NCOAUG
North Central Oracle Apps User Group

Required Technologies of Highly Available Database and Data Center Redundancy

Presented By Chad Dimatulac

Principal Database Architect

United Airlines

October 24, 2011

The Cost Of Downtime

How much are the losses of a potential business when a downtime occurs during a planned maintenance and unexpected site outage?

	YEARLY DOWNTIME DURATION		
UPTIME SERVICE LEVEL	DAYS	HOURS	MINUTES
95 %	18	6	0
99 %	3	15	36
99.9 %	0	8	46
99.99 %	0	0	53
99.999 %	0	0	5

Businesses lose an average of about $ 5,000 per minute in an outage. At that rate, $300,000 per hour is not something to dismiss lightly…

- Research from Emerson Network Power, eWeek Magazine May 13, 2011

Business Continuity

Business Continuity is the ability for a company to resume its e-commerce or operational systems in the event of disasters, system failures and planned maintenance.

- **Emergency Response Plan**

Data Center specific plan for responding to a system failure in the first hour of such event.

- **Business Resumption Plan**

An action plan for resuming critical business to an alternate site when disruptions occur at the primary site.

- **Business Disaster Plan**

A plan that deals with the recovery of the entire technology stack within an agreed timeframe during disastrous event that disables the entire operation of the company.

The Solution

ORACLE

CISCO

EMC2

Sun microsystems

CITRIX

HITACHI
Inspire the Next

Build a good infrastructure foundation that can scale to accommodate growing system throughput and with capabilities of providing business and operations continuity.

Symantec. Confidence in a connected world.

f5

FORTINET

QUEST SOFTWARE
Simplicity At Work™

QLOGIC

BARRACUDA NETWORKS

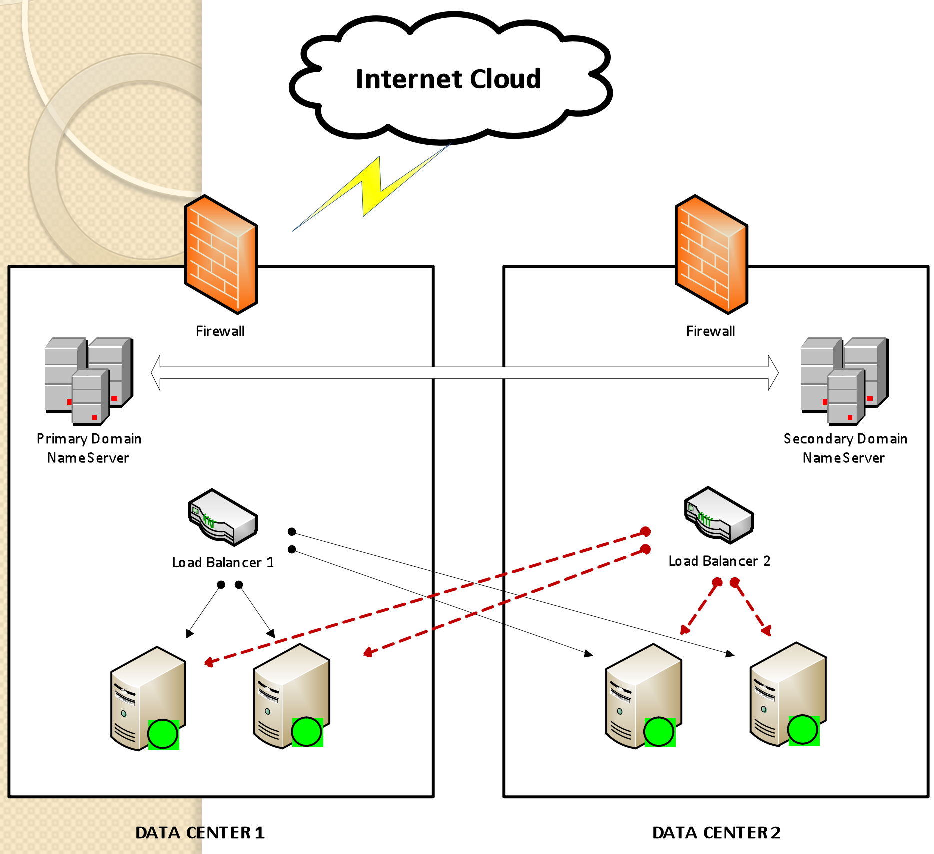
Topics

- High Availability Technology Overview
  - Oracle Data Guard
  - Symantec Cluster Server and Volume Replication
  - Oracle RAC
  - Logical Replication: Quest SharePlex
- Network Load Balancer
- Oracle Data Guard
- Symantec Global Cluster Server Failover System
- Oracle RAC Infrastructure
- Quest SharePlex Logical Replication – How It Works
- Quest SharePlex Components
- Quest SharePlex Implementation Types
- The High Availability Infrastructure
- Summary, Q & A

High Availability Technology Overview

- Symantec Cluster Failover and Disk Volume Replication (Active-Passive availability)
  - Monitors critical processes that if one fails, it starts the database instance on the next surviving node along with the VIP required for the database listener.
  - Disk Volume Replication is required for failover across data centers.
  - Short outage occurs during failover process. Application servers may be required to either manually reconnect (often by restart) or it has to validate existing connections to create new ones to replenish the stale connections.
- Oracle Data Guard (Active-Semi-passive System)
  - Provides standby database that can be configured for Non-Updatable Read-Only Database.
- Oracle RAC (Local Database availability)
  - Availability achieved by running multiple database instances across nodes on a shared disk. If a database instance is down, transactions can reconnect to the surviving nodes.
- SharePlex Data Replication (Global Active-Active multi-database availability)
  - Monitors the Oracle Database Redo Logs for DML/DDL applied to tables that were tagged for replication. Captures the SQL statements and transform it into compact messages and sends the messages to target servers for reassembling of SQL statements and apply them to the database.
  - Supports ETL type data propagation, horizontal and vertical data replication.
  - Supports replication of varying database versions and host platform between source and target.

Dual Data Center



The Need For A Redundant Data Center

To achieve an effective highly available application system, it is just right to host your failover site in a separate data center and if possible, across geographical distances. This not only provide better high availability but the failover site can also serve as your pre-production testing site such as load testing, functional testing of new code deployments, all these without affecting the main production system. The concept of a new environment - Failover and Pre-Production Site - addresses the short comings of a staging environment which are mostly scaled-down version of production site. With this, Load Tests are now much more realistic.

What is a Network Load Balancer?

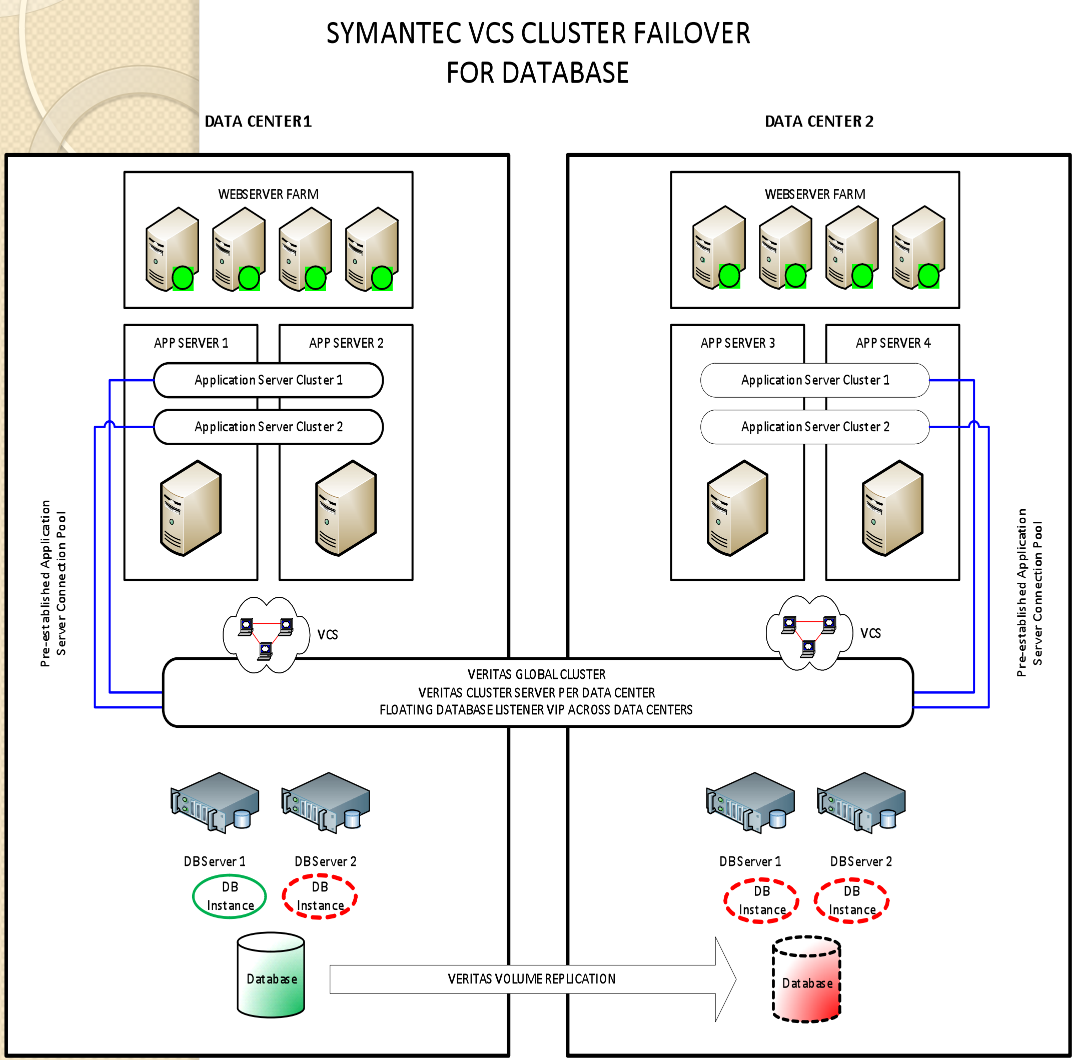
A software or a hardware device that distributes in-coming workload across multiple computers or a computer cluster. It is commonly implemented to provide a single internet service from multiple servers across geographic server farms. This includes large websites, Internet Relay Chat networks, FTP sites, DNS servers and database servers.

Common load balancing methods are weighted (distribution based on load) and round-robin (sequential distribution).

Popular devices:

- Citrix Netscaler
- F5 Big IP
- Barracuda Load Balancer
- Cisco Content Services Switch
- Fortinet FortiController

Symantec Global Cluster Server Failover System


SYMANTEC VCS CLUSTER FAILOVER
FOR DATABASE
DATA CENTER 1
DATA CENTER 2
WEBSERVER FARM
WEBSERVER FARM
APP SERVER 1
APP SERVER 2
APP SERVER 3
APP SERVER 4
Application Server Cluster 1
Application Server Cluster 2
Application Server Cluster 1
Application Server Cluster 2
Pre-established Application Server Connection Pool
Pre-established Application Server Connection Pool
VCS
VCS
VERITAS GLOBAL CLUSTER
VERITAS CLUSTER SERVER PER DATA CENTER
FLOATING DATABASE LISTENER VIP ACROSS DATA CENTERS
DBServer 1
DBServer 2
DBServer 1
DBServer 2
DB Instance
DB Instance
DB Instance
DB Instance
Database
VERITAS VOLUME REPLICATION
Database

What is Veritas Cluster Server?

A VCS cluster is a linking of multiple servers under a management system providing application failover and control of resources.

- Controls startup and shutdown of an application
- Provides application switching across nodes
- Application failover across nodes.
- Monitors health of critical application processes and underlying resources (file systems, network) as basis for a failover.

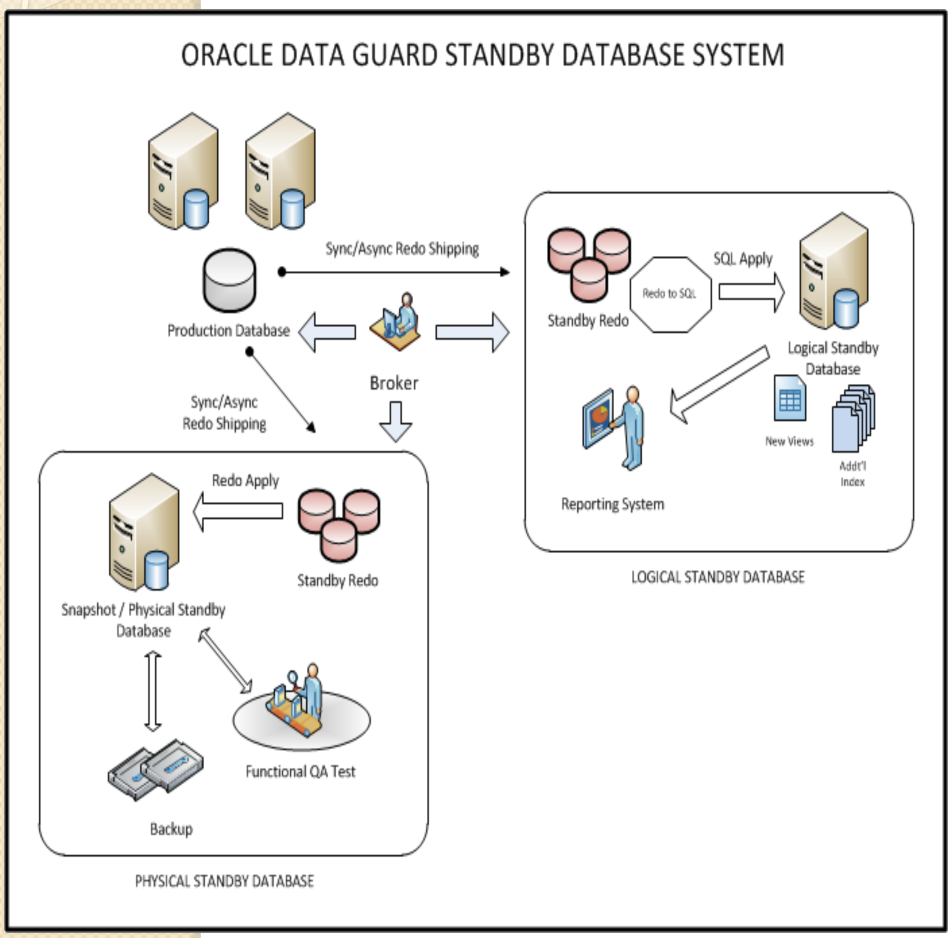
In VCS, an application runs on one node at a time and uses a virtual IP associated to the application.

What is Global Cluster Server?

A Global Cluster Server is a cluster of multiple local VCS providing service control failover across geographical locations.

- Requires a network that supports long distance communication.
- Requires Veritas Volume Replication to mirror the disk volumes from one SAN system to another.

Oracle Data Guard



What Is Data Guard?

A feature in Oracle database providing services to create, maintain and monitor standby databases as an updated copy of production database. It's primary purpose is to protect the primary database from total loss during disaster situations by having an identical standby site.

Physical Standby Database

An identical copy of the primary database on a block for block level. Synchronization is done through Redo Apply process by applying the redo data received from the primary database. It requires both the primary and standby database to be of the same version and release level. As of 11g R1, it can be opened as read-only while receiving and applying redo data.

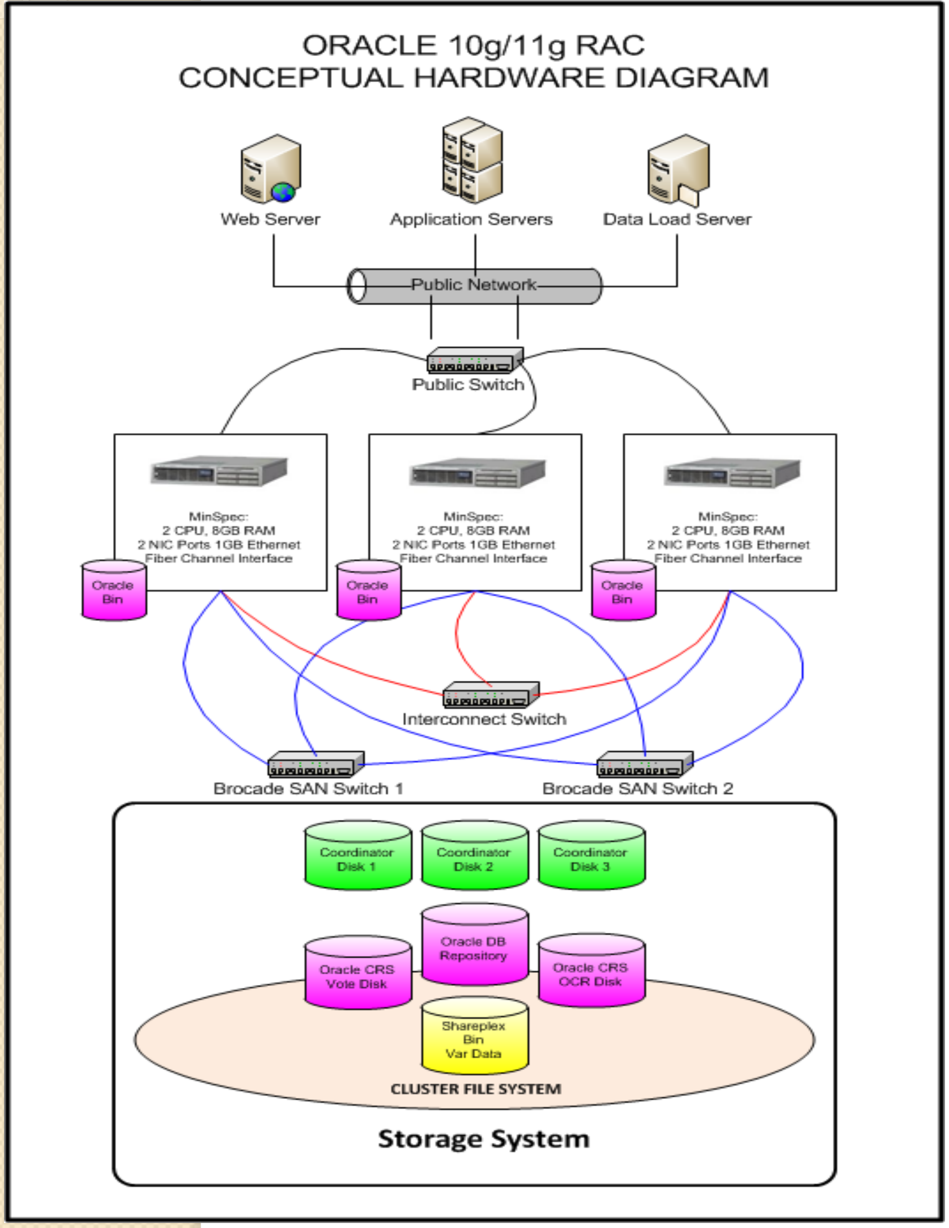
Logical standby database

Contains the same data found in the primary database but has the flexibility of having a different underlying physical structures to support a specific querying and business reporting application. Synchronization is done through SQL Apply process that transforms the redo data of the primary database the into SQL statements (DML and supported DDL only) and applies them at the standby database. In this configuration, the standby database can be used not only for reporting and data protection but also for database upgrades.

Snapshot Standby Database

Similar to physical standby, it receives and archives the redo data from a primary database but, it holds-off the redo apply until it is converted from Snapshot mode into Physical Standby mode. Snapshot standby provides an updatable physical standby database. It requires Flashback Database to be configured in order to revert back to the original physical standby and discarding any changes made.

Oracle RAC Infrastructure Setup



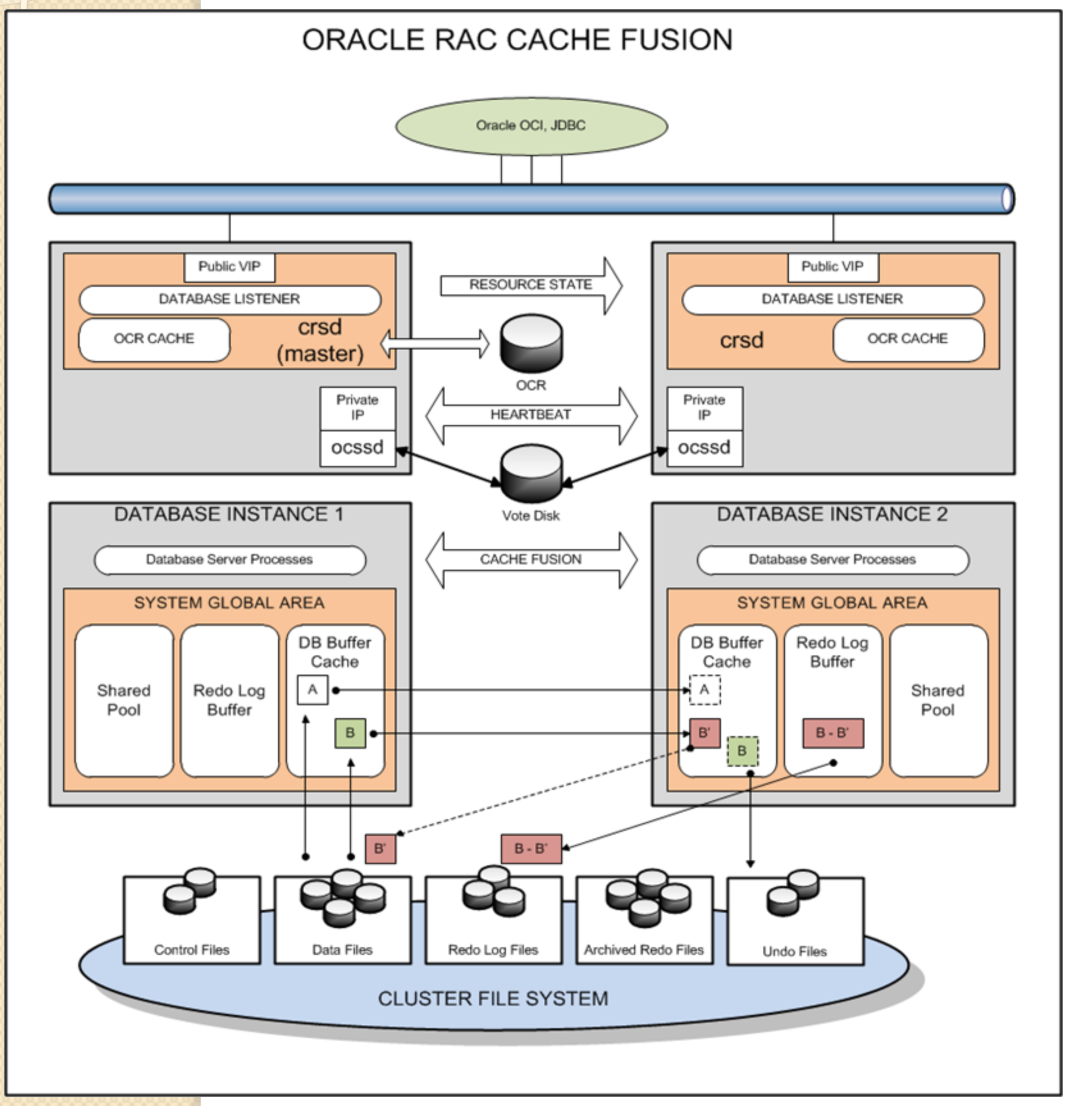
What is Oracle RAC?

A database system having multiple server instances accessing a shared database repository.

Required System Components

- At least 2 network interface per RAC server. One for public IP and another for private IP.
- Network switch dedicated for interconnect.
- Cluster Server
  - Oracle CRS (now called Oracle Clusterware)
  - 3rd Party Vendor Cluster Server (i.e. Symantec SFRAC)
- Network interface redundancy and load balancing for interconnect. (NIC bonding)
- Storage system supporting SCSI-3 Persistent Reservation
- Shared disk system (Oracle ASM or 3rd Party Vendor Cluster File System, or NFS)

Oracle Cache Fusion



What is Cache Fusion?

A technology for transferring data blocks between database instances in a cluster through the interconnect.

Block Sharing Scenarios

- ❖ *Concurrent Reads On Multiple Nodes*. Occurs when two or more clustered database instances are required to read the same data block. The first instance that fetches the data block from the data files becomes the owner (master) of the block.
- ❖ *Concurrent Read and Writes On Multiple Nodes*. A mixture of read and write operations against a single block.
- ❖ *Concurrent Writes On Multiple Nodes*. This is a situation where multiple database instances request modification of the same data block.

NOTE: Your interconnect need to be sized to have a larger bandwidth to accommodate the estimated traffic of Cache Fusion.

Quest SharePlex Logical Replication

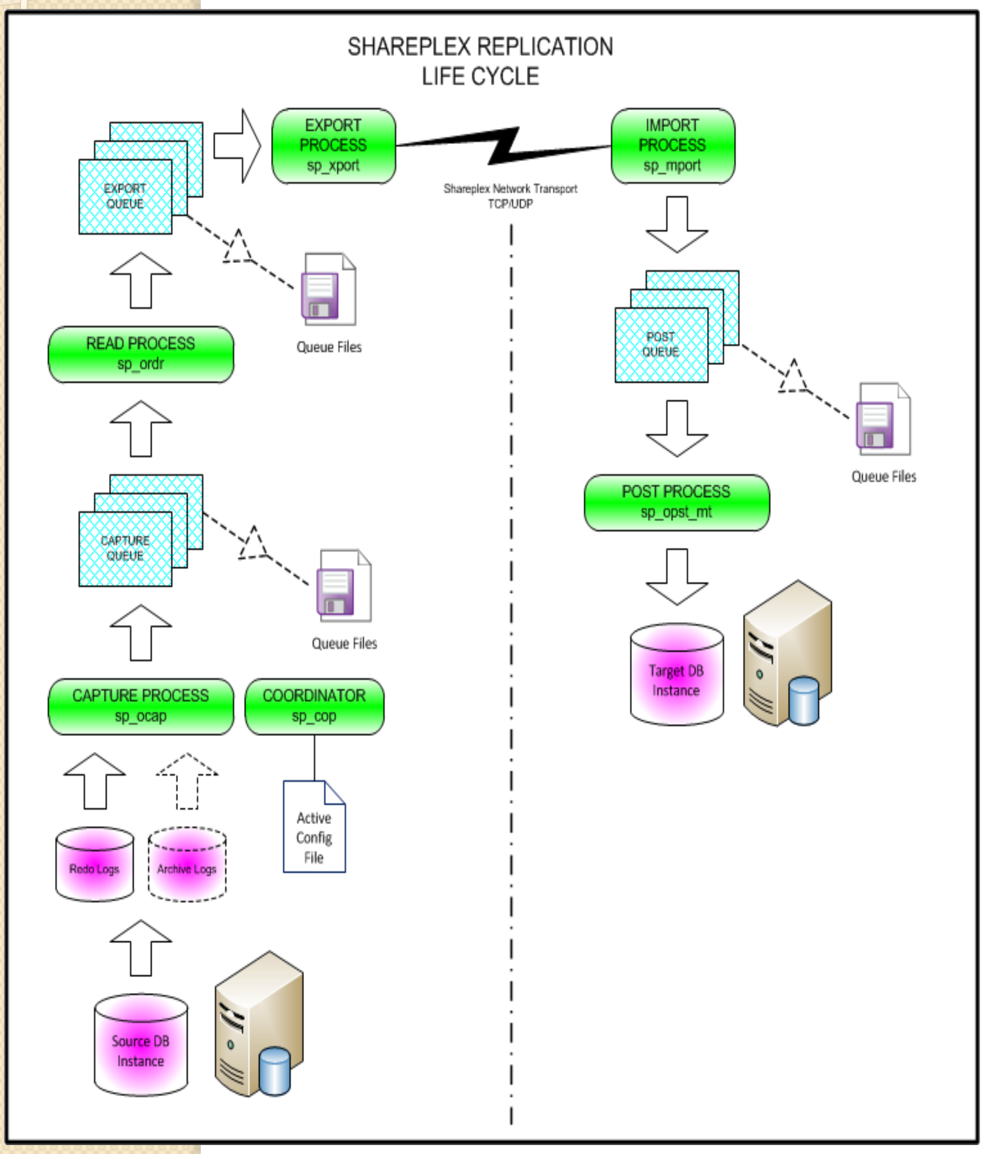
Setup

- For the source system, SharePlex instance need to be running on the host where it can read the redo logs and archive logs of the source database.
- For the target system, SharePlex instance can run remotely from the target database but most implementation has the SharePlex running on the same host where the target database instance is running.
- For both source and target system, an Oracle Client is required to be installed.

Process Flow

- Reads the blocks of online redo logs for operations that need to be replicated based on the definitions in the configuration file.
- SharePlex replicates only the changed data which makes the replication messages more compact and light.
- Large operations involving LONG and LOB data type may require several messages because of the message size limitation.
- On the target system, a Post process receives the messages and assembles the SQL statements and applies it to the database.

Quest SharePlex Components



CAPTURE PROCESS

Reads the log entries from the redo log or archive logs that are relevant to the objects bound for replication as defined in the config file. The entries are stored in the Capture Queue.

READ PROCESS

Reads the entries in the Capture Queue and strips the entries into light weight Shareplex message data and puts them into the Export Queue.

EXPORT PROCESS

Sends the Shareplex message data across the network to the target system.

IMPORT PROCESS

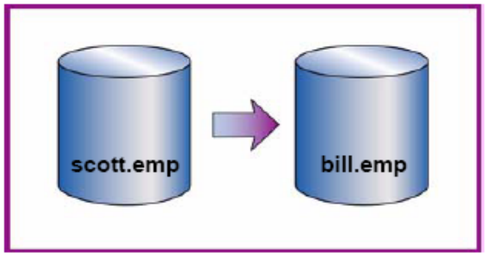
Receives all exported messages and puts them into the Post Queue.

POST PROCESS

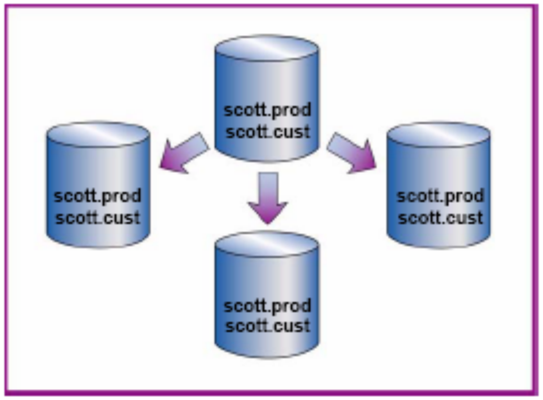
Reads the Post Queue and assembles the SQL statement and applies it to the designated Oracle database.

Quest SharePlex Implementation Types

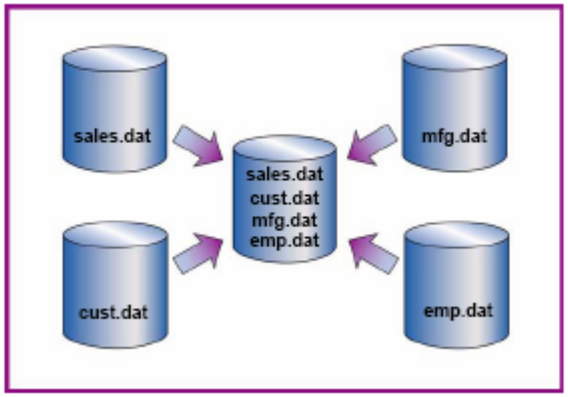
Replication to Reporting Instance



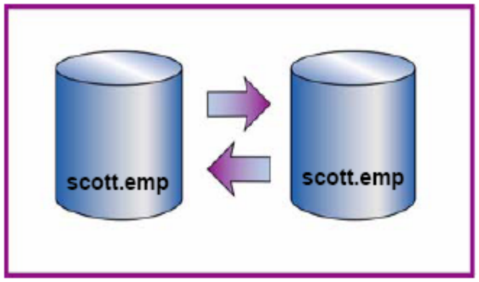
Replication to Distribute Data



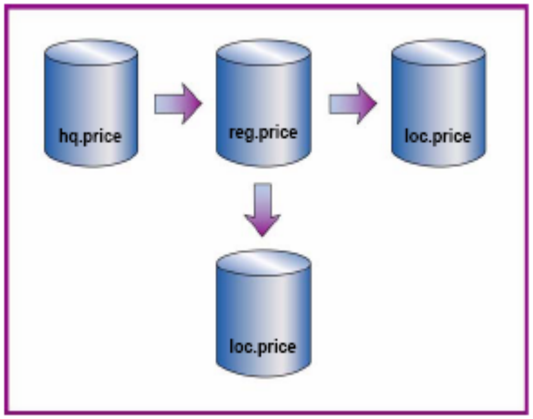
Replication For Centralized Reporting



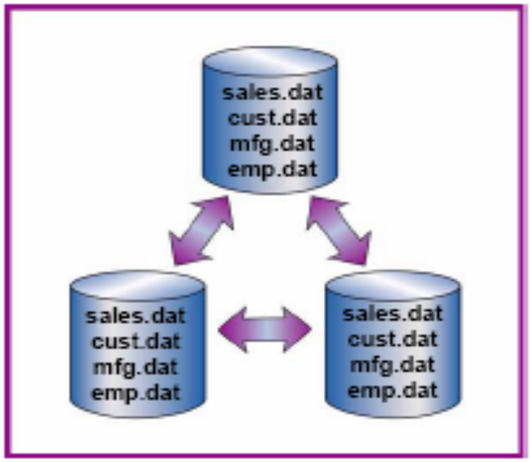
Replication For High Availability



Replication Using Intermediary System



Peer-to-peer Replication



Peer-To-Peer Replication Guidelines

- SharePlex replicates both DDL and DML changes written on redo logs. Any object bound for replication need to be in logging mode. Optionally, DDL replication can be turned off.
- Use supported data types and operations. See SharePlex Release Note for details.
  - Some unsupported operations: materialized views to materialized views, nested tables, clustered tables, tables with compress option, LONG and LONG RAW data type in transformation and conflict resolution, index organized table (IOT) with LONG and VARRAY data type, replication between non-IOT tables and IOT tables, XML type tables, data in transparent data encryption.
- Table structure requirements:
  - Primary key or unique key (replc_key). This also applies to child tables.
  - Transaction timestamp column (replc_timestamp)
  - Site priority column (replc_db_priority)
  - Triggers to populate the timestamp and site priority column
- Unique number sequencing between source and target database for conflict avoidance. (Ex. odd number on primary database then even number in secondary database)
- Identify running balance columns for conflict resolution.

Required Technologies For A Successful Database HA

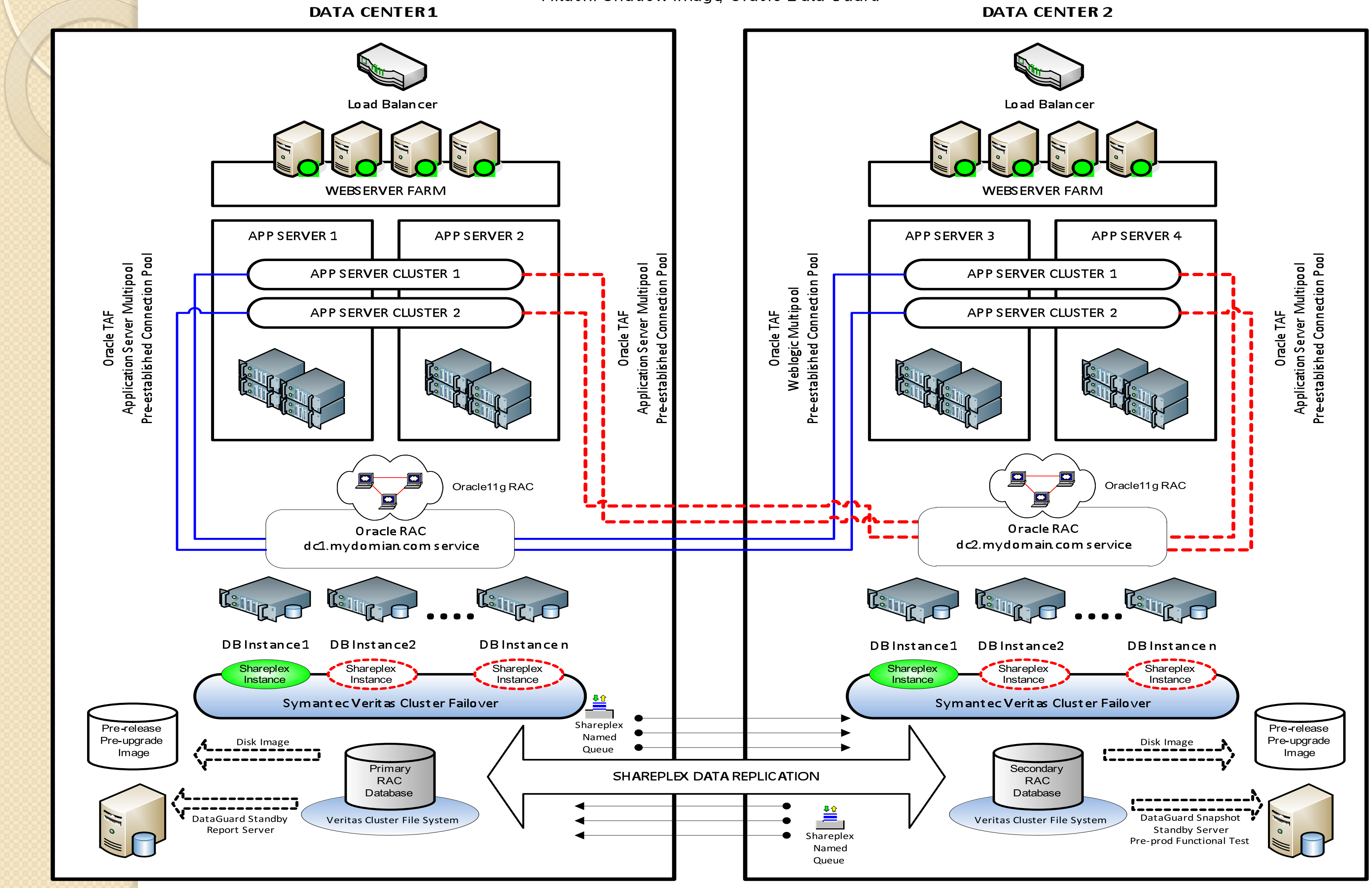
```
MULTIPOOL
     CONNECTION POOL 1
jdbc:oracle:thin:@(DESCRIPTION =(LOAD_BALANCE=ON)(FAILOVER=ON)
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc1rac1-vip.mydomain.com)(PORT = 1521))
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc1rac2-vip.mydomain.com)(PORT = 1521))
                              … more entries …
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc1rac#-vip.mydomain.com)(PORT = 1521))
     )
    (CONNECT_DATA =
      (SERVICE_NAME = PRODDB.mydomain.com)
     ))

    CONNECTION POOL 2
jdbc:oracle:thin:@(DESCRIPTION =(LOAD_BALANCE=ON)(FAILOVER=ON)
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc2rac1-vip.mydomain.com)(PORT = 1521))
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc2rac2-vip.mydomain.com)(PORT = 1521))
                              … more entries …
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc2rac#-vip.mydomain.com)(PORT = 1521))
     )
    (CONNECT_DATA =
      (SERVICE_NAME = PRODDB.mydomain.com)
     ))
```

```
# File: tnsnames.ora

DC1PRODDB =
  (DESCRIPTION = (LOAD_BALANCE=ON) (FAILOVER=ON)
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc1rac1-vip.mydomain.com)(PORT = 1521))
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc1rac2-vip.mydomain.com)(PORT = 1521))
                               … more entries …
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc1rac#.vip.mydomain.com)(PORT = 1521))
    )
    (CONNECT_DATA =
      (SERVICE_NAME = PRODDB.mydomain.com)
      (FAILOVER_MODE = (TYPE=select)(METHOD=preconnect)(RETRIES=3)(DELAY=5)(BACKUP=DC2PRODDB))
    )
  )

DC2PRODDB =
  (DESCRIPTION = (LOAD_BALANCE=ON) (FAILOVER=ON)
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc2rac1-vip.mydomain.com)(PORT = 1521))
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc2rac2-vip.mydomain.com)(PORT = 1521))
                               … more entries …
      (ADDRESS = (PROTOCOL = TCP)(HOST = dc2rac#.vip.mydomain.com)(PORT = 1521))
    )
    (CONNECT_DATA =
      (SERVICE_NAME = PRODDB.mydomain.com)
      (FAILOVER_MODE = (TYPE=select)(METHOD=preconnect)(RETRIES=3)(DELAY=5)(BACKUP=DC1PRODDB))
    )
  )
```

- **Network Load Balancer**
  - Placed at the upper stack of your application servers or web servers.
  - Its role is to provide control of network traffic coming from the users for distribution across data centers.
- **Oracle RAC**
  - Provides high availability within a data center.
- **Quest SharePlex**
  - Provides database replication across data centers
- **Oracle TAF and Connection Load Balancing**
  - Transparent Application Failover for SQL queries across RAC nodes.
  - Connection Load Balance based on node load and number of instance connections.
- **Application Server Transaction Failover**
  - Data Source Connection Multipool in failover mode. This provides connection pool failover to a replicated database.
- **Symantec Storage Foundation 5 (Veritas SFRAC 5)**
  - Provides server clustering, cluster failover, cluster file system and ODM API.

The High Availability Infrastructure Architecture

HIGH AVAILABILITY ARCHITECTURE
SHAREPLEX DATABASE REPLICATION
with Oracle TAF, Weblogic Multipool, Oracle RAC, Veritas SFRAC,
Hitachi Shadow Image, Oracle Data Guard



The High Availability Infrastructure Summary

NOTE: Please see High Availability Infrastructure Architecture diagram in previous slide as reference

- ❖ Each data center has an Oracle RAC that provides high-availability and scalability at the database instance level.
- ❖ Server-side connection load balancing is configured in RAC, thus, transaction connections are load balanced across the RAC nodes.
- ❖ SharePlex is configured to replicate in peer-to-peer between RAC databases on each data center. Named Queue replication stream is configured for each schema providing applications to have their own replication flow and also help improve replication speed overall.
- ❖ Cluster Server Failover is configured for Shareplex providing high availability of the replication instance during a server failure or server maintenance.
- ❖ Disk Image Snapshot is configured for the RAC database repository to capture a disk image of a Pre-upgrade or Pre-release database. The disk image is useful during fallback situations of a planned maintenance.
- ❖ The application server has multiple connection pools each pointing to a RAC database of a data center. Multipool is configured in failover mode for the automatic data center failover and failback of application transaction connections. In this mode, all transactions goes to the primary database in data center 1 and only when there are connection issues that a failover to the next connection pool (data center 2) occurs. Failback is automatic the moment the primary database is available.

Q & A

Thank You.

Chad Dimatulac
Principal Database Architect
United Airlines