Technical Introduction to Oracle Fail Safe

A Guide to Concepts and Terminology

An Oracle Technical White Paper

October 2001

Laurence Clarke

Oracle New England Development Center

Server Technologies

failsafe_us@oracle.com

Technical Introduction to Oracle Fail Safe

TABLE OF CONTENTS

INTRODUCTION

In today's global Internet economy, nearly continuous 24 x 7 operation has become a basic business necessity, and even a brief server outage can quickly compound into thousands or millions of dollars in lost revenue and missed opportunities. Don't wait for a downtime disaster—learn how Oracle Fail Safe ensures high availability (through fast failover) for databases and applications deployed on Windows NT and Windows 2000 clusters. Understand the key concepts and features behind highly available e-business solutions configured with Oracle Fail Safe. Find out why armed forces, breweries, call centers, courier and postal services, hotels, energy utilities, financial services, government agencies, insurance agencies, law enforcement agencies, manufacturing plants, health care providers, retail chains, telephone services, and warehouses—thousands of customers—all use Oracle Fail Safe today!

BASIC CONCEPTS

The goal in designing a highly available e-business solution is to eliminate as many potential points of failure as possible within a reasonable cost constraint (typically a function of the cost of downtime associated with the solution). One way to gain a measure of high availability is to use redundant hardware components such as RAID storage and multiple power supplies, network cards, and CPUs. Two or more such redundant systems can be combined to form a cluster. A cluster is a group of independent computing systems (nodes) that operates as a single virtual system. Clusters eliminate individual host systems as points of failure.

Microsoft Cluster Server (MSCS--included with Windows NT Enterprise Edition and the Windows 2000 Advanced Server and Datacenter releases) works with the Windows operating system to configure Windows systems into clusters. Oracle Fail Safe works with MSCS to ensure that if a failure occurs on one node of a Windows cluster, then the workloads running on that node move (fail over) quickly and automatically to a surviving node, usually in seconds.

MSCS provides a basic cluster environment, known as a shared-nothing cluster, in which disks, IP addresses, and other cluster resources can be owned and accessed through only one cluster node at a time. When a database or application fails over from one cluster node to another, ownership of the disks, IP addresses, and other cluster resources associated with the database or application is quickly and automatically transferred to the new node. Two cluster nodes cannot write to the same disk at the same time (because all access is exclusively through the single node that currently owns the disk), and a given application workload cannot scale across multiple cluster nodes. In the event of a node failure, the surviving system must be able to handle its own normal workload plus all failed over workloads.

Oracle Fail Safe is optimized for the MSCS environment and includes two main components, a server and a manager. The server component, Oracle Services for MSCS, works with the MSCS cluster software and a set of resource libraries to ensure fast automatic failover during both planned and unplanned outages. The management component, Oracle Fail

Safe Manager, provides an easy-to-use graphical interface that works with the Oracle Fail Safe server software on one or more clusters to perform configuration, management, verification, and static load balancing. Together, these components provide a rich set of features and integrated troubleshooting tools that enable rapid deployment of highly available database and e-business solutions on commodity Windows clusters.

Resource

A cluster resource is any physical or logical entity configured and managed on a cluster node that provides a service to clients. There are standard MSCS resource types (such as IP Address and Physical Disk) as well as an application programming interface (API) to create custom resource types (such as Oracle Database). Each resource type is associated with a resource Dynamic Link Library (DLL) that MSCS calls to manage the resource. High availability solutions configured with Oracle Fail Safe typically involve a variety of cluster resources configured to work together.

Time Service Resource and Quorum Resource

Each cluster has a time service resource and a quorum resource. These two resources are critical to the successful creation and operation of a cluster and deserve special mention.

The role of the time service is intuitive: it keeps the system time synchronized across the cluster nodes. For Windows NT, the time service is explicitly included as a resource in the MSCS Cluster Group. For Windows 2000, the time service is subsumed into the cluster software.

The role of the quorum resource is more complex. The quorum resource is created at the time the first node of a cluster is defined and provides a physical storage area used to log changes made to the cluster configuration database. Currently, it must be located on one of the cluster disks (future MSCS releases may support other locations). When a cluster node comes online, it attempts to gain control of the quorum resource. If the node successfully gains ownership of the quorum resource, it then creates the cluster. If another node already owns the quorum resource, then the node joins the already existing cluster. In situations when normal network communication between the cluster nodes is not possible, the node that owns the quorum resource continues operating while any nodes that cannot communicate with or gain ownership of the quorum resource immediately shutdown. This avoids potential “split-brain” situations with multiple cluster nodes operating in isolation as if each controlled the cluster.

Group

A group is a logical container that holds zero or more cluster resources; it is the minimum unit of failover. At any given time, a group and all the resources it contains are owned by and accessed through the same cluster node. Each cluster resource must be a member of exactly one group.

The relationships among the individual resources in a group are specified as dependencies. Resource dependencies determine the order in which the cluster software brings the resources online or offline. Figure 1 illustrates the typical resource dependencies within a group containing an Oracle database. In this case, the database is not brought online before the IP address and the disks that contain the data, log, and control files are online, and the listener is not brought online until the network name is online, and so forth. The opposite sequence is followed when bringing the resources offline.

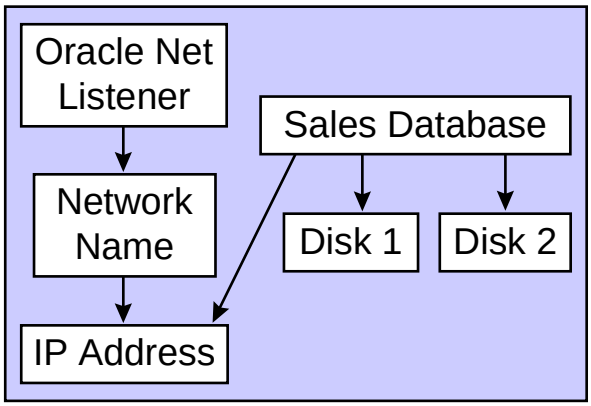

Figure1: Resource Dependencies Within a Group Containing a Database

Generally, a group should only contain resources that are in some way related to each other, such as a database resource and the related disks, IP addresses, network names, and other resources required to ensure that the database functions correctly. Note that a resource can only depend upon resources within the same group—dependencies upon resources in another group are not permitted because there is no guarantee that both groups will always exist together on the same node. Whenever possible, resources associated with different independent workloads should be configured into separate groups to ensure that a given resource failure does not unnecessarily cause a loss of service to unrelated workloads or resources.

Oracle Fail Safe Manager works with the cluster software to automate the creation, configuration, and management of resources and groups on Windows clusters. Figure 2 shows the Oracle Fail Safe Manager tree view of resources and groups on a typical cluster, while Figure 3 shows the initial screen of the Add Resource to Group wizard. When a resource is added to a group, Oracle Fail Safe performs all steps necessary to configure that resource to run in the cluster environment. A status report, like the one shown in Figure 4, records each action performed by Oracle Fail Safe and optionally can be saved to disk for future reference. If a clusterwide operation cannot complete successfully, the operation is rolled back and any problems encountered during the operation are documented in the status report.

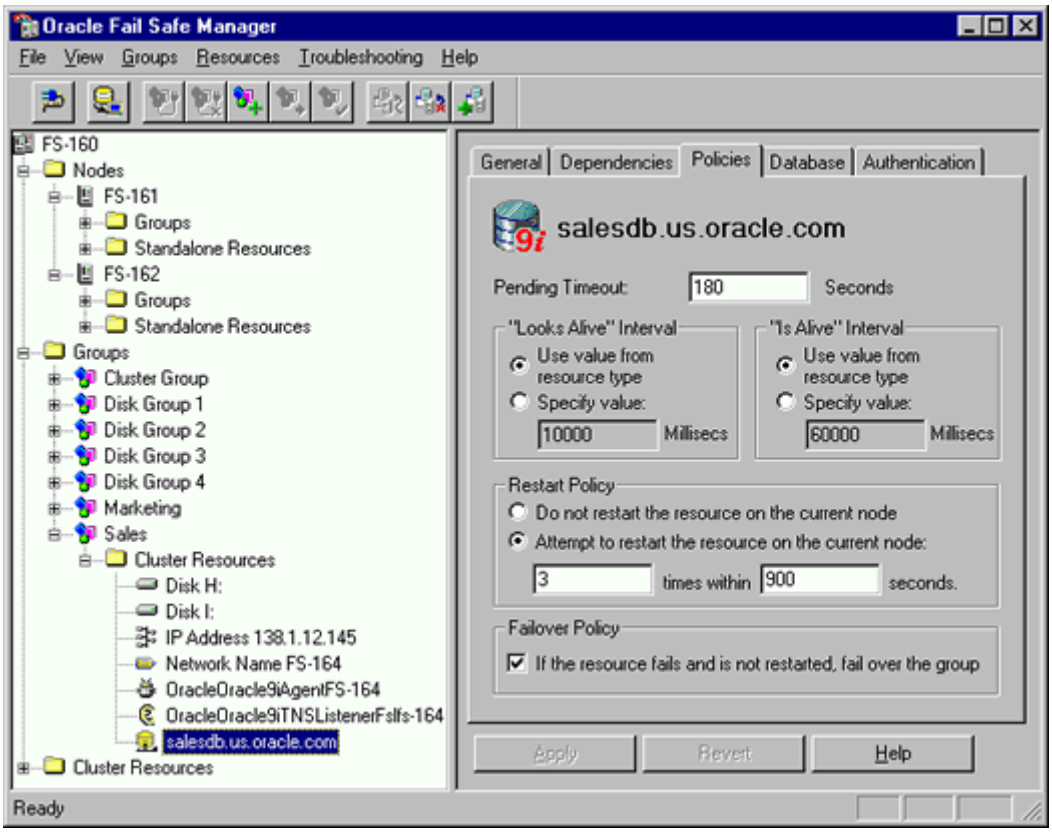

Figure 2: Oracle Fail Safe Manager Tree View of Resources and Groups

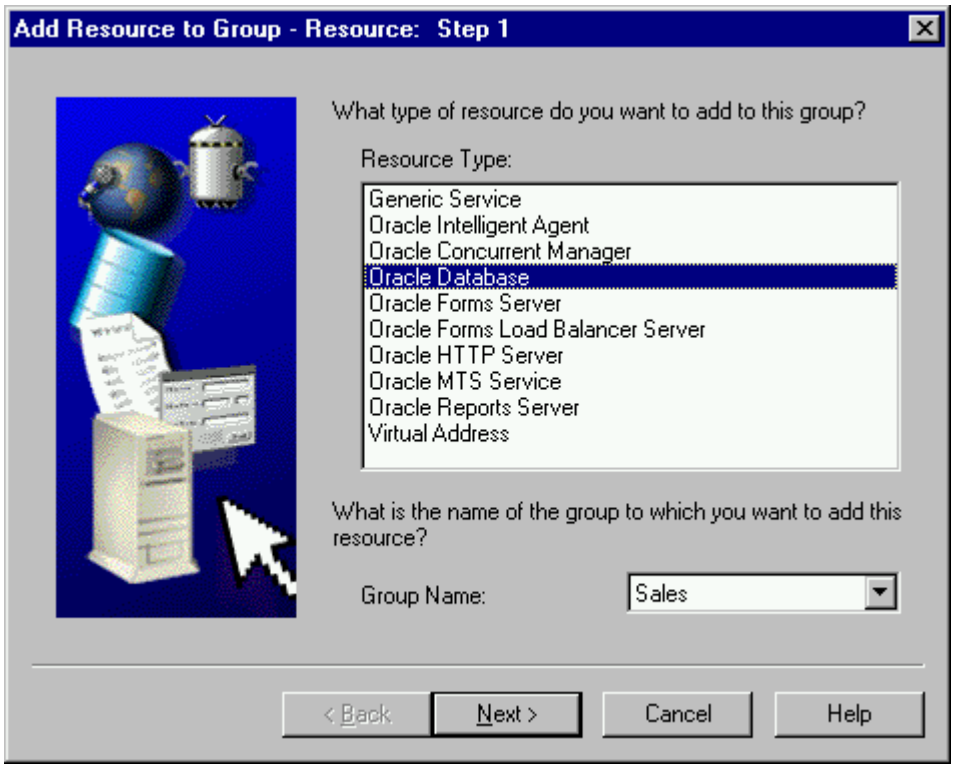

Figure 3: Oracle Fail Safe Manager Add Resource to Group Wizard

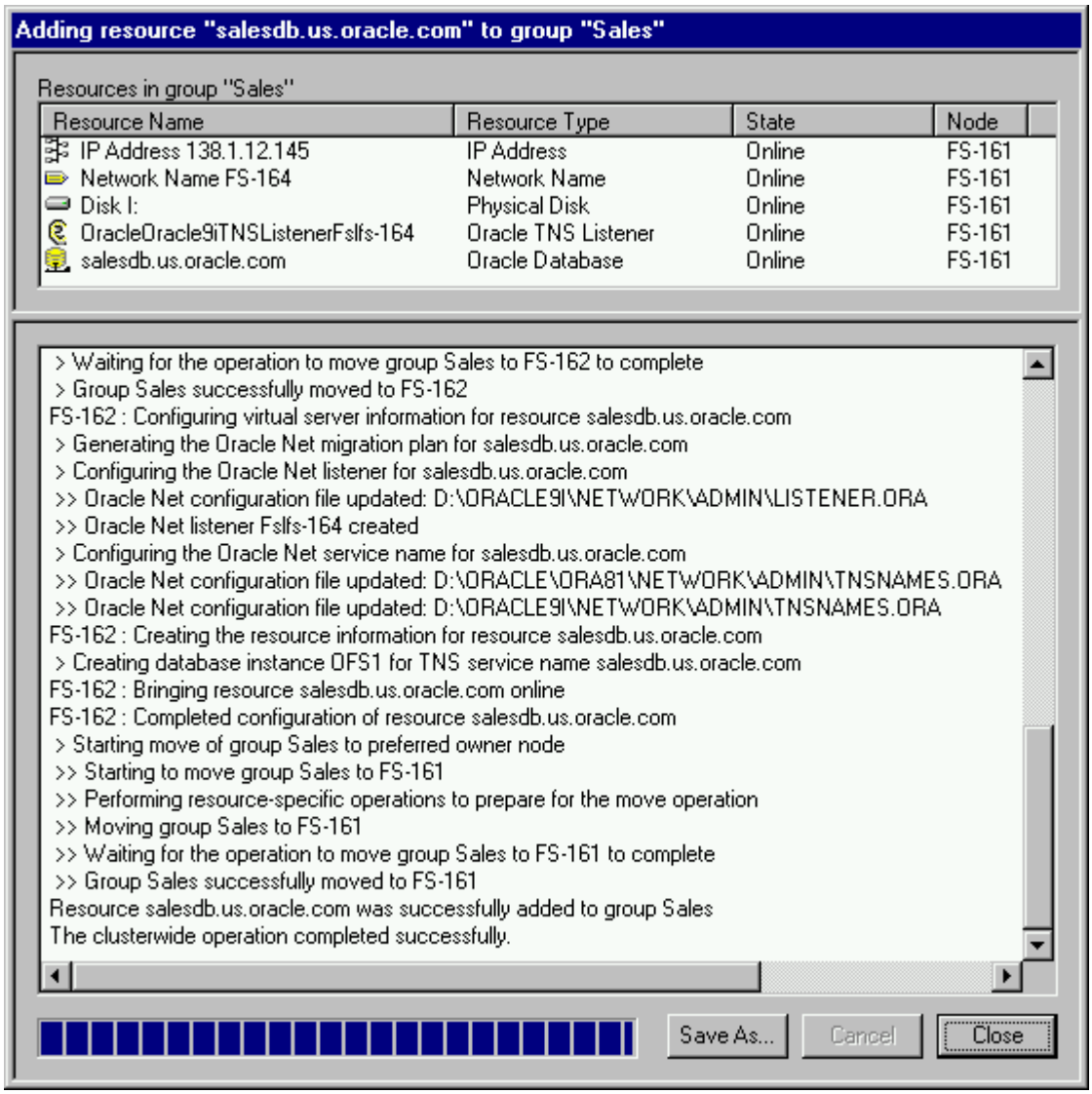

Figure 4: Add Database to Group Status Report

Failover

The process of taking a group of resources offline on one cluster node and bringing them online on another cluster node is called failover. A failover can be unplanned (automatic failover in response to an unexpected problem, such as a resource or system failure) or planned (user initiated Move Group command in response to a planned event, such as a hardware or software upgrade).

The behavior of resources and groups during an unexpected outage is governed by sets of parameters called failover policies. Failover policies are specified for both individual resources and for groups. A typical resource failover policy specifies whether or not to restart the resource on the current node if it unexpectedly fails and whether or not the failure of a resource (assuming it cannot be restarted) should affect the group that contains it. If a resource fails and the resource failover policy indicates that the resource failure should affect the group, then the group is failed over. The group policy specifies how many times an unplanned failover is allowed to happen within a given time interval before the entire group is taken offline (stops a group from failing back and forth between nodes indefinitely). Figure 2 shows typical resource policy parameters for a database and Figure 5 shows typicalfailover policy parameters for a group.

Figure 5: Group Failover Policy Property Sheet

During normal cluster operations, the cluster software periodically polls each resource (IsAlive and LooksAlive resource polling) to ensure it is functioning correctly. If a problem is detected, the actions specified in the resource and group failover policies are automatically implemented.

Planned failovers, by contrast, are initiated by the user, either interactively through the Oracle Fail Safe Manager Move Group menu command, or by executing a script that uses the Oracle Fail Safe FSCMD command line interface. Planned failovers can significantly reduce downtime during system maintenance and during hardware and software upgrades. In addition, planned failovers can be used to achieve optimal cluster load balancing. If one node becomes heavily utilized, one or more of the groups hosted by that node can be moved to a less utilized node. Load balancing can be made automatic if a detection tool, such as Oracle Enterprise Manager, is configured to automatically execute a script that redistributes workloads whenever cluster node usage becomes unbalanced (for example, if CPU resources are overutilized on one node and underutilized on another). Since clients are disconnected each time a group is moved (and may potentially lose uncommitted work), load balancing decisions are effectively tradeoffs between the cost of a brief service interruption versus the subsequent benefit of using the available cluster resources more efficiently.

Failback

Failback is the process of returning a group of resources back to their preferred node once that node comes back online. The conditions under which failback occur are specified at the group level. For failback to be enabled, a group must have a preferred node defined. Each time a node comes online, the cluster software will check to see if it is the preferred node for any of the cluster groups. Then, based on the failback policy for each group, it will move groups to the new cluster node.

The Oracle Fail Safe Manager Create Group wizard collects information to create the initial group failback policy. After initial creation, the failback policy for a group can be modified by changing the information in the Failback and Nodes tabs of the group property sheet. A group can be configured to:

- never fail back (run on whatever node is hosting it until that node goes down or the group is manually moved to another node).
- fail back immediately (move immediately to the preferred node as soon as it comes online).

- fail back between a specific time interval (move the preferred node after it comes online, but only between specific hours—for example, when there is less chance of disrupting client sessions).

Figure 6 shows the failback properties for a typical group. In this example, failback is scheduled to occur only between 3AM and 4AM after the preferred node comes online. If the preferred node came online at 6AM, failback would not occur until between 3AM and 4AM the next day.

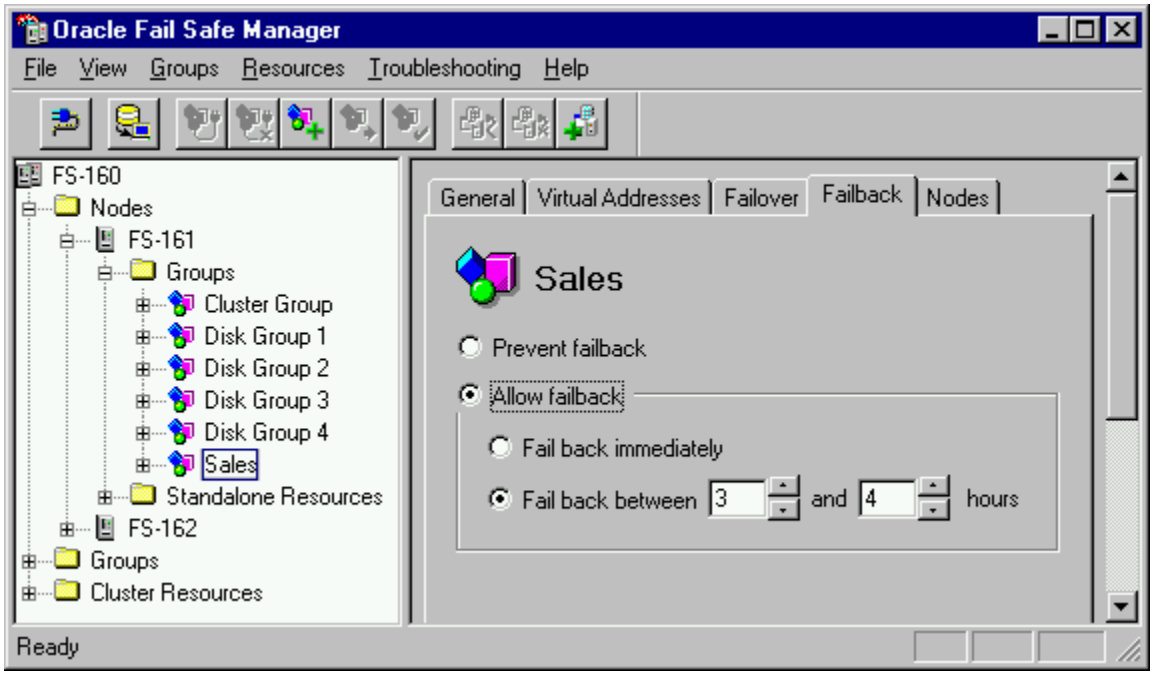

Figure6: Failback Policy Parameters for a Typical Group

ADVANCED CONCEPTS

Group, resource, failover, failback, and the various cluster configurations described in the previous section are basic concepts that apply to any high availability solution deployed on a shared-nothing cluster. The following section describes more advanced cluster concepts that have features or implementations specific to Oracle Fail Safe solutions.

Virtual Address

A virtual address is a fixed node-independent network address (network name and associated IP address) through which clients can access the resources in a group regardless of the specific hardware server hosting those resources. The Oracle Fail Safe Manager Add Resource to Group wizard, shown in Figure 7, can be used to add one or more virtual addresses to a group.

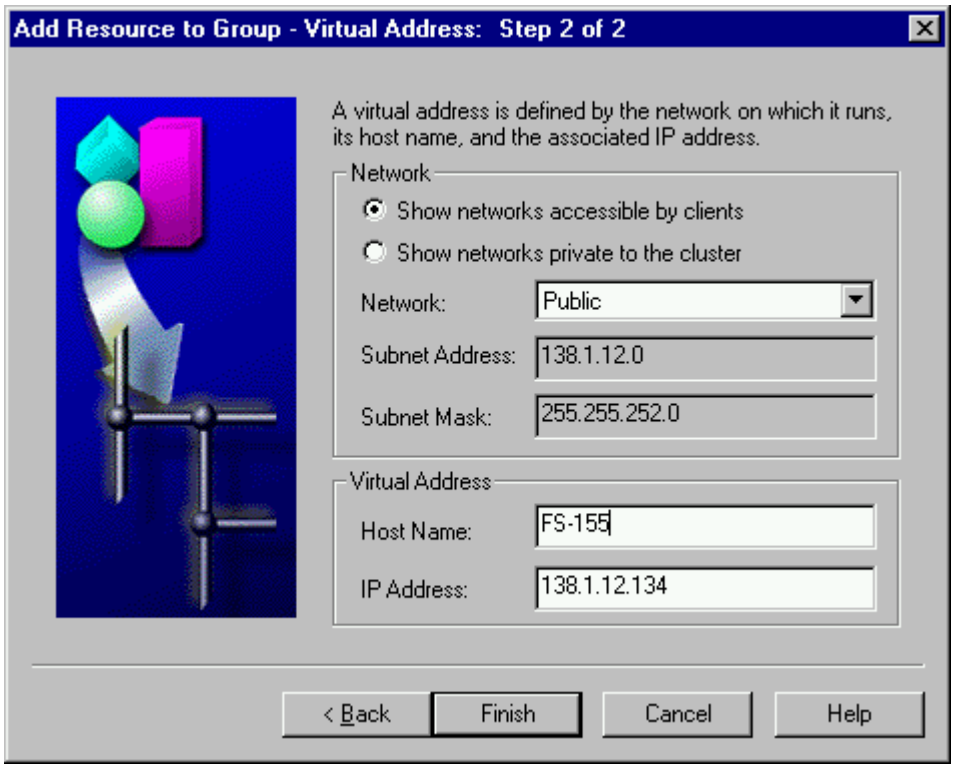

Figure 7: Add Virtual Address Resource to Group Wizard

Virtual Server

A virtual server is a logical server available to clients at a fixed IP address and configured to operate on any one of a set of physical servers. For MSCS clusters, a virtual server is implemented as a group with one or more virtual addresses. At any given time, a client user cannot tell which physical cluster node is hosting a virtual server.

Oracle Fail Safe automates the creation of virtual servers for databases, Oracle Forms and Reports Services, Oracle HTTP Server, and other applications. It takes only three wizards to configure a highly available e-business solution:

- Create Group wizard to create a group.
- Add Virtual Address wizard to add one or more virtual addresses to the group.
- Add Resource to Group wizard to add databases, Forms and Reports Services, Oracle HTTP Server, Oracle Applications Release 11*i* components, and other resources to the group to complete the solution.

Clients then use a fixed highly available virtual address instead of a node-specific address to access the resources configured to run in the virtual server environment. If a failure occurs on one node, the virtual server automatically fails over to a surviving cluster node. Clients then reconnect using the same virtual address and can quickly resume work. For database clients, the Oracle Call Interface transparent application failover features can make reconnection automatic and can automatically replay interrupted SELECT statements. For many users, a virtual server failover may not even be detectable or may appear only as a brief pause in activity, giving the illusion of continuous availability. Figure 8 shows client access before and after a failover to a virtual server that hosts a highly available database.

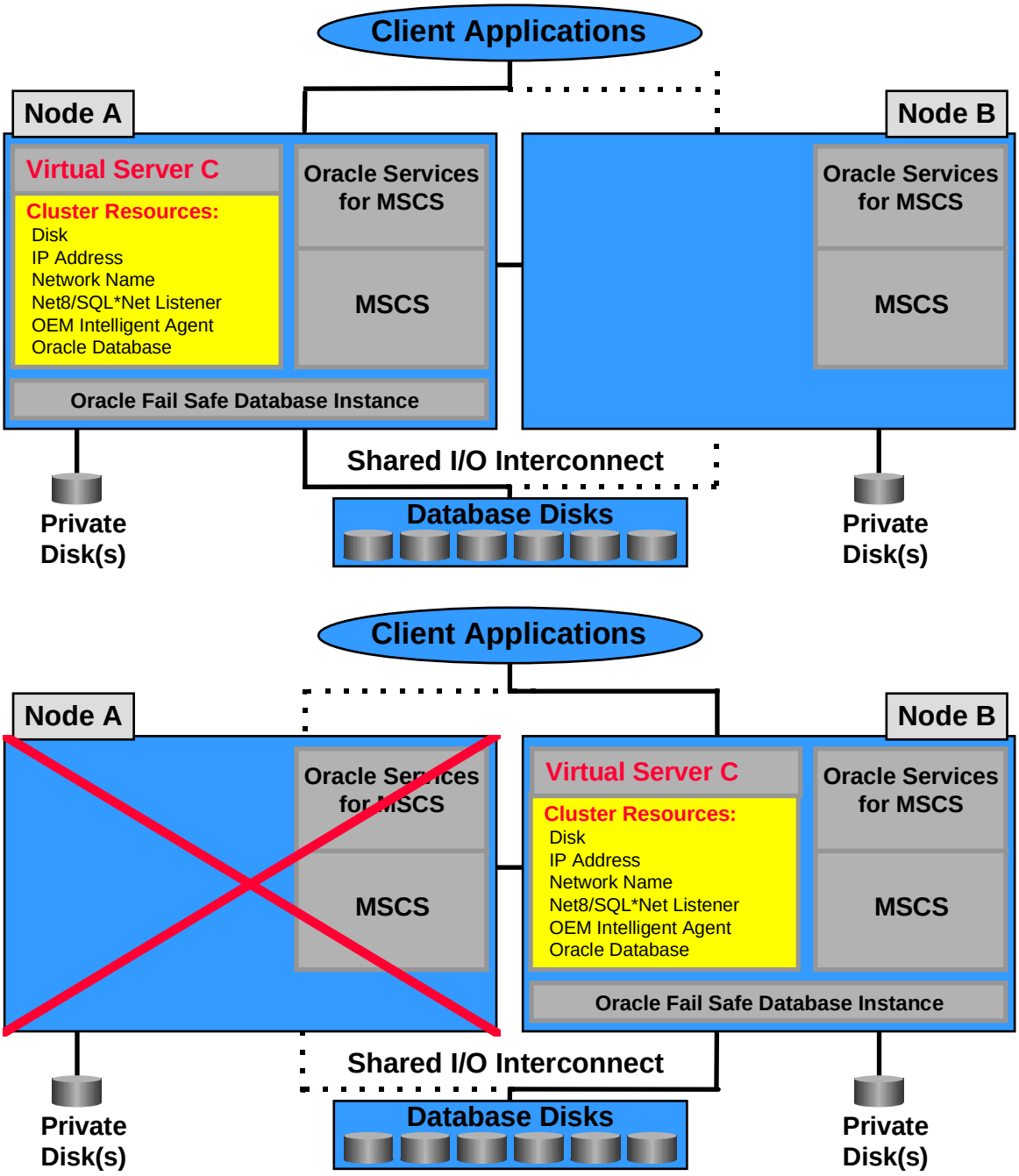

Figure 8: Client Database Access Through a Virtual Server Before and After a Failover

Clusterwide Operations

Many of the configuration and management operations performed through Oracle Fail Safe Manager take place across multiple cluster nodes. Oracle Services for MSCS coordinates the work across the cluster nodes to perform these clusterwide operations and to ensure that all clusterwide operations roll back if problems are encountered, returning the cluster to the original state. Examples of clusterwide operations include adding a resource to a group or using any of the troubleshooting tools (Verify Cluster, Verify Group, or Verify Standalone Database). The Oracle Fail Safe

Manager console displays the ongoing status of clusterwide operations across all nodes. Figure 9, for example, shows a typical status report displayed during a Verify Group clusterwide operation.

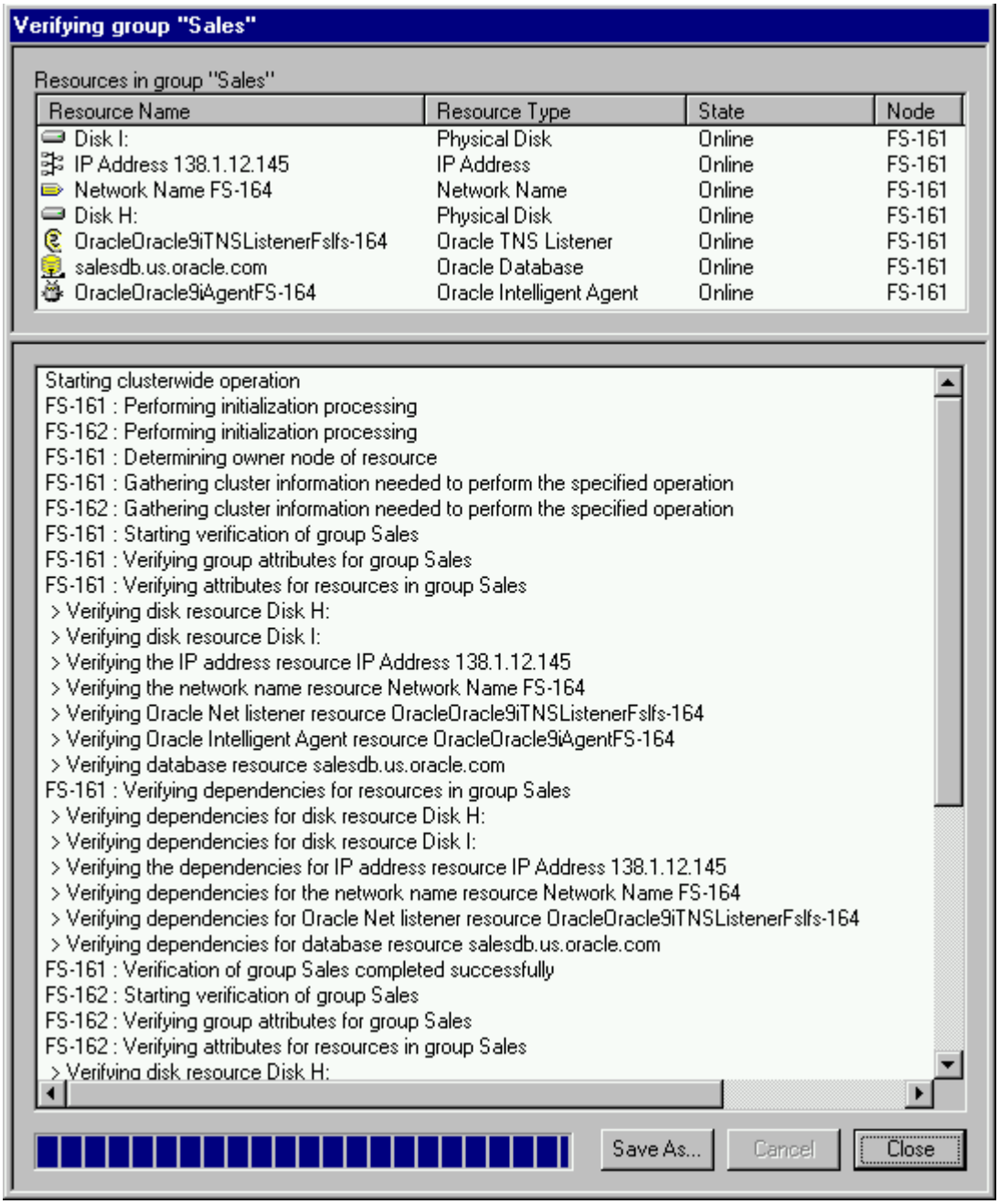

Figure 9: Oracle Fail Safe Manager Report Showing Ongoing Status of a Verify Group Operation

Rolling Upgrades

A key benefit of all Oracle Fail Safe high availability solutions is the ability to minimize downtime through rolling upgrades. During initial installation and configuration, separate copies of all program executable files are placed on a private disk on each cluster node and only data files are located on the cluster disks. This allows administrators to upgrade the program executable files on one cluster node while users continue working on the other node using a different copy of the program executable files. Once the first node is upgraded, users can be failed back to it, and the

rest of the upgrade can be completed while users again continue working. Some downtime may be unavoidable in situations where data files need to be modified as part of an upgrade. However, with careful planning, rolling upgrades can eliminate significant downtime from many common maintenance operations, including:

- Rolling upgrades of the Windows operating system
- Rolling upgrades of system hardware
- Rolling upgrades of Oracle Fail Safe
- Rolling upgrades of database and application software

Non-Oracle application and database solutions that place both program executable files and data files on the cluster disks do not have this benefit, because users will experience potentially long periods of downtime each time any application or database software component is upgraded.

Configuration Change Management and Troubleshooting Tools

The Oracle Fail Safe Manager Troubleshooting menu commands (shown in Figure 10) and the FSCMD commandline parameters DUMPCLUSTER, VERIFYGROUP, and VERIFYALLGROUPS make it easy to diagnose and repair problems and to proactively monitor the health of cluster resources. In addition, they can be used to automatically reconfigure cluster resources when changes (such as adding more disks to a database or adding a new cluster node) are made after the initial cluster setup and configuration steps have been completed. Detailed information on how to best use each of these tools is provided in the online help and documentation included with Oracle Fail Safe.

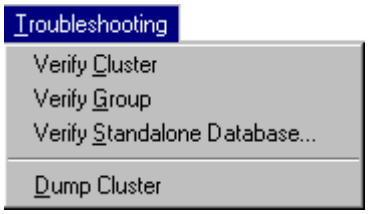

Figure 10: Oracle Fail Safe Manager Troubleshooting Menu Commands

CLUSTER ARCHITECTURE AND CONFIGURATION

Architecture

A detailed understanding of the internal architectures of Microsoft Cluster Server and Oracle Fail Safe is not required to use or deploy highly available database or application solutions. Oracle Fail Safe wizards automate most cluster and network configuration tasks and ensure that databases and applications will operate correctly on the cluster. However, some familiarity with key cluster concepts and architecture features is helpful when deciding how to deploy a solution for optimal high availability or when troubleshooting unexpected behavior. Figure 11 shows the major components of Windows, MSCS, and Oracle Fail Safe. Oracle Enterprise Manager (optional) and Oracle Fail Safe Manager (required) are usually installed on a separate (non-clustered) administration console, while the other components are installed on each node of the cluster.

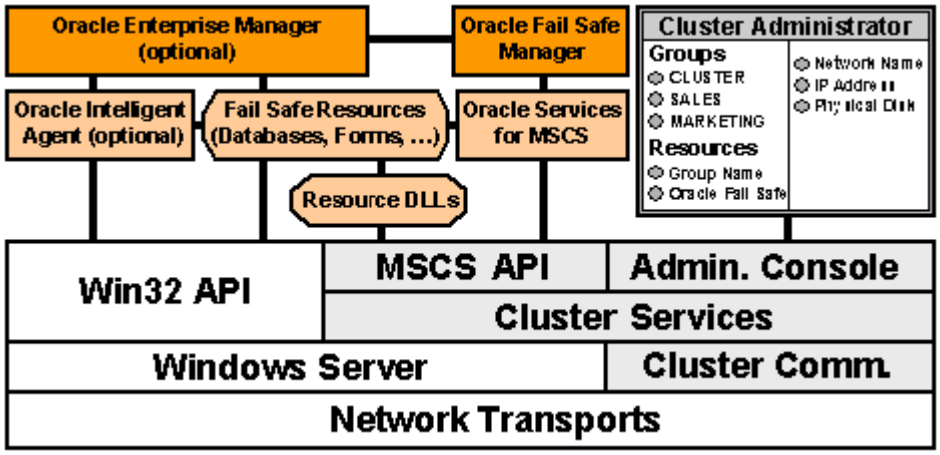

Figure 11: Oracle, MSCS, and Windows Software Component Architecture

Microsoft Cluster Server includes its own administration console, Microsoft Cluster Administrator. However, because Cluster Administrator has no intrinsic knowledge of Oracle databases and applications, the suggested way to configure and administer Oracle software components on MSCS clusters is to use Oracle Fail Safe Manager.

Clusters designed for use with MSCS are generally similar in configuration. They differ primarily in price, vendor-specific added value features, and storage interconnect (typically either SCSI or fibre channel). SCSI clusters are generally less expensive than fibre channel clusters, but all nodes and storage arrays must all be located within a few feet of each other due to SCSI cable length restrictions. Fibre channel clusters allow individual nodes or storage arrays to be separated by distances of 10 kilometers or more, providing a measure of additional (though still limited) disaster protection. In addition, SCSI clusters are limited to two nodes, while fibre channel clusters can have more than two nodes (for example, Windows 2000 Datacenter clusters support fibre channel clusters with up to four nodes). Most clusters are configured with both a public network connection (for client access) and a private internode network connection (for intra-cluster "heartbeat" communications). Figure 12 shows the major physical components of a typical fibre channel cluster.

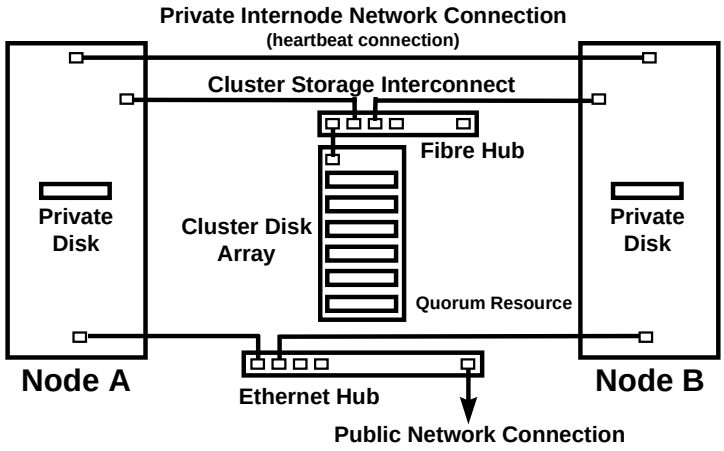

Figure 12: Fibre Channel Cluster Hardware Components

Cluster Configuration

Most Windows clusters are configured similarly, differing primarily in choice of storage interconnect (SCSI or fibre channel) and in the software deployed on the cluster nodes. Oracle Fail Safe solutions can be deployed on any cluster configuration listed on the Microsoft Hardware Compatibility List.

Each cluster node must have Windows NT Enterprise Edition or a release of Windows 2000 that supports clustering installed along with Microsoft Cluster Server. Product-related files that must be accessible to either cluster node (for example, the data, control, and log files associated with a database instance) are installed on cluster disks attached to the shared storage interconnect between the nodes. Oracle Services for MSCS and all other executable Oracle software are then installed on a private disk (usually the system disk) on each node. In addition, Oracle Fail Safe Manager is usually installed on the management system for the Windows domain containing the cluster.

Most Oracle Fail Safe solutions can be described as active/passive, active/active, or multitier. For all of these configurations, RAID storage is recommended to prevent downtime or loss of data due to media failure. Using separate RAID arrays for the quorum resource and for each independent workload (application or database) further minimizes the potential effects of a data corruption or disk failure. In addition, by configuring each independent workload to use a separate group, a resource failure in one group will not unnecessarily affect the unrelated resources in another group. The key features and benefits of these three cluster configurations are summarized in the next sections.

Active/Passive Software Configuration

In an active/passive configuration (see Figure 13), one cluster node performs all work while the other node stands by to pick up work in the event there is a failure on the first node. This configuration is less expensive than traditional standby solutions (no second disk farm or data replication costs) and provides the fastest failover response, because the entire second node is immediately available to process a failover.

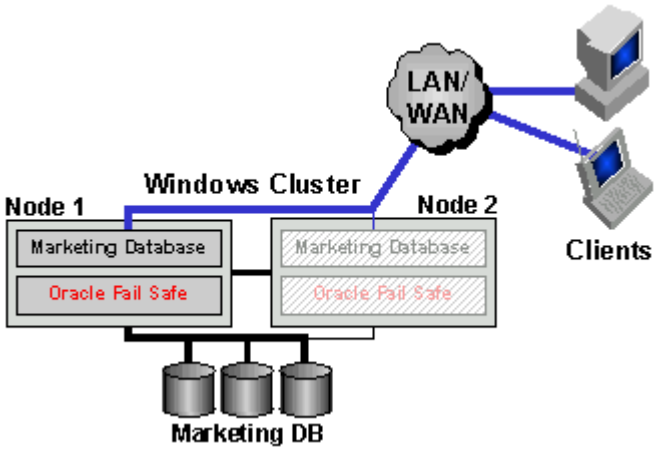

Figure 13: Typical Active/Passive Database Configuration

Active/Active Software Configuration

In an active/active configuration, both nodes perform useful work. There are tradeoffs between the resources used by the workloads normally running on each node and the response time during and after a failure. Several performance optimizations are possible that allow users to get more work out of active/active configurations during normal operations, yet still ensure that critical workloads will fail over successfully. For example:

- Only the critical workloads on a system (such as a database) may be configured to fail over.
- A database can use a different parameter file on each node for optimal tuning and configuration.

Another benefit of active/active configurations is that when each node runs at around 50% capacity, there is plenty of additional capacity available on each node to handle transient workload spikes. For active/passive configurations, by contrast, the active node typically runs much close r to full capacity and has only minimal additional resources available. Figure 14 illustrates a typical active/active configuration with a Marketing database normally running on one

cluster node and a Sales database running on the other cluster node, each configured for high availability using Oracle Fail Safe.

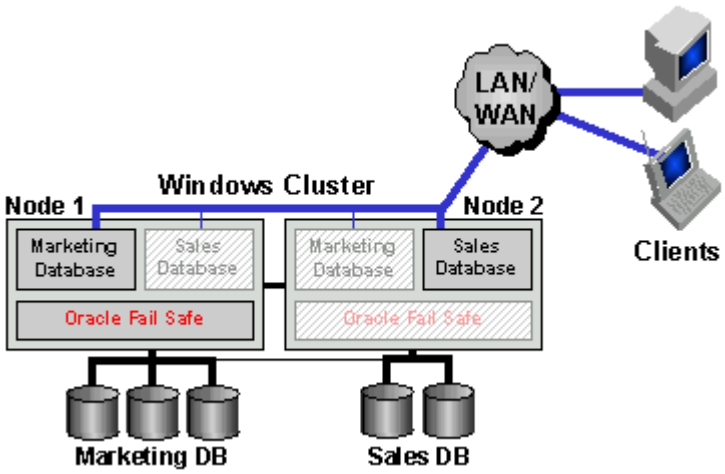

Figure 14: Typical Active/Active Cluster Configuration

Multitier E-Business Solutions

Oracle Fail Safe provides wizards to automate the configuration of both application and data tier software components and supports many flexible deployment options for Windows-based e-business solutions. Figure 15 shows a highly available Oracle Reports solution, for example, that combines both clustered and standalone systems. As reporting requirements grow, more slave Reports Servers can be added, allowing the solution to scale. Oracle Fail Safe eliminates potential points of failure in both the application and data tier, such as the master Reports Server and Database Server, to ensure that end users, as much as possible, have the illusion of zero downtime. For critical reporting functions, availability can be enhanced even further by configuring each slave Reports Server with Oracle Fail Safe on a cluster. If a failure occurs, both the slave Report Server and its associated job queue will fail over together, ensuring that any previously queued reports will execute as scheduled on the surviving node.

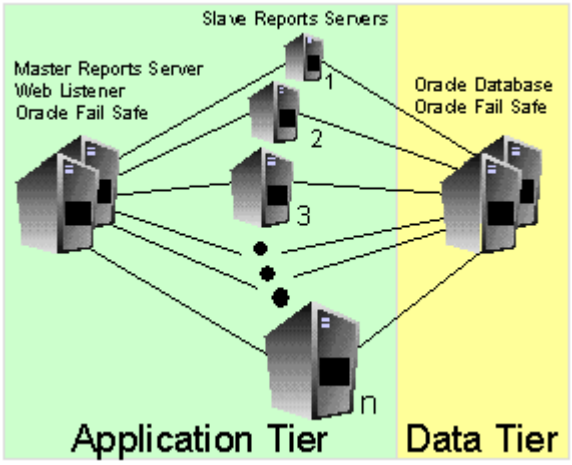

Figure 15: Scalable Highly Available Multitier Reporting Solution

Similarly, Oracle Applications Release 11*i* e-business solutions can be configured for high availability using Oracle Fail Safe. Figure 16 shows a typical multitiered configuration. In this case, the database, Reports Server, Concurrent Manager, Web Server, and Forms Load Balancer Server components are all configured with Oracle Fail Safe on

clusters to ensure high availability and to eliminate what would otherwise be potential points of failure in the overall solution. Additional standalone Forms Servers can be added as needed to provide application tier scalability.

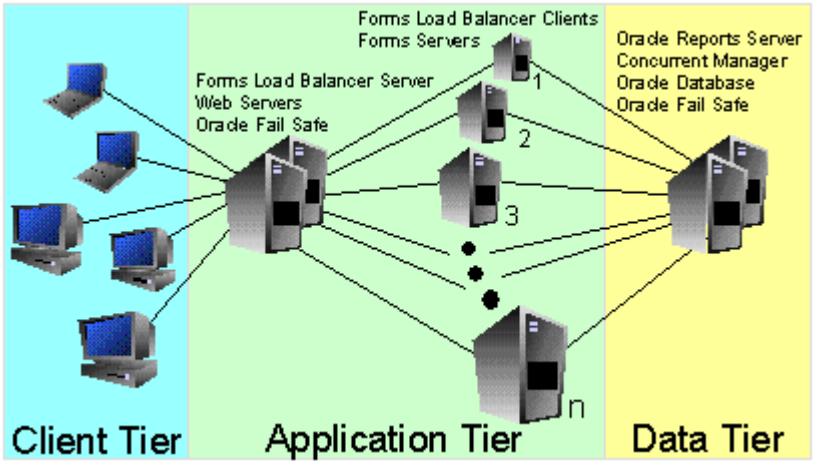

Figure 16: Scalable Highly Available Oracle Applications Solution

Economies of Scale

Oracle Fail Safe supports Windows 2000 Datacenter clusters with up to four nodes. Larger clusters can provide significant economies of scale. Replacing a collection of two-node clusters with a single Windows 2000 Datacenter Server cluster can substantially reduce the hardware cost associated with otherwise "idle" systems. Management and administration tasks are also consolidated into a single cluster environment. Figure 17 shows a 4-node Windows 2000 Datacenter cluster configured so that a single cluster node serves as the backup system in the event that any of the other nodes fails. By contrast, if each of the three workloads is instead deployed on its own separate 2-node active/passive cluster, then three of the six total systems would be idle.

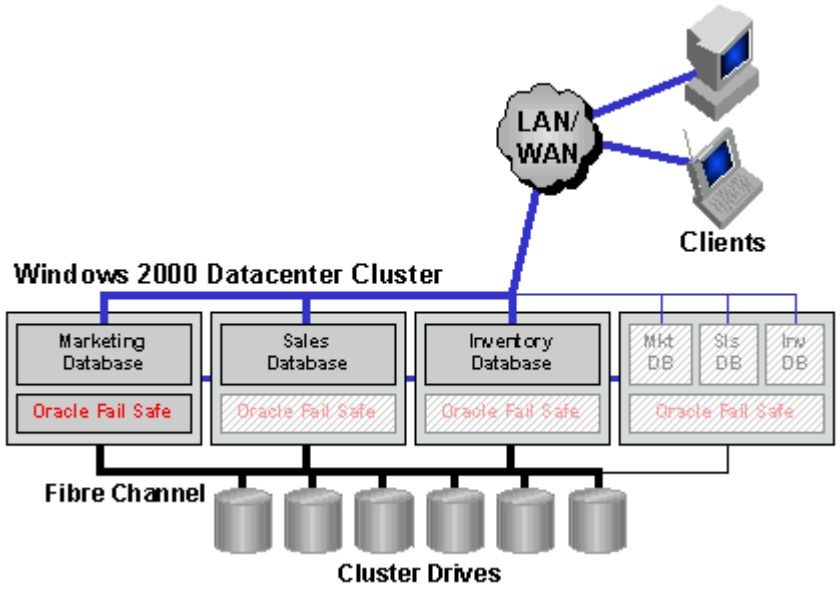

Figure 17: Economies of Scale with Larger Clusters

SUMMARY

Oracle Fail Safe minimizes or eliminates many potential sources of downtime. By understanding key concepts such as virtual server, failover, and failback, developers and administrators can design and implement a large variety of highly available application and database solutions on Windows clusters. The automated wizards and configuration tools included with Oracle Fail Safe make deployment and maintenance easy and error-free. For many customers, the client downtime prevented during a single server outage provides an immediate return of investment for their entire high availability solution.

For more information, refer to the following online sources:

TOPIC	ONLINE LOCATION
High Availability	http://www.oracle.com/ip/deploy/database/oracle9i/index.html?ha_home.html
Oracle Databases	http://www.oracle.com/ip/deploy/database/oracle9i/index.html?content.html
Oracle Fail Safe	http://www.oracle.com/ip/deploy/database/features/failsafe/ http://technet.oracle.com/tech/windows/failsafe/

Technical Introduction to Oracle Fail Safe:
A Guide to Concepts and Terminology
October, 2001
Author: Laurence Clarke

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
www.oracle.com

Oracle Corporation provides the software
that powers the internet.