Enterprise 10000. The specific hardware that is supported will change over time. A Sun representative will know which specific configurations are supported.

Sun Cluster software is layered on top of the cluster nodes, and provides the illusion of a single system to clients accessing applications running on the cluster.

Cluster Interconnect

All cluster nodes are connected through a private cluster communications interconnect. This interconnect is used for any information exchange that is needed between cluster nodes. Examples are heartbeat health checks, cluster membership information, application data, file system traffic, load-balanced workload distribution, and IP service resource management.

The cluster interconnect has been designed to meet the requirements for high availability, resiliency, and scalability.

High availability and resiliency: Clusters are configured with a minimum of two interconnect links per node. If one link fails, the cluster will still operate using the remaining interconnect link(s). Interconnect links can be added into or removed from the cluster dynamically, without any disruption to the services deployed on the cluster.

Scalability: Multiple interconnect links can be added dynamically for increased interconnect bandwidth between cluster nodes. The cluster interconnect transport software transparently aggregates and utilizes the bandwidth from the multiple links.

To achieve optimal performance, it is best to use interconnect hardware with very low-latency and high-bandwidth characteristics.

Shared Storage Devices

Access to shared data is made available by using dual-port or multiport storage devices connected to multiple cluster nodes. If access to one physical path becomes unavailable (due to failure of the connecting cluster node, for example), the storage device will still be available through another node. Sun Cluster 3 can hide these types of failures without applications even noticing, aside from a brief pause when the other node assumes control. The cluster framework guarantees that disk operations that didn't complete will be restarted, assuring the *exactly-once* transactional semantics required for mission critical environments where data corruption needs to be avoided at all cost.

Applications that utilize multiple paths to storage devices concurrently, such as Oracle Parallel Server (OPS), are also possible when appropriately configured.

Sun Cluster 3 Software Overview

The following figure presents a high-level overview of the Sun Cluster 3 architecture.

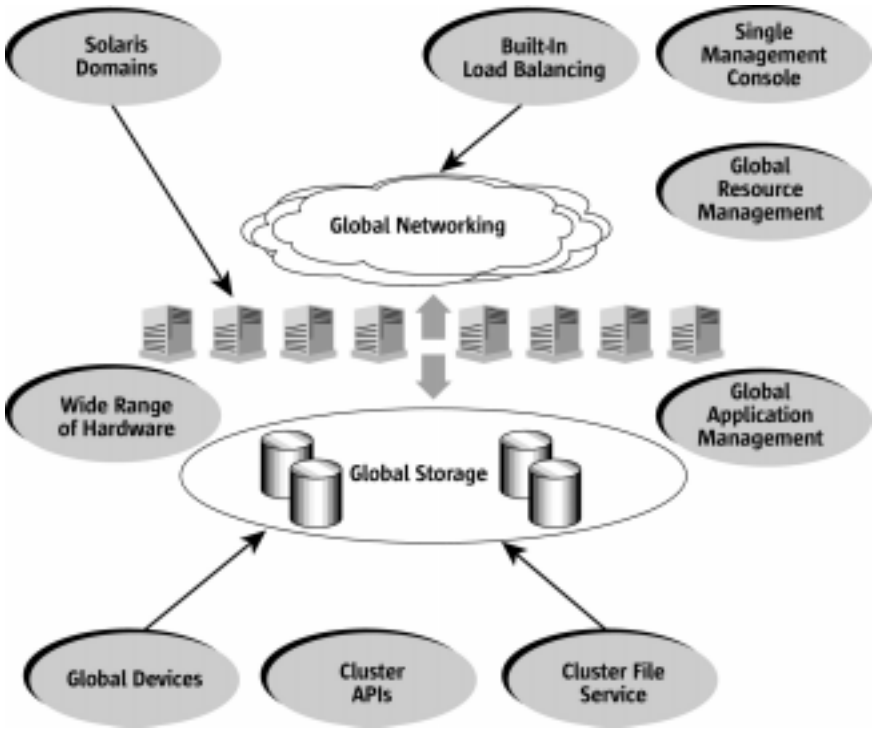

FIGURE 1 Sun Cluster 3 Architecture

Cluster Services

As illustrated in Figure 1, the software in Sun Cluster 3 can be viewed as a set of several key components. These components work together so that a collection of tightly coupled nodes presents the illusion of a single, coherent system. The Sun Cluster 3 framework provides a set of key cluster services that make this possible:

- Cluster-wide Global Devices
- Cluster-wide Global File Service (with near-continuous access to data)
- Scalable, Global Network Service (with load balancing and IP failover)
- Centralized application management
- Centralized and simplified cluster management
- Rich set of clustering, communication, load-balancing, and management APIs

Global Devices

While Sun Cluster 3 nodes maintain their independent nature, devices such as disks, tapes, and CD-ROM drives are made available globally to all members of the cluster. Each device appears as if it were local, maintaining identical and familiar access semantics and interfaces. If devices can be connected using multiple hardware paths, the Global Device path will always give access to the device, regardless of which physical hardware path or node the device is currently being accessed through. Failures are shielded as much as possible from the users that are accessing these devices, usually without the users even noticing that a failure has occurred.

Volumes can be created from Global Devices using familiar tools such as Solaris Logical Volume Manager (SLVM, formerly known as Solstice DiskSuite™ software) or VERITAS Volume Manager, and can also be made available globally. File systems can be created on top of these volumes, or applications such as databases can use the raw volumes directly.

Hot-plugging of new devices is supported as well. The new devices can be configured into the system dynamically (as in any Solaris system), and can be made available globally at the same time, after which the devices are automatically available on all other cluster nodes as if they were attached locally.

Global File Service

Sun Cluster 3 provides global access to data anywhere in the cluster by providing a cache-coherent Global File Service. Any existing and familiar file systems such as UFS can easily be made available globally when specified by the user. The file service maintains its semantics, but now has globally-available file system data and near-continuous access to such data in case of failure.

Global Network Service

In the Sun Cluster 3 environment, each node in the cluster retains its own IP address(es) and can be addressed separately from the LAN. Applications can be deployed in either of two modes: *failover* or *scalable*. In failover mode, the application runs on one cluster node at a time. If, in the case of a failure, that node cannot host the service any longer, it will automatically be rehosted on another node.

In addition, Sun Cluster 3 adds the concept of a scalable application. A scalable application can be hosted on more than one node at a time. The cluster framework load-balances client requests dynamically across these nodes. This allows for scalability beyond a single node, while maintaining the view of a single system to the clients.

Application Resource Management

The term *agent* is used to describe a third-party application such as Oracle or the iPlanet™ Web Server that has been configured to run on a cluster rather than a single server. No changes to these applications are needed in order to become cluster agents. They are merely “wrapped” with scripts that allow the cluster framework to understand how to start, stop, and monitor the health of the given service.

The Resource Group Manager (RGM) component of the cluster framework supports the registration and operation of applications. Resources that have dependencies on each other can be grouped together so that in case of failure, the correct reconfigurations can be made without impacting any service not affected by this failure.

The RGM also monitors the health of the application, and determines when failures happen and how to react to them.

Cluster Management

Sun Cluster 3 provides centralized management for many of its key components. Multiple clusters can be monitored remotely using Sun™ Management Center plug-in modules. For more details about cluster management, see Chapter 8.

Cluster APIs

A rich set of resource management APIs is available to deploy and manage applications. Existing off-the-shelf applications can be wrapped and deployed as either Failover or Scalable Services. Applications can take full advantage of the cluster communications infrastructure. A full SDK is available to aid the developer, including an agent builder wizard, which requires no additional programming and can build an agent in minutes.

Cluster Infrastructure Components

The aforementioned user-visible cluster services are built on top of a number of key cluster components. These components, mostly hidden from the user, can be divided into the following:

- A high-speed transport mechanism with transactional communications
- A high-availability and replica management system
- A cluster membership framework
- A distributed cluster configuration database

Cluster Transport and Communications

Sun Cluster 3 uses a private cluster interconnect for intra-node cluster communications. The cluster framework provides a software transport layer on top of the interconnect hardware that hides the hardware implementation details from the rest of the framework. The Sun Cluster 3 Architecture currently defines two types of interconnects:

- DLPI-based interconnects
- Remote Shared Memory (RSM)-based interconnects

Supporting DLPI[2]-based interconnects ensure that a wide array of DLPI-compliant communications devices can be used as interconnect technology, especially off-the-shelf network interface cards such as Fast Ethernet or Gigabit Ethernet. DLPI drivers for other interconnect technologies such as SCI are also supported. The DLPI interconnects leverage the standard TCP/IP protocols.

RSM-based interconnects enable direct memory-to-memory communication between cluster nodes. By accessing a remote node's memory directly, dramatic performance advantages can be achieved for certain types of applications. RSM will be used within the cluster framework internally, but will also be made available through APIs so applications can take full advantage of this. RSM support and APIs will be made available in a subsequent release of Sun Cluster 3.0 software. Early access to RSM is available on the Sun Cluster 3.0 Cool Stuff companion CD bundled with the Sun Cluster 3.0 software. This version supports SCI interconnects and can be used for application prototyping.

The cluster transport is designed with hardware independence in mind. Future technologies such as Infiniband can be integrated into the Sun Cluster 3 architecture very easily without changing other layers in the framework. This provides investment protection as well as a path for future growth.

Layered on top of the cluster transport is a framework for cluster communications, providing transactional semantics. Many of the cluster services, such as the Global File Service, are based on this framework. Transactions that are in progress can be rolled back in case of any failure that prevents them from completion, and can subsequently be retried elsewhere.

2. Data Link Provider Interface (DLPI) is the standardized interface between the data link layer (typically a network device driver) and the network layer (typically IP) in the STREAMS environment on which the Solaris networking architecture is based.

High Availability Framework

The Sun Cluster 3 architecture ensures that each component on the path between users and data is highly available, including network interfaces, applications, file systems, and multiported disks. In general, a cluster component is highly available if it survives any single (hardware or software) failure in the system.

This high availability framework detects node failures, and creates a new equivalent server for the framework resources on a remaining node in the cluster.

Most highly available framework resources are recovered transparently to the applications (data services) using the resource. The semantics of framework resource access are fully preserved across node failure. The applications cannot tell that the framework resource server has been moved to another node.

Cluster Membership Framework

In order to guarantee data consistency and service availability, it is extremely important that the cluster as a whole agrees on cluster membership-related issues. Cases like split-brain syndrome, where multiple isolated parts of the cluster think they are the only surviving members and take ownership of data independently, must be avoided. Sun Cluster 3 uses a comprehensive, distributed membership algorithm, heartbeat message framework, and fencing to achieve this. Data consistency is of highest priority and must never be compromised.

CHAPTER 5

Deployment Examples

Applications can be deployed in one of several manners with Sun Cluster 3, each providing a different trade-off between the level of integration required on the one hand, and availability and performance benefits on the other.

Unmodified Applications

Off-the-shelf applications can be deployed on Sun Cluster without any additional work required. Because Sun Cluster provides continuous availability of key services, unmodified applications can take full advantage of some of these, specifically, data and network availability.

Data Availability

If application data is stored on a Global File Service built on top of highly available Global Devices, the data will remain available even in the face of certain failures. For example, a hardware path to storage could become unavailable without impacting the application.

Network Availability

Applications can benefit also from network high availability within individual cluster nodes. If one network adapter fails, a standby network adapter can take over its IP address, without breaking any network connections.

Failover Services

Applications can be integrated more tightly into the cluster framework by using the service API to tie the application into the cluster framework. This allows applications to be started, stopped, and monitored by the cluster framework. If, for example, the node running the application fails, the application can immediately be restarted on another node.

Failover Services are characterized by the fact that at any point in time, only one instance of the application is running in the cluster.

The development work required to make applications into Failover Services does not have to be done by the application developer. The functionality wraps around applications in the form of an agent, without requiring modifications to the application itself. The development might be done by Sun, by the application developer, or by the customer directly using the SDK and agent builder wizard.

An example of a Failover Service would be an LDAP server application. The application would be registered in the cluster framework, and be brought online. The application's health would be continuously monitored, and corrective action taken when required to keep the application online. Certain types of failure would be shielded from the application. Others would cause the application to be brought down and restarted elsewhere in the cluster, for example, after a node failure.

Scalable Services

For certain types of applications, it is also possible to run multiple instances of the same application concurrently on multiple nodes in the cluster. This provides higher levels of scalability and availability.

Increased Scalability

Using Scalable Services, total service capacity can be increased by adding more nodes in the cluster that hosts an active instance of the service. The cluster framework distributes the workload across the available instances.

Increased Availability

The deployment of Scalable Services increases overall availability. If one instance of the application fails, any workload that is processed by any of the other instances is unaffected. Any new connections will bypass the failed instance and be sent to one of the remaining instances for processing. There is no need to wait for the failed instance to come back online; there will never be a moment when the application is unavailable.

The Scalable Services API is a superset of the Failover Services API, although the extra amount of work required is typically not significant.

Not all applications are suitable to be deployed as Scalable Services. Issues related to data dependencies and synchronization might prevent this from being possible. For example, multiple instances of the LDAP service are not likely to be able to update the same data store concurrently, unless the application has been designed specifically to allow this.

A good example of a Scalable Service can be a Web service. Both Apache and iPlanet Web Server scalable cluster agents are bundled with the Sun Cluster 3.0 product.

Tightly Integrated Applications

Applications can be integrated more tightly into the Sun Cluster 3 framework by custom application development. This can lead to even greater levels of performance, scalability, and availability.

A good example of this category is Oracle Parallel Server (OPS). This application allows a single instance of an Oracle database to be distributed across multiple active nodes. The Oracle application implements its own data synchronization techniques by communicating over the cluster interconnect.

Additional performance benefits can be achieved by implementing these data synchronization techniques using the Remote Shared Memory (RSM) technology, which allows applications to directly access memory on remote cluster nodes. A future version of Oracle OPS is scheduled to utilize RSM, which can potentially lead to greatly reduced synchronization latencies, a critical factor for overall OPS performance and scalability.

The advantages of a tightly integrated application like OPS are:

- **Increased scalability**. Multiple active instances of the application can run concurrently by providing data consistency at the application level.
- **Increased availability**. In case of a node failure, another instance of the application is online and already has a physical path to disk, so there is no need to wait for volumes to come online.
- **Increased performance**. RSM provides a very fast and efficient memory to memory communications infrastructure for performance-critical pieces of the applications.

CHAPTER 6

Sun Cluster 3 Implementation Details

Global Devices

Sun Cluster 3 uses Global Devices to provide cluster-wide, highly-available access to any device in a cluster, from any node, without regard to where the device is physically attached.

Disk, CD-ROM, and tape devices on each cluster node are accessible using their familiar device paths such as `/dev/dsk/c0t0d0s0`. In addition, a Global Device name space is maintained for each device in the `/dev/global` directory. This global name space is created automatically when the cluster software is first installed, and is maintained automatically by the cluster framework. By accessing the unique Global Device ID instead of the local device name, applications such as volume managers or those using raw devices can access the device in exactly the same, consistent manner from any node in the cluster.

Multiported disks that are connected to multiple hosts can, at the same time, be made highly available to their users. If the application accesses these devices by their Global Devices ID, node failures can be resolved in a completely transparent manner. The user doesn't even have to be aware that a node failure happened, since the ID continues to provide access to the device, even though device access now flows through a different hardware path.

Volume Managers

Sun Cluster 3 uses volume management software, involving mirrors and hot spare disks, to increase the availability of data. If an entire storage array fails, the data is still available using a mirrored set of the same data. Sun Cluster 3 does not require different, unfamiliar volume manager software; either Solaris Logical Volume Manager (SLVM) or VERITAS Volume Manager are supported, ensuring that volume management objects are automatically brought under cluster control and made available globally.

In addition to improving availability, volume managers can provide increased performance by using striping techniques, as in non-clustered systems.

Global File Service

The Global File Service acts as a proxy between the kernel on one node and the underlying file system and volume manager on another node (that has a physical connection to the disks).

File systems made from a Global Device can be mounted globally with `mount -g` or locally with `mount`. By mounting a file system globally, files are accessible on each node in the cluster using the exact same file name path.

The Global File Service uses coherency protocols to preserve the UNIX® file access semantics, even if the file is accessed concurrently from multiple nodes.

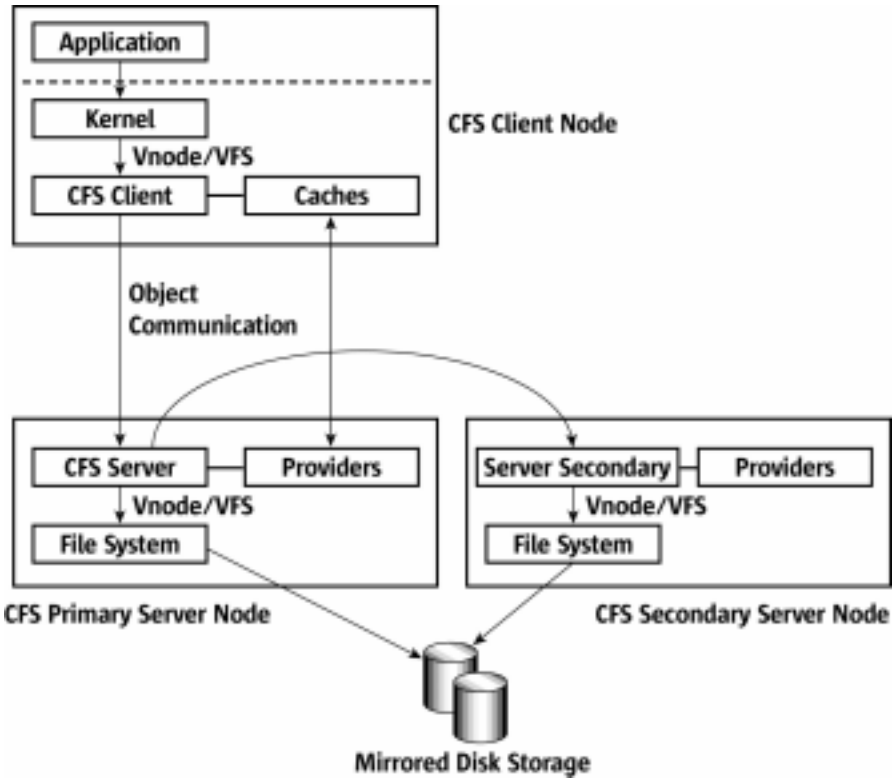

FIGURE 1 Global File Service Overview

The Global File Service is implemented at the *vnode* layer in the kernel. File system operations are intercepted at the client proxy layer and transported to a server proxy layer that can be on a local or remote node. The transport path is tuned extensively for performance, and a high degree of file system caching is implemented at the client proxy layer, while still maintaining full cache coherency.

There is no special file system type for the Global File Service. To users, it appears like the underlying file system is directly accessed. This is a great advantage in the investment protection of tools, such as backup, and personnel expertise.

Cluster Membership

In a tightly-coupled cluster environment such as Sun Cluster 3, keeping close track of cluster membership issues is important.

Cluster Membership Monitor

The Cluster Membership Monitor (CMM) is a distributed set of kernel agents, one per cluster member. The agents exchange information over the cluster interconnect relating to cluster membership. These messages are given very high priority, so that even under a very high load, it is possible to maintain consistency regarding cluster membership. If cluster nodes fail to communicate their status, they will be forced out of the cluster.

Cluster Configuration Repository

The Cluster Configuration Repository (CCR) is a distributed database for storing cluster configuration and state information. Each node maintains its own independent copy of this database, and updates are made in a failure resistant transactional (two-phase commit) manner. The CCR is updated based on information exchanged by the Cluster Membership Monitor.

Quorum

One critical aspect of cluster membership is the *split-brain* issue. Split-brain occurs if a cluster gets split into two or more independent parts, where each set of nodes thinks they are the only surviving nodes and takes ownership of data and resources. This can lead to data corruption. To solve this situation, Sun Cluster uses a method called quorum voting to achieve a consistent cluster membership view. It is a complex distributed algorithm to achieve majority voting. Only when a majority vote is achieved is the cluster allowed to continue to operate. Since there can only be one majority, it prevents split-brain from happening.

Failure Fencing

In shared-storage environments such as Sun Cluster, it is critical to maintain data integrity under all circumstances. *Failure fencing* is the technique used to ensure that failed nodes which are no longer part of the cluster cannot access and corrupt shared data. Sun Cluster 3 shields unhealthy nodes from accessing shared data using several failure fencing techniques.

Continuous Service Availability

Cluster Transport

One of the key differentiators of the Sun Cluster 3 architecture is to provide near-continuous service availability, even in case of hardware or software failure. The foundation for this is in the cluster communications architecture, depicted in Figure 2.

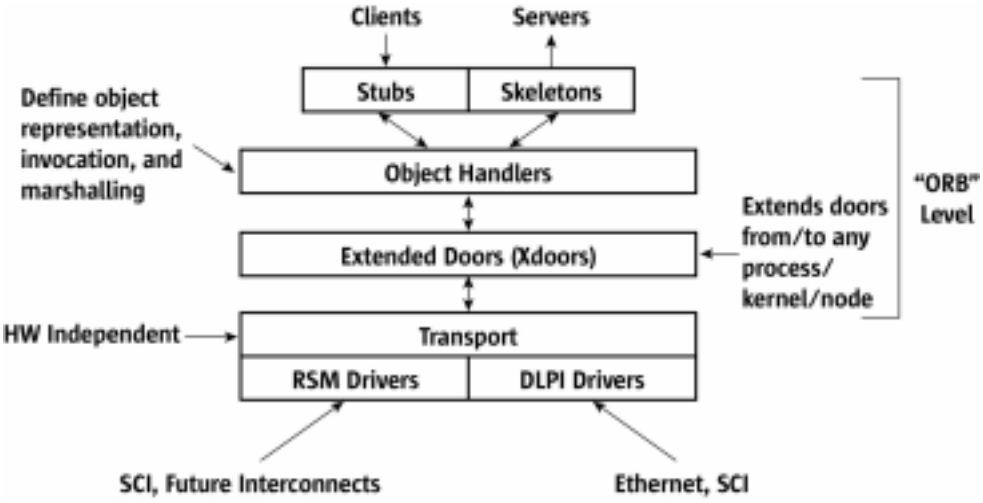

FIGURE 2 Cluster Transport

At the bottom layer, there is the lower-level transport layer. It supports DLPI drivers and interfaces as well as RSM interconnect hardware. The transport layer is mostly hardware-independent, and all hardware-specific details are hidden from the upper layers. This makes supporting future hardware interconnect functionality, such as Infiniband, very straightforward.

The transport implements the reliable sending of arbitrary-length messages. It provides a simple yet powerful interface to upper layers, taking full advantage of underlying hardware features such as zero-copy data transfer.

Doors and Xdoors

The layer directly above the transport layer provides an object-oriented mechanism to invoke objects independent of their location in the cluster. Many of its architectural details were borrowed from CORBA, and any similarities are not a coincidence.

The CORBA-like Object Request Broker (ORB) layer consists of the following three layers:

- The client stub and server skeleton
- The object handler
- The door/Xdoor layer

The stub and skeleton handlers deal with marshalling and unmarshalling object parameters in forms suitable for transport. The *client* in this context refers to the client who is invoking an operation on an object. The *server* is the actual object being invoked.

The object handler deals with locating and handling object references. The handler knows where the object lives that is being invoked.

The door and Xdoor layers[3] are the mechanisms through which object invocations are transported. Doors is a standard Solaris software interprocess communication feature that provides very fast communication between address spaces. Xdoors is a Sun Cluster specific extension to doors enabling intra-node communications using the cluster transport. Figure 3 shows an Xdoors extension.

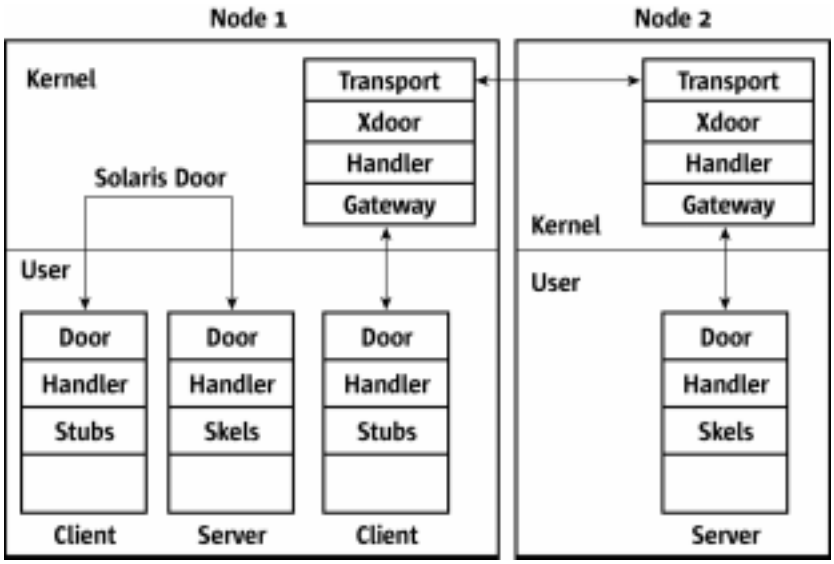

FIGURE 3 Sun Cluster 3 Door/Xdoor Layer

3. The Xdoor mechanism is only used internal to the Sun Cluster 3 framework. Its interfaces are not available or documented to external parties, and are only described to illustrate the inner workings of the Sun Cluster 3 architecture.

Replica Management

Service high availability is achieved by using server (object) replication. Replicated servers are called replicas, and are usually started on different nodes within the cluster. At any point in time, one replica will be designated the *primary* for each service; there can be one or more *secondaries* as well as *standbys*. The client only contacts the primary server. A standby is a passive secondary that can become an active secondary.

In order to be able to recover transparently from server failures, mini-transactions are used, as illustrated in Figure 4. For each object invocation, the secondaries (but not the standbys) receive checkpoints to enable instantaneous failover if needed. If an object invocation does not complete before a failure prevents the primary from doing so, the transaction can be rolled back. A secondary will be promoted to become the primary and the transaction will be completed on this new primary.

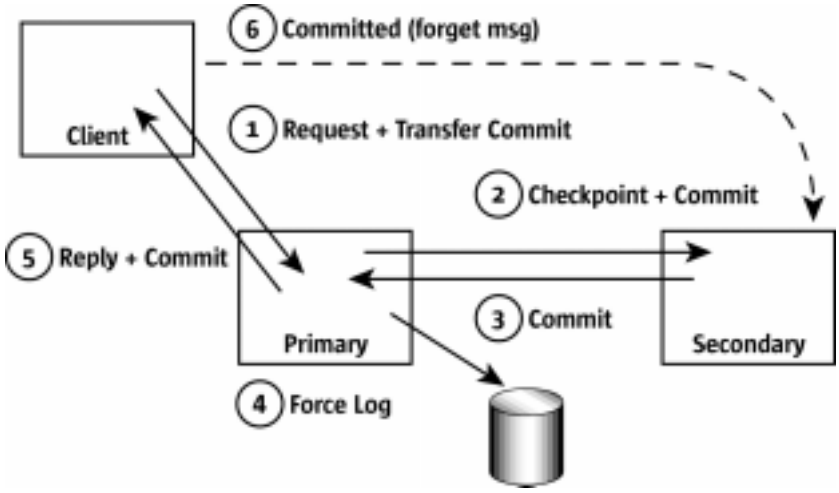

FIGURE 4 Sun Cluster 3 Mini-transactions

The object handler shields the client from all these details. The client only sees the primary server, and all transaction and replica-related housekeeping is done by the handler. The client has no knowledge or concern about which physical node the object lives on.

As an example of this, in a dual-ported storage array that is connected to two cluster nodes, one cluster node acts as the primary, the other as the secondary. A third node, without any shared storage directly attached, manipulates files through a global mount using the Global File Service. If, during the updating of some files, the node containing the primary replica crashes, the secondary node takes over control of the storage array and the volumes associated with it. File system consistency checks can also be performed. The object handler temporarily blocks all object calls. The secondary is then promoted to the role of the new primary, and any blocked object invocations are released, including the ones that had to be rolled back. From the client's perspective, there was no failure except a brief interruption. Error messages such as EIO are only propagated back to the client if the Global File Service cannot handle the failure (i.e., when all primaries and all secondaries failed).

Cluster Networking

Sun Cluster supports multiple network adapters per node, allowing for increased flexibility, performance and availability. Network interfaces can be divided into two categories:

- **Private interconnect interfaces.** These are under the control of the cluster framework. Multiple interconnect interfaces are required to avoid a single point of failure in the cluster. The cluster framework monitors health and automatically spreads the load across the available interfaces. Interconnect faults are isolated transparently without affecting the rest of the interconnect communications.
- **Public network interfaces.** These are the interfaces used by the clients to access the applications running on the cluster.

Global Network Service

Public network interfaces are used to expose applications to the clients. The Resource Group Manager knows the node the IP address needs to be hosted on, and can migrate the IP address (as well as the application) to another cluster node in case the primary node fails.

Each node can configure multiple public network adapters in a group. If a single network adapter fails, the IP address can be migrated to another adapter in the group without requiring the file service and IP address to be rehosted on another. Network interfaces are monitored continuously to determine health. If all interfaces in the group fail, the IP address will be migrated to another node in the cluster. Active network connections will be broken, and it is up to the clients accessing the cluster to reissue these.

Scalable Services

One of the unique features of Sun Cluster 3 is that certain applications can be made to run simultaneously on multiple cluster nodes. The main benefits of doing this are:

- Increased application service **scalability**, where more service capacity can be added by adding more nodes in the cluster (in addition to being able to add more capacity to individual nodes, e.g., by adding more CPUs/memory).
- Increased application service **resilience**, where the service will continue to be available without any interruption when one instance fails (existing connections that were maintained on the failing node will have to be retried, but any connections handled on any other node, as well as any new connections are unaffected).

A Scalable Service consists of:

- A set of processes by which each process runs on a different node
- A protocol type (TCP or UDP) and port numbers used by the service
- A set of one or more global IP addresses

A Scalable Service is defined such that any process that is part of the service may accept TCP connections (or UDP packets) sent to the global IP address(es) of the service. New connections are transparently forwarded to one of these processes over the cluster interconnect. All subsequent traffic from that same connection will be forwarded to that same process.

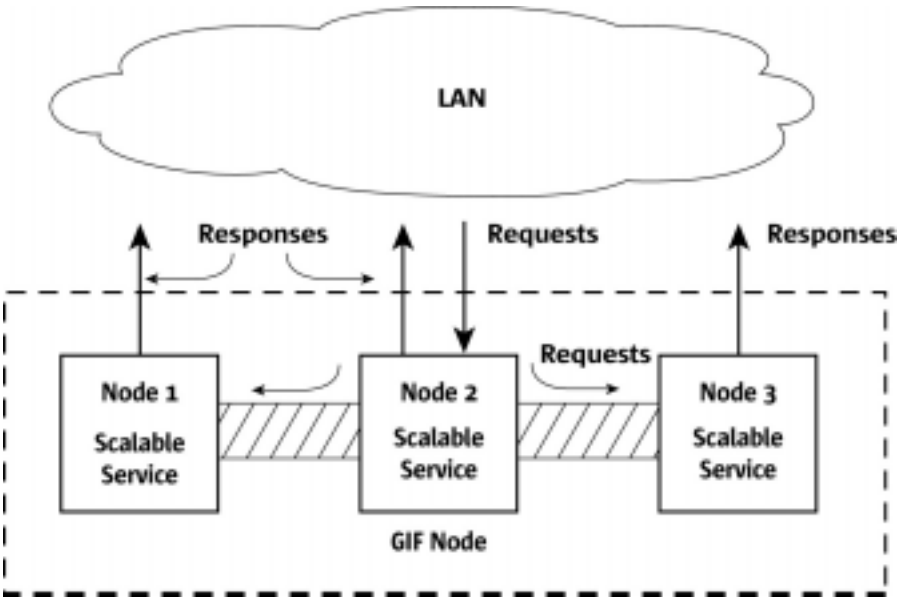

FIGURE 5 Global Network Service Overview

One interface in the cluster is assigned to be the global interface for a particular IP address, and is known as the *GIF* node. Incoming requests enter the cluster at the global interface. The GIF node is responsible for distributing the request to the most appropriate node in the cluster for processing. Outgoing network traffic leaves through a local interface, and does not have to be routed back through the GIF node.

The GIF node maintains a dispatch table for efficient lookups to determine where connections are processed. This mechanism is a highly available service by itself. If a GIF node fails, another node in the cluster will immediately be promoted to act as the new GIF node. Since any transactional state associated with the GIF is replicated, a GIF node failure will not break any existing network connections, except the ones which had their endpoint on the GIF node itself.

CHAPTER 7

Creating Cluster-Aware Applications

Sun Cluster 3 provides a platform for deploying existing off-the-shelf applications and making them continuously available. These applications — Oracle, iPlanet Web Server, or DNS — are tied in the cluster framework by means of the Sun Cluster Resource Management API, in the form of a cluster agent. An SDK is available to create these agents.

Resources

The Sun Cluster 3 Resource Management framework is built around object-oriented semantics. There are three types of objects that can be distinguished: resource types, resources, and resource groups.

A *resource type* is semantically similar to an object class. It describes the application to be run on the cluster, as well as how to start, stop, and monitor that application. A resource type can also represent other system resources such as network addresses.

A *resource* can be compared to a resource type (object class) *instantiation*. There can be several resources instantiating the same resource type, each with unique values. For example, there could be several different Web server instantiations, each instantiating a Web service resource type. Each resource could have different parameters, such as the IP address they are listening to.

A *resource group* is a collection of related resources. Each resource belongs to a resource group. The Resource Group Manager (RGM) brings all resources in a resource group online and offline together on the same node. This allows dependencies to be created between services.

Resource Group Manager

The Resource Group Manager is a system service responsible for keeping resource groups online in case of failures. If a particular resource fails, the RGM can try to restart it on the same node it was running on, and if that is not possible, it can restart the resource on another node. The RGM has administrative interfaces for registering resources and controlling their behaviors.

Creating Agents

Customers can create their own cluster agents using the Resource Management APIs. This can be done in a variety of ways, ranging from low-level API access for fine-grained control, to point-and-click GUIs allowing data services to be created without writing a single line of code. Figure 1 provides an example of the latter.

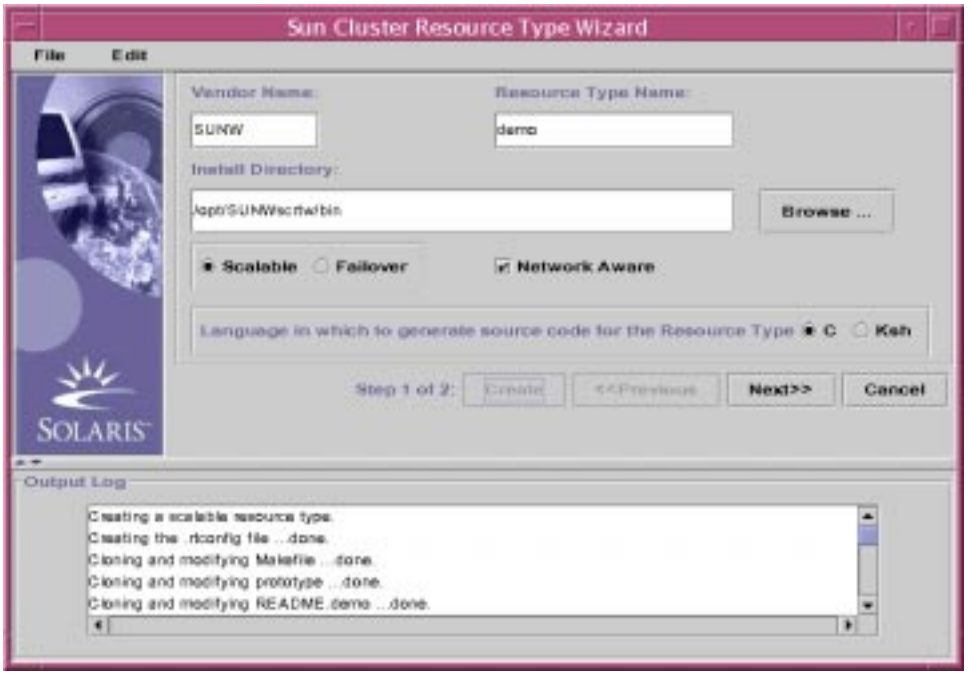

FIGURE 1 Creating Sun Cluster 3 Agents

CHAPTER 8

Cluster Management

Sun Cluster 3 provides a centralized management view of its key services. It enables administration through a rich and complete command-line interface. Cluster node installation can be integrated on top of the Solaris JumpStart™ framework, enabling automated installs of cluster nodes, saving time and reducing the chance of human error.

Devices are managed mostly automatically by the cluster framework. When a device is added, the framework takes care of making it globally available at all cluster nodes, without the administrator having to do anything special.

Sun Management Center Integration

Sun Cluster 3 provides powerful management integration with Sun Management Center. Multiple clusters can be monitored from a single integrated console, which also provides management for generic Sun hardware and software components.

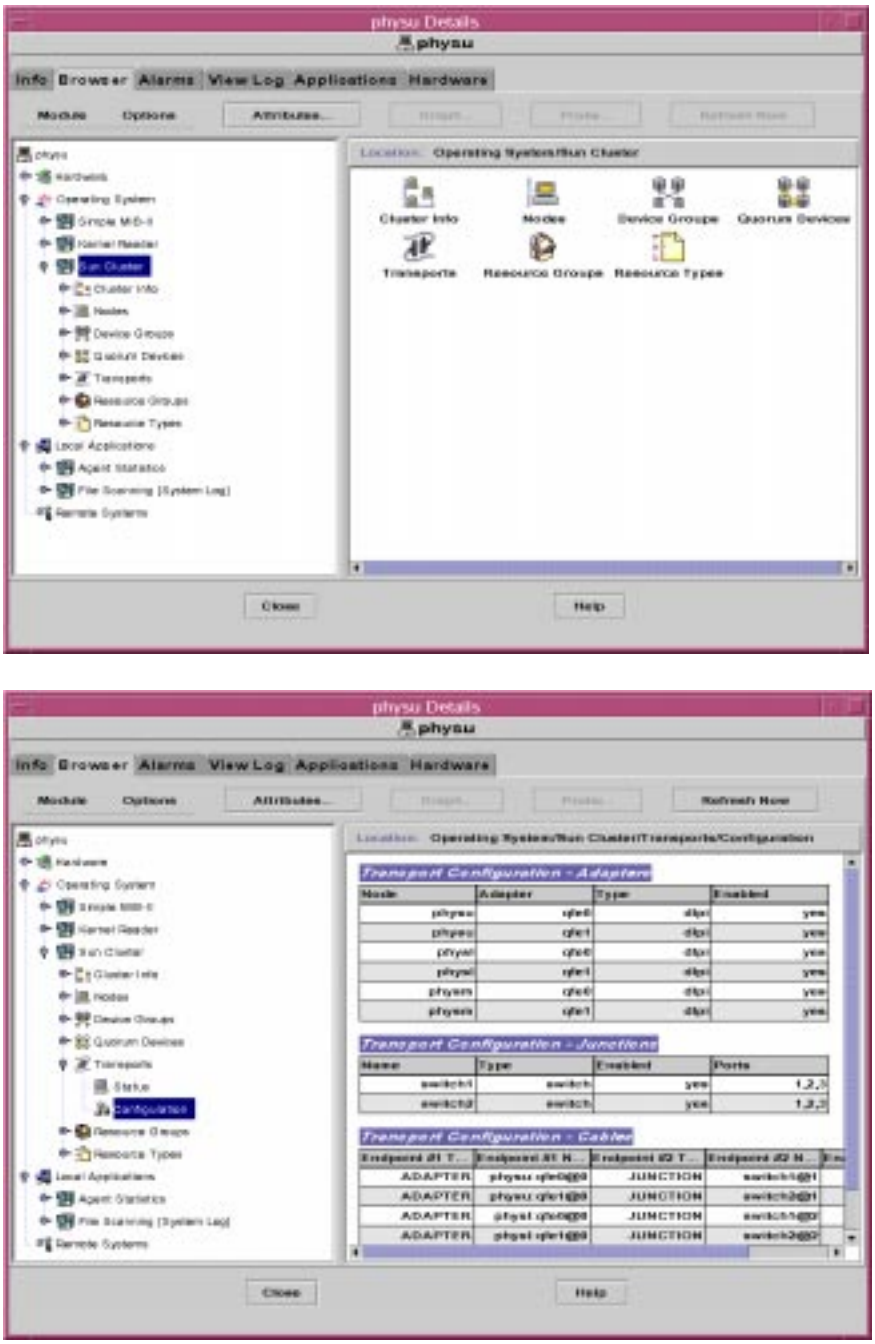

FIGURE 1 Managing Sun Cluster 3 Using Sun Management Center

All Sun Management Center features can be applied to managing Sun Cluster environments, including managing across geographies, advanced rules and alarms, and a high degree of customization.

CHAPTER 9

Service Offerings

Sun's expert consulting services, flexible training, and support services can help harness the explosive growth and availability demands associated with mission-critical services. Sun Cluster Services are designed to help increase service levels while decreasing service level costs, resulting in lower risk. Sun puts the best minds and technology in the industry together to help build massive scalability and continuous availability into Internet solutions, right from the start.

Sun consultants plan, design, and implement innovative solutions to help manage data across the Net cleanly, quickly, and without impacting business, even while it grows. Sun's commitment to availability is real. It's built into our culture with fully integrated availability policies and initiatives designed to address the three major causes of downtime — people, process, and products.

In addition, Sun Educational Services offers comprehensive training solutions, including courses specific to Service Management and Sun Cluster 3, that can help business staff proficiently operate their new environment, reducing downtime and increasing employee satisfaction and retention.

SunReady[SM] Sustaining Services focus specifically on post-implementation support of the people, process, and products that make IT infrastructure click. Through global closed-loop support planning, managed service delivery, and continuous business/IT alignment, SunReady Sustaining Services help businesses achieve customer-defined service levels and reliably scale and evolve their systems.

The Sun[SM] Remote Services portfolio comprises a series of proactive remote monitoring and event reporting services designed to identify potential problems before they occur, to help provide greater system uptime for mission-critical environments. This service is available to SunSpectrum Platinum[SM] and SunSpectrum Gold[SM] service plan customers.

Sun Cluster Services are designed as an integral part of Sun's company-wide availability initiatives. These services are designed to build availability into the installation of Sun Cluster software. With project management, configuration, and testing, Sun helps ensure that an application is up and running within a robust operational environment — right from the start.

Sun Cluster Application Readiness Service

The Sun Cluster Application Readiness Service provides the essential implementation and project management services required to configure new installation of the Sun Cluster software, and establish and test basic operational parameters. This service is a prerequisite for Sun Cluster Data Center Readiness. And because it is key to the availability of client systems, it is mandatory with each Sun Cluster installation.

Sun Cluster Data Center Readiness Service

This service enhances the Sun Cluster Application Readiness Service by providing more stringent testing and customization of procedures and documentation aimed at providing a robust operational environment suitable to support use of the cluster within a data center. This service is designed to meet the needs of clients who have have already purchased the Sun Cluster Application Readiness Service and are establishing mission-critical data centers, are not already proficient in data center operations, or wish to have help clustering new applications that are supported by the standard framework. This service includes expanded design, implementation, configuration of Standard Framework APIs and testing phases (e.g., application testing), and provides setup and best-practice guidance for essential operational and procedural issues around the cluster configuration. Extensive customized online operational documentation is also included.

SunReady Sustaining Services

SunReady Production Services provides an account-planning architecture consisting of a rich mix of preemptive, proactive, and reactive support services to help the client use cluster technologies to achieve, manage and sustain high availability in the software environment. The sustaining services methodology is based on a continuous cycle of planning, services execution, measurement, and feedback. Through experience in building and supporting many of the world's largest and most complex Net-connected infrastructures, Sun has refined market-standard IT management practices to meet the demands of new data center applications.

Sun Remote Services

The Sun Remote Services portfolio comprises a series of proactive, preemptive remote monitoring and event-reporting services designed to help provide greater system uptime to mission-critical environments. Remote, 24x7 real-time systems monitoring helps provide proactive event detection and resolution on all monitored Sun systems anywhere in the world. Sun Remote Services also provides status and trend analysis reports, vital for strategic planning. This service is available to SunSpectrum Platinum and SunSpectrum Gold service plan customers.

Sun Remote Services 2.1

Sun Remote Services is a new operational enhancement to Sun's remote monitoring solution set. This upgrade is designed with new reporting features and future Web view capabilities, and is available to customers via the MySun[SM] center at http://mysun.sun.com.

Sun Remote Services Net Connect Software

Sun Remote Services Net Connect software, the newest offering from Sun Remote Services, is an availability monitoring software tool that provides:

- Faster problem detection and *Internet-time* reporting for data collection
- Alarm collection and availability reporting
- Web-based reports for viewing Sun system environment status and the data collected

Sun Educational Services

As a leading supplier of training solutions around the world, Sun Educational Services is ready to help clients dot-com their staff's skills. Sun teaches how to get the greatest advantage from multiple technologies as they become the IT backbone of dot-com organizations in the Net economy.

Sun offers two new instructor-led courses, one specific to the administration of Sun Cluster 3, and one specific to Service Management:

- ES-330, which covers administration of Sun Cluster 3
- Sun Educational Services also offers a new instructor-led course on Sun Management Center.

CHAPTER 10

Additional Information

More information on Sun Cluster 3 can be found on the following Web sites:

- http://www.sun.com/clusters
- http://www.sun.com/sunmanagementcenter
- http://www.sun.com/service/solutions/suncluster

Other Web sites for Sun information:

- http://www.sun.com/solaris
- http://www.sun.com
- http://www.sun.com/aboutsun/media/presskits/servicemgmt
- http://reseller.sun.com
- http://www.sun.com/bigadmin

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303

1 (800) 786.7638
1.512.434.1511

http://www.sun.com/clusters

December 2000