Building Highly Available Database Servers Using Oracle Real Application Clusters

An Oracle White Paper
May, 2001

ORACLE

Building Highly Available Database Servers Using Oracle Real Application Clusters

EXECUTIVE OVERVIEW

High Availability (HA) is becoming a requirement for e-businesses that cannot afford system down time. The Internet makes it easier to reach customers around the world and around the clock. Since companies must always be prepared to serve their customers, this expanded reach also makes it more costly when service is not available, either planned or unplanned.

ORAC (Oracle Real Application Clusters) is the multi-node extension to Oracle database server that enables e-businesses to build database servers across multiple nodes that are highly available and highly scalable. An e-businesses does not need to sacrifice scalability or performance for high availability with ORAC.

This white paper presents the architecture that makes ORAC databases highly available while functioning as highly scalable database servers. It starts with the basic availability issues faced by a generic cluster system, followed with a discussion on how ORAC addresses those issues. The paper continues by discussing enhancements in ORAC which distinguish it from generic cluster systems on availability capabilities. The paper also discusses limitations associated with using cluster systems as HA solutions and gives suggestions on how to address them.

This white paper also addresses practical issues with using ORAC, including planning, installation and configuration. It discusses options for applications to use ORAC so that the applications can be made more available when building a highly available application system.

This paper also shows why ORAC, with active instances on all nodes, is also a much better choice for database fail-over solution than those offered at operating system level for generic application fail-over.

Finally, it demonstrates how Oracle Corporation has used ORAC to save costs and expand into new markets in two typical cases. First, Oracle uses ORAC to consolidate its global email systems that makes it much more robust and performing. Oracle also uses the same architecture for its email hosting business. Second, Oracles uses ORAC as the central repository for Oracle hosted exchanges which offer Internet based business services.

ARCHITECTURE OF REAL APPLICATION CLUSTERS

Typical Configurations Supported by Real Application Clusters

Real Application Clusters (RAC) runs on top of a hardware cluster. A cluster is a group of independent servers (nodes) that cooperate as a single system.

The primary cluster components are processor nodes, a cluster interconnect, and a shared storage subsystem. The nodes share access to the storage subsystem and resources that manage data, but they do not physically share main memory in their respective nodes. ORAC combines the memory in the individual nodes to provide a single view of the distributed cache memory for the entire database system.

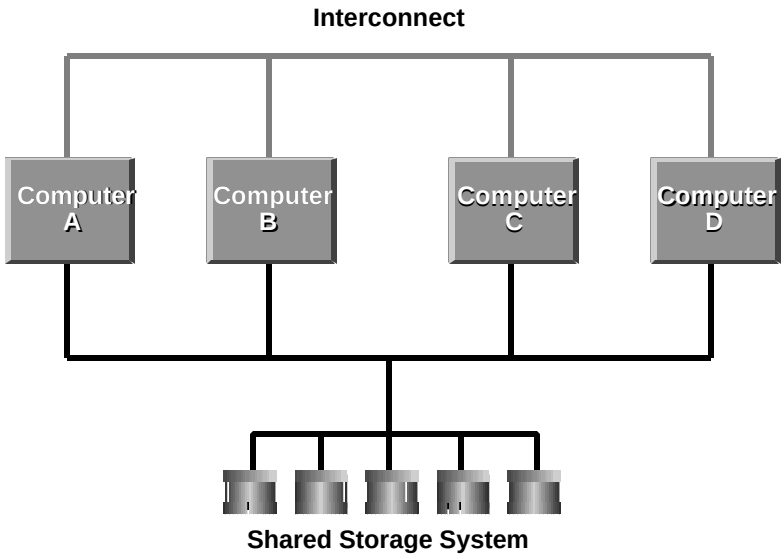

Figure 1. A Cluster Comprises Processor Nodes, The Cluster Interconnect, and A Disk Subsystem

A node can be made up of multiple processors. A common type of node is a Symmetric Multi-Processor (SMP) node. Each node has its own dedicated system memory as well as its own operating system, database instance, and application software.

Advantages of Clusters

The benefits of using clusters versus a single larger node are-

Flexibility and cost effectiveness in capacity planning, so that a system can scale to any desired capacity

Fault tolerance to partial failures within the cluster, especially the node failures.

Scalability

ORAC gives users the flexibility to add node(s) to the cluster as the demands for capacity increases, scaling system up incrementally to save costs in capital investments and eliminating the need to replace smaller single node systems with larger ones. It makes capacity upgrade process much easier and faster since, in most case, one or more nodes with similar or identical configuration are added to the cluster, compared to using complete new and larger nodes to upgrade systems.

The Cache Fusion technology implemented in ORAC enables capacity to be scaled up close to linearly. The Cache Fusion technology is described in details in the Cache Fusion technical white paper.

High Availability

Another main advantage of the cluster architecture is the inherent fault tolerance provided by multiple nodes. Since the physical nodes run independently, the failure of one or more nodes should not affect other nodes in the cluster. In the extreme case, a cluster system can still be available even when all but one node survives, making a system based on cluster highly available. This architecture also allows group of nodes to be taken off-line for maintenance while the rest of the cluster continues to provide services online.

ORAC takes full advantages of this inherent fault tolerance architecture to provide a highly available database server.

The following sections describe how ORAC achieves high availability for Oracle databases.

Availability Framework of Cluster Systems

Introduction

To take full advantage of the fault tolerance afforded by cluster architecture, ORAC enables the Oracle database server to function in the face of various failure scenarios in the cluster. Furthermore, ORAC is able to recover failed nodes while the database server is online.

Before going into details of the availability framework for cluster systems, we need to understand the differences between availability within a single node and within a cluster. In a single node Oracle database system, availability refers to the ability to survive various application and operation failures within the database instance. In the extreme case of the failure of the node, availability refers to the ability to recover the database to a transaction consistent state as fast as possible. Extensive discussion on this subject is provided in the white paper for Oracle Database High Availability.

For a cluster system, aside from handling failure scenarios in a single node, it needs to handle failure scenarios associated with node or network , while providing required performance. ORAC builds on top of the fault tolerant

capabilities of the single instance Oracle database and enhance the database server to handle failure scenarios unique to a cluster system.

Common Failure Scenarios in a cluster system

The flexibility of a cluster system for each node to function relatively independently does come with some unique problems the cluster system must deal with when failure occurs.

Fault Isolation

The cluster system maintains a consistent system image at all times, especially during failures of individual nodes or the cluster interconnect. The biggest challenge for the cluster system is to be able to quickly and reliably isolate faults and take corresponding actions.

For example, in the case of network failure, a cluster with many nodes can end up as isolated groups of connected nodes. The cluster system must be able to decide that this condition is due to network failure and be able to decide which connected group of nodes (sub-cluster) should continue to operate for the cluster, and which should be temporarily retired from the cluster. This decision is critical to prevent the cluster system from developing the “split brain” syndrome in which different isolated groups of connected nodes (sub-clusters) all claim to represent the whole cluster and all working on the same set of data, unaware of each other’s continuing existence.

Recovery

Since a failed node or nodes may contain global information to the whole cluster, the cluster system must be able to re-construct the information as quickly as possible. It can either maintain a hot standby repository for the global information, or re-create the information from the live nodes and the information stored in the storage system. ORAC’s fast recovery is especially crucial to maintain high availability for Oracle database servers. This fast recovery capability is discussed in more depth later in the ORAC recovery section .

IO Fencing

IO fencing refers to the situation that occurs when the left-over write operations from failed database instances (cluster function failed on the nodes, but the nodes are still running at OS level) reach the storage system after the recovery process starts. Since these write operations are no longer in the proper serial order, they can damage the consistency of the stored data. ORAC utilizes facilities provided by the underlying cluster system to prevent this.

Limitations of the Cluster System’s Availability Capability

As highly available as a clustered system is, it is not the solution to all availability problems. A cluster’s availability is mostly a system solution that enables a system

to function when some of its components fail, such as its instances, physical nodes, or cluster interconnect. To maintain a highly available system, other precautions must be taken.

For example, redundant hardware, such as separate power supplies for each of the sub components or redundant interconnect hardware, can help to avoid a problem caused by a failure of one of these components.

A cluster system cannot prevent failures resulting from operator errors either. Every precaution much be taken to guard against these user errors, just like in the case of a single database instance.

Two types of these additional failures need special attention.

Disaster Recovery

Cluster systems can not be used to guard against disasters that completely shut down a site. Disasters such as earthquakes, fires, power outages or flooding that destroy a physical site can not be guarded using cluster systems at the same site. Solutions for database disaster recovery such as a standby database, either physical or logical, provide the right protection against such failures. For details about Oracle's offerings in this area, please refer to relevant white papers.

Storage System Protection

Since the one of the intrinsic functions of a cluster system is for cluster nodes to utilize a shared storage system, the cluster system itself does not guard against failures of the shared storage system. Even though a plain bunch of disks will work with a cluster system, a more sophisticated storage system that is highly reliable and available should be used as the shared storage sub-systems for cluster systems to guard against storage related failures.

Availability Framework of ORAC

Architecture

approximately depicted in *Figure 2*. Cluster ware is the component that provides

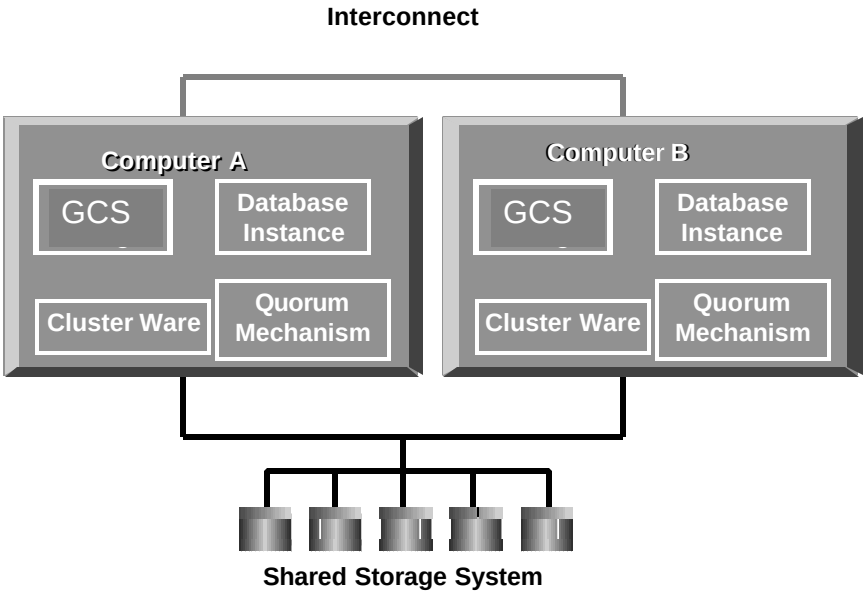

Figure 2. ORAC software components

the generic cluster functions at the operating system level. Cluster ware monitors and handles group membership related events, such as when to include a node into or exclude a node from the cluster when a node starts or fails.

The Quorum Mechanism is used by ORAC to enhance failure detection at database level. For example, cluster ware does not know if a database instance crashes.

Cluster Ware in conjunction with Quorum Mechanism provides reliable fault detection and isolation at both OS level and database level. Fault scenarios are more complicated than a generic cluster system with database as part of it. For example, when a network failure occurs, both Cluster Ware and Quorum Mechanism need to work together to ensure both the right nodes and instances are chosen to act for ORAC. Another example would be when a database instance fails, Quorum Mechanism and Cluster Ware need to coordinate to make sure no left over writes from the failed instances reach the storage system after the recovery process starts.

GCS (Global Cache Service) insures that a consistent single database image is maintained. It maintains the consistency of the database at the cluster level. Database blocks accessed concurrently by cluster nodes have corresponding GCS resources to insure the same data block is not updated without coordination by

different nodes. When the information is lost during to node failure, it must be reconstructed first before anything else within the database can proceed.

Database Recovery in ORAC environment

The Database Instance relies on all three of these components to implement instance recovery for failed instances, in addition to handling normal database operations.

When a database instance or a node in the cluster fails, ORAC needs to do recovery for the database just as it does for a single instance database. Since other nodes in the cluster is still providing services to the clients, ORAC makes sure the recovery time is as little as possible through concurrent recovery operations.

The global cache service (GCS) maintains global cache resources status to insure the consistency of database. If a node fails, it needs to rebuild the global cache resource information. Recovery cannot start until the GCS finishes rebuilding the information. Effectively, the whole database is "frozen" during this time. To reduce the time needed for rebuilding GCS information, the cache resources for database blocks are distributed among the cluster nodes. Only the cache resources that reside or are mastered by the GCSs on the failed nodes need to be rebuilt or re-mastered.

Furthermore, Oracle9*i* uses a two-pass database recovery scheme, where the first pass of redo log scan decides data blocks to recover and then the second pass only accesses the marked blocks to speed up instance crash recovery. Oracle9*i*, can initiate the first pass of this recovery process concurrently with GCS rebuild process. After the first pass, database is made available for service for data blocks that are not impacted by the failure.

Oracle9i also gives you the ability to specify the amount of time a recovery should take, which eliminates potential problems caused by uncertain times for recovery.

Fail over of TCP/IP Connections

To ensure fast fail over of clients connecting to the failed nodes/instances, database application clients can connect to ORAC through Oracle Net using load balancing and application fail-over. Client connection load balancing distributes client connections to all nodes of the clusters, alleviating the impact of an individual node failure. With the Transparent Application Failover (TAF) option, Oracle Net will re-connect the failed connections to the fail-over node in the cluster without the client even being awareness of the failure.

Today, most client connections are made using the TCP/IP networking protocol. Since TCP/IP is a reliable communication protocol for wide area network, it ensures every message is either reliably delivered or a failure event is generated. This can cause problem for ORAC during node failure. In some condition, if a message is sent to a node that fails and no response is generated back to the client

before the failure, the TCP/IP protocol will enforce a time-out period of up to a couple hours before reporting the error condition. This scenario can leave the impression of a hanging client.

To address this problem, ORAC floats the IP address of the failed nodes to a live node as part of the recovery procedure, ensuring that requests even to the failed node have a destination so that clients do not need to endure long time-outs.

Online Configuration

To minimize the impact of ORAC configuration on availability, nodes can be added to or taken out of an ORAC cluster without the database server being shut down.

Though DBCA (database configuration assistant), a new instance can be added to or deleted from an existing ORAC cluster while the database is online. The GUI-based utility makes a complicated operation of adding and deleting a node online much simpler.

BUILDING HIGHLY AVAILABLE DATABASE SERVERS

Understanding the availability features of ORAC lays a common ground for understanding the issues involved in constructing highly available Oracle9*i* database servers. This section discusses the practical aspects of building database systems using ORAC.

The information presented here is intended to serve only as general guidelines. For detailed, step by step procedures, please refer to the ORAC document set.

Availability Requirements

The most critical steps in constructing a highly available database server are to have a clear objective, to set the right expectations, and to collect the right requirements.

As discussed in previous section, ORAC is ideal for providing system availability solutions when a site must be operational even if some nodes of the clusters are taken off-line due to maintenance or failures. Using ORAC to provide scalable database services while making it resilient to partial system failure is the best use for ORAC. Online stores, general web sites, corporate databases, and most Web based business portal fit into this category.

If the requirement is for the system to survive disasters, ORAC alone is not enough. A standby backup system that is not physically co-located is the can provide extended protection from disasters. One ORAC system acts as the main server while the other ORAC system acts as the standby in another location. These systems that are loosely coupled through database standby operations can survive most disastrous events. Considerably more resources are needed for this type of availability requirement. Highly critical applications such as banking

systems or airline reservations that are extremely time sensitive may require this type of availability.

Installation Considerations

Cluster Systems

Since cluster systems require additional hardware and software components than a single node, users are advised to pay special attention during installation. In addition to the nodes and the system software required by single instance Oracle database, a cluster requires an interconnect network to link the cluster nodes together and a storage system that is shared accessed by all nodes in the cluster. In addition, most cluster system vendors provide cluster software to manage group membership of the cluster and to monitor the cluster. This adds some complexity to the installation process before installing ORAC.

To ensure a smooth installation, ORACLE works with system vendors to certify cluster platforms as well as storage systems that support ORAC. Specific certification information is published on www.oracle.com under the Real Application Clusters section.

Peripherals

Considerations should be taken for preparation of the peripherals as well. For example, separate power sources should be used for separate nodes if possible to prevent a single power source failure crashing down the whole system.

We also strongly recommend using a highly available storage system with ORAC when high availability is the primary concerns for using ORAC. As pointed out earlier, a cluster system itself does not guard against storage system failure, it relies on the storage system to provide guards against disk failures.

In addition, you should use redundant networking links for the network interconnect when supported by the vendors. This practice can circumvent a hardware-based problem with the interconnect which could disrupt operations.

Client Connections

Connecting clients to ORAC is the same as connecting to a single Oracle database. No changes are needed to move clients from single database to ORAC based Oracle databases. This makes migrating to ORAC easier. In addition, there are a few connection options that can make the applications more failure resilient.

Connection Fail-over

Connection fail-over uses a low overhead connection option to fail over connections in the event of a failure. When the connection fails during the connect time, applications can fail over the connection to another live ORAC node transparently. This option prevents applications from trying to connect to

failed nodes repeatedly. It does not preserve states for sessions or queries. Applications need to handle the failure recovery during queries or transactions.

Transparent Application Fail-over (TAF)

Applications using the Transparent Application Fail-over option to ORAC can fail over transparently to other ORAC instance when the instance they connect to fails for most cases. Currently, session fail-over and SELECT operation fail-over are supported. In progress queries can be picked up where they left off prior to failure at the failed over instance. Transactions in progress during failure needs to be rolled back. A callback function is provided for applications so that they can continue on after fail-over without quitting the applications due to node failure.

ENHANCED FAIL-OVER WITH UNINTERRUPTED CAPACITY

For regular ORAC operations, when a node in an ORAC cluster fails, ORAC continues to operate at a reduced capacity, with the only decrease in performance due to the loss of the processing power of the failed node.

For customers who are primarily concerned about database availability, ORAC can also be configured to function as a much better fail-over cluster than those offered by OS vendors, e.g., FailSafe on Windows platform. ORAC enables Oracle database servers to provide guaranteed capacity even during node failure. Using this configuration option, one ORAC node provides reserved capacity as active standby for the other. When one node fails, the other node takes over immediately without lost of any database server capacity.

This enhanced fail-over capability provides much better availability than regular fail safe capability offered by OS vendors. Regular fail safe needs to restart all the database related services, including volume groups, network resources, application processes and the database instance from scratch after failure is detected, and database recovery must be completed before service is available. These operations can be time consuming when database grows larger. Since services are not available when these operations are ongoing, the benefits of such solutions are reduced. Also, they do not allow the standby node to be used for secondary tasks, such as administration and diagnostics.

With ORAC, the fail-over (standby) node is ready to take requests without restarting database all over again, since an active database instance is already running on the fail-over node. When the main node fails, the fail-over node can take over the operations also instantly. This configuration also has the added benefits of using the active standby node to do maintenance works on the database. This is only possible since the active standby shares the same database in the ORAC cluster. You can not do this with regular fail safe configuration since the database is not available in the standby node.

CASE STUDIES

Utilizing Oracle's own technologies, Oracle Corporation has been able to gain great cost saving benefits and enter into new markets that directly contributed to Oracle's bottom line. The following two cases demonstrate how ORAC is used to provide highly available and scalable database services in vastly different business applications.

Oracle Email Services

Oracle Email Services handle email communications for all Oracle employees over the world. There are 54,000 named users with a daily traffic load of many hundreds of gigabytes of data. Since email is a global service, it must be up 24 hours a day. The service requires a 3 second response time and 2 minute mean time to recovery.

Oracle recently consolidated its email servers. Before the consolidation, the email systems consisted of 97 servers that scattered around over 50 countries. There were 120 Oracle databases with 13 different character sets, different kinds of hardware and software and different email protocols. Email addresses had to be identified with different country codes. There were no backup services when some of them go down. To support the services with all the different configurations and operating environments required 70 system and database administrators, and an equal number of help desk and backup staff.

After the consolidation, Oracle email services use 11 servers, one common UTF8 character set, 5 databases and a set of common protocols such as LDAP, POP3, IMAP5. The service is hosted at one central location with disaster recovery standby at another location. For the first time, email addresses can all use one single domain oracle.com without the country code sub domains. A system diagram is shown at Figure 3.

A simple comparison can show the obvious great cost savings: 97 servers to 11, 120 databases to 5, standardized client software, all inclusive Oracle server software, data center space and staffing.

The server system is built completely using Oracle's own products. The 5 Oracle databases that are used as repositories for the email system are hosted by a two node ORAC cluster. The ORAC server is configured for one node to provide standby for the other to guarantee the capacity of the database in case of one node failure.

Aside from satisfying the performance requirements, the databases are required to recover within 60 seconds on average so that the email system can recover within 2 minutes. There is no way for regular fail safe to achieve the required recovery time, given that the sizes of database instances range from several GB to 18 GB, with database sizes up to 1.6 TB. The only option available is to use ORAC. The ORAC node acting as standby database has all the necessary resources running

and ready to take over while the main node is active. Since the database instance on the failed over node share the same database as the active one, it is ready to switch over as soon as the database recovery allows processing to proceed when the active node fails. Combining ORAC with Oracle Database's fast start recovery scheme, the ORAC database server deployed as Oracle Email Service's repository easily meets the recovery time requirements.

Oracle's email hosting business will be utilizing the same architecture to provide fast and reliable email hosting services.

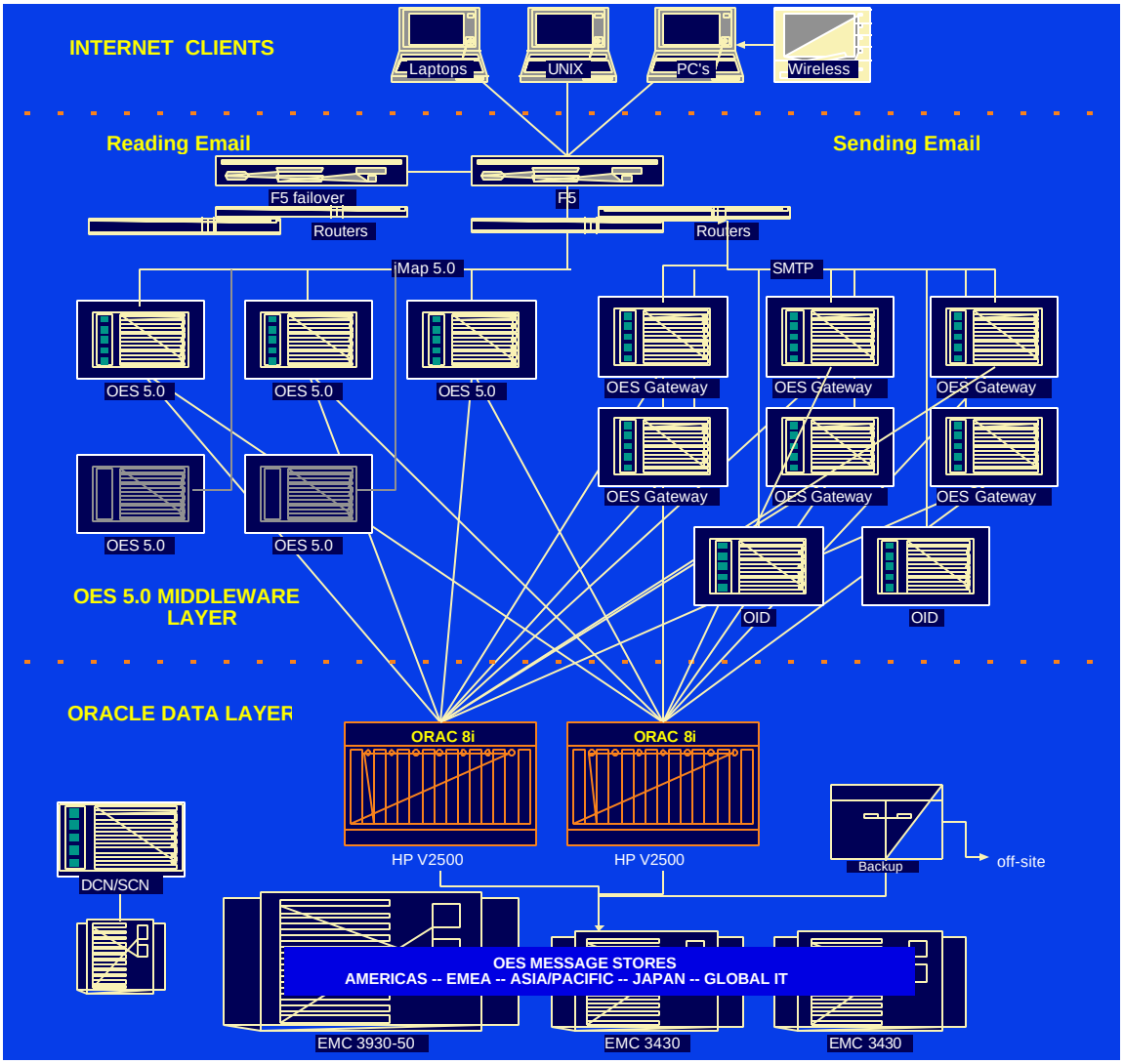

Figure 3. Oracle Email Service System Configuration

Oracle Exchange

Oracle Exchange acts as a single-source strategic partner to companies that want to transform to e-business by creating online business-to-business exchange and connecting their supply chains via the Internet. Oracle is the only platform provider that offers all the necessary products and services to create, run, support and maintain global B2B exchanges. Not only is it capable of ensuring the privacy and security of data for all exchange participants, Oracle Exchange is capable of managing the hundreds of millions of transactions annually that may be processed on an exchange.

The Oracle Exchange platform also offers all the necessary business transactions to support an entire industry's or a company's supply chain. It is based on Oracle's industry leading e-business suite, which supports a supply chain from the initial contact with the prospect, to manufacturing planning and execution, to post sales on-going service and support. This ensures that no matter how complex a company or an industry's business requirements, they can all be managed on the exchange. Figure 4 shows the architecture of Oracle Exchange.

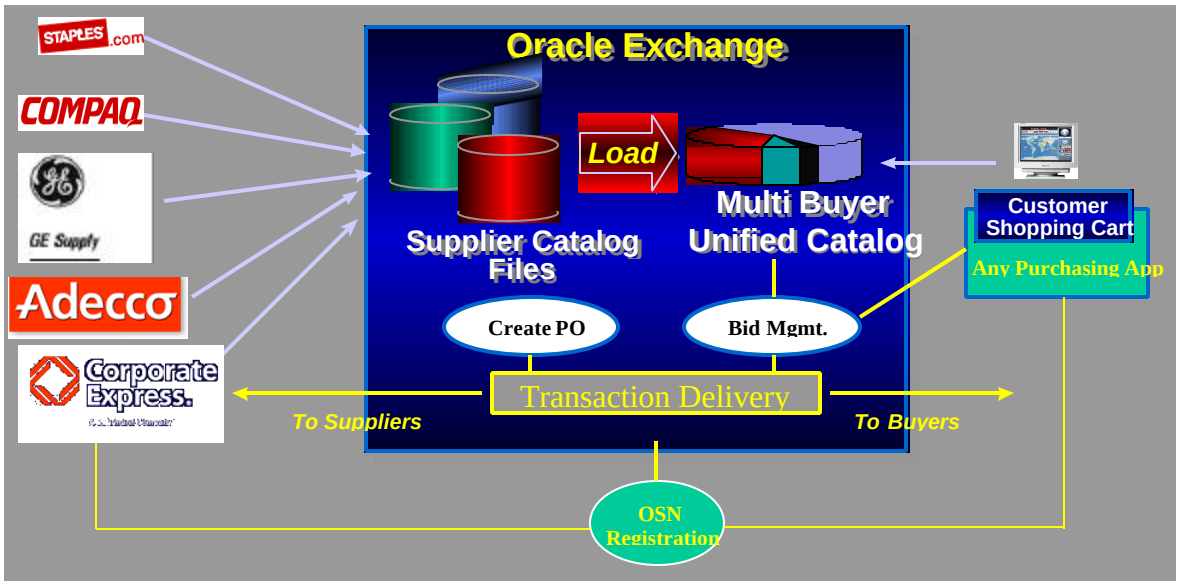

Figure 4. Oracle Exchange Architecture

Oracle not only provides the Exchange product suite, but also provides exchange hosting services for customers what want to focus on business aspects of using the exchanges.

Currently, Oracle hosts over 80 exchanges from all over the industry fields and all over the world from the US facilities. There are over 20 million catalog items in the exchanges. Database sizes are about 2.5 TB for all the exchanges. In addition

to delivering the required performance, the exchanges have a required recovery time of less than a minute.

ORAC plays a critical role for Oracle's exchange hosting services. For example, using one two-node ORAC cluster, Oracle exchange hosting service is able to host over 20 databases while meeting the performance and availability requirements. Similar to Oracle Email services, the ORAC cluster is configured to provide guaranteed capacity so that in case of failures, ORAC is still able to deliver the same performance. Again, this is not possible without ORAC as the deployment option for similar reasons cited in previous case. It is more so since the exchanges are providing services all over the world. They are active around the clock due to the time zones in which users reside.

CONCLUSION

Oracle Real Application Clusters are both highly available and highly scalable. ORAC achieves both goals without sacrificing one for the other.

This paper has attempted to give readers a clear idea on how ORAC addresses generic availability issues associated with cluster architecture and how ORAC is enhanced to provide better availability capabilities.

Real world examples in this paper demonstrated how ORAC is used in the business world to save costs and help business enter new markets as a highly available, highly scalable and cost effective database server platform.

Building Highly Available Database Servers Using Oracle Real Application Clusters
May 2001
Author: Jack Cai
Contributing Authors:

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.