Oracle Real Application Clusters on VMware vSAN

October 13, 2017
Table of Contents

1. Executive Summary
   1.1. Business Case
   1.2. Solution Overview
   1.3. Key Results

2. vSAN Oracle RAC Reference Architecture
   2.1. Purpose
   2.2. Scope
   2.3. Audience
   2.4. Terminology

3. Technology Overview
   3.1. Technology Overview
   3.2. VMware vSphere
   3.3. VMware vSAN
   3.4. VMware vSAN Stretched Cluster
   3.5. Oracle 11gR2 RAC
   3.6. Oracle Extended RAC
   3.7. Oracle Data Guard
   3.8. Oracle RMAN

4. Deploying Oracle Extended RAC on vSAN
   4.1. Overview
   4.2. Deployment Considerations on vSAN Stretched
   4.3. Oracle Cluster ware and vSAN Stretched
   4.4. Behavior during Network Partition
   4.5. Deployment Benefits

5. Solution Configuration
   5.1. Solution Configuration
   5.2. Architecture Diagram
   5.3. Hardware Resources
   5.4. Software Resources
   5.5. Network Configuration
   5.6. VMware ESXi Servers
   5.7. vSAN Configuration
   5.8. Oracle RAC VM and Database Storage Configuration

6. Solution Validation
   6.1. Solution Validation
   6.2. Test Overview
   6.3. Test and Performance Data Collection Tools
   6.4. Oracle RAC Performance on vSAN
   6.5. Oracle RAC Scalability on vSAN
   6.6. vSAN Resiliency
   6.7. vSphere vMotion on vSAN
   6.8. Oracle Extended RAC Performance on vSAN
   6.9. vSAN Stretched Cluster Site Failure Tolerance
   6.10. Metro and Global DR with vSAN Stretched
   6.11. Oracle RAC Database Backup and Recovery on vSAN

7. Best Practices of Virtualized Oracle RAC on vSAN
   7.1. Best Practices of Virtualized Oracle RAC on vSAN
   7.2. vSAN Storage Configuration Guidelines
   7.3. CPU Configuration Guidelines
   7.4. Memory Configuration Guidelines
   7.5. Network Configuration Guidelines

8. Conclusion
   8.1. Conclusion

9. Reference
   9.1. White Paper
   9.2. Product Documentation
   9.3. Other Documentation

10. About the Author and Contributors
    10.1. About the Author and Contributors
1. Executive Summary

This section covers the Business Case, Solution Overview and Key Results of the Oracle Real Application Clusters on vSAN document.
1.1 Business Case

Customers deploying Oracle Real Application Clusters (RAC) have requirements such as stringent SLA’s, continued high performance, and application availability. It is a major challenge for business organizations to manage data storage in these environments due to the stringent business requirement. Common issues in using traditional storage solutions for business critical application (BCA) include inadequate performance, scale-in/scale-out, storage inefficiency, complex management, and high deployment and operating costs.

With more and more production servers being virtualized, the demand for highly converged server-based storage is surging. VMware vSAN™ aims at providing a highly scalable, available, reliable, and high performance storage using cost-effective hardware, specifically direct-attached disks in VMware ESXi™ hosts. vSAN adheres to a new policy-based storage management paradigm, which simplifies and automates complex management workflows that exist in traditional enterprise storage systems with respect to configuration and clustering.

1.2 Solution Overview

This solution addresses the common business challenges discussed in the previous section that CIOs face today in an online transaction processing (OLTP) environment that requires availability, reliability, scalability, predictability and cost-effective storage, which helps customers design and implement optimal configurations specifically for Oracle RAC Database on vSAN.

1.3 Key Results

The following highlights validate that vSAN is an enterprise-class storage solution suitable for Oracle RAC Database:

- Predictable and highly available Oracle RAC OLTP performance on vSAN.
- Simple design methodology that eliminates operational and maintenance complexity of traditional SAN.
- Sustainable solution for enterprise Tier-1 Database Management System (DBMS) application platform.
- Validated architecture that reduces implementation and operational risks.
- Integrated technologies to provide unparalleled availability, business continuity (BC), and disaster recovery (DR).
- Efficient backup and recovery solution using Oracle RMAN.
2. vSAN Oracle RAC Reference Architecture

This section provides an overview of the architecture validating vSAN's ability to support industry-standard TPC-C like workloads in an Oracle RAC environment.
2.1 Purpose

This reference architecture validates vSAN’s ability to support industry-standard TPC-C like workloads in an Oracle RAC environment. Oracle RAC on vSAN ensures a desired level of storage performance for mission-critical OLTP workload while providing high availability (HA) and DR solution.

2.2 Scope

This reference architecture:

- Demonstrates storage performance scalability and resiliency of enterprise-class 11gR2 Oracle RAC database in a vSAN environment.
- Shows vSAN Stretched Cluster enabling Oracle Extended RAC environment. It also shows the resiliency and ease of deployment offered by vSAN Stretched Cluster.
- Provides an availability solution including a three-site DR deployment leveraging vSAN Stretched Cluster and Oracle Data Guard.
- Provides a business continuity solution with minimal impact to the production environment for database backup and recovery using Oracle RMAN in a vSAN environment.

2.3 Audience

This reference architecture is intended for Oracle RAC Database administrators and storage architects involved in planning, architecting, and administering an environment with vSAN.

2.4 Terminology

This paper includes the following terminologies.

<table>
<thead>
<tr>
<th>TERM</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle Automatic Storage Management (Oracle ASM)</td>
<td>Oracle ASM is a volume manager and a file system for Oracle database files that support single-instance Oracle Database and Oracle RAC configurations.</td>
</tr>
<tr>
<td>Oracle Clusterware</td>
<td>Oracle Clusterware is a portable cluster software that allows clustering of independent servers so that they cooperate as a single system.</td>
</tr>
<tr>
<td>Oracle Data Guard</td>
<td>Oracle Data Guard ensures high availability, data protection, and disaster recovery for enterprise data. Data Guard provides a comprehensive set of services that create, maintain, manage, and monitor one or more standby databases to enable production Oracle databases to survive disasters and data corruptions.</td>
</tr>
<tr>
<td>Oracle RAC</td>
<td>Oracle Real Application Clusters is a clustered version of Oracle database providing a deployment of a single database across a cluster of servers.</td>
</tr>
<tr>
<td>TERM</td>
<td>DEFINITION</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Oracle Extended RAC</strong> (Oracle RAC on Extended Distance Cluster)</td>
<td>Oracle Extended RAC is a deployment model in which servers in the cluster reside in locations that are physically separated.</td>
</tr>
<tr>
<td><strong>Primary database</strong></td>
<td>Also known as the production database that functions in the primary role. This is the database that is accessed by most of your applications.</td>
</tr>
<tr>
<td><strong>Physical standby database</strong></td>
<td>A physical standby uses Redo Apply to maintain a block for block, replica of the primary database. Physical standby databases provide the best DR protection for Oracle Database</td>
</tr>
<tr>
<td><strong>RMAN</strong></td>
<td>RMAN (Recovery Manager) is a backup and recovery manager for Oracle Database.</td>
</tr>
</tbody>
</table>
3. Technology Overview

This section provides an overview of the technologies used in this solution.
3.1 Technology Overview

This section provides an overview of the technologies used in this solution:

- VMware vSphere®
- VMware vSAN
- VMware vSAN Stretched Cluster
- Oracle 11gR2 RAC
- Oracle Extended RAC
- Oracle Data Guard
- Oracle RMAN

3.2 VMware vSphere

VMware vSphere is the industry-leading virtualization platform for building cloud infrastructures. It enables users to run business-critical applications with confidence and respond quickly to business needs. vSphere accelerates the shift to cloud computing for existing data centers and underpins compatible public cloud offerings, forming the foundation for the industry's best hybrid-cloud model.

3.3 VMware vSAN

VMware vSAN is VMware's software-defined storage solution for hyper-converged infrastructure, a software-driven architecture that delivers tightly integrated computing, networking, and shared storage from a single virtualized x86 server. vSAN delivers high performance, highly resilient shared storage by clustering server-attached flash devices and hard disks (HDDs).

vSAN delivers enterprise-class storage services for virtualized production environments along with predictable scalability and all-flash performance—all at a fraction of the price of traditional, purpose-built storage arrays. Just like vSphere, vSAN provides users the flexibility and control to choose from a wide range of hardware options and easily deploy and manage it for a variety of IT workloads and use cases. vSAN can be configured as all-flash or hybrid storage.
vSAN supports a hybrid disk architecture that leverages flash-based devices for performance and magnetic disks for capacity and persistent data storage. In addition, vSAN can use flash-based devices for both caching and persistent storage. It is a distributed object storage system that leverages the vSphere Storage Policy-Based Management (SPBM) feature to deliver centrally managed, application-centric storage services and capabilities. Administrators can specify storage attributes, such as capacity, performance, and availability, as a policy on a per VMDK level. The policies dynamically self-tune and load balance the system so that each virtual machine has the right level of resources.

3.4 VMware vSAN Stretched Cluster

vSAN 6.1 introduces the new feature Stretched Cluster. vSAN Stretched Clusters provides customers with the ability to deploy a single vSAN Cluster across multiple data center. vSAN Stretched Cluster is a specific configuration implemented in environments where disaster or downtime avoidance is a key requirement.

vSAN Stretched Cluster builds on the foundation of Fault Domains. The Fault Domain feature introduced rack awareness in vSAN 6.0. The feature allows customers to group multiple hosts into failure zones across multiple server racks in order to ensure that replicas of virtual machine objects are not provisioned onto the same logical failure zones or server racks. Similarly, vSAN Stretched Cluster requires three failure domains and is based on three sites (two active—active sites and witness site). The witness site is only utilized to host witness virtual appliance that stores witness objects and cluster metadata information and also provide cluster quorum services during failure events.
The nomenclature used to describe a vSAN Stretched Cluster configuration is $X+Y+Z$, where $X$ is the number of ESXi hosts at data site A, $Y$ is the number of ESXi hosts at data site B, and $Z$ is the number of witness host at site C. Data sites are where virtual machines are deployed. The minimum supported configuration is three nodes (1+1+1). The maximum configuration is 31 nodes (15+15+1).

vSAN Stretched Cluster differs from regular vSAN Cluster in the following perspectives:

- **Write latency**: In a regular vSAN Cluster, mirrored writes incur the same latency. In a vSAN Stretched Cluster, you need to prepare the write operations at two sites. Therefore, write operation needs to traverse the inter-site link, and thereby incur the inter-site latency. The higher the latency, the longer it takes for the write operations to complete.
- **Read locality**: The regular cluster does read operations in a round robin pattern across the mirrored copies of an object. The stretched cluster does all reads from the single-object copy available at the local site.
- **Failure**: In the event of any failure, recovery traffic needs to originate from the remote site, which has the only mirrored copy of the object. Thus, all recovery traffic traverses the inter-site link. In addition, since the local copy of the object on a failed node is degraded, all reads to that object are redirected to the remote copy across the inter-site link.

See more information in the [vSAN 6.1 Stretched Cluster Guide](#).
Oracle RAC is a clustered version of Oracle database based on a comprehensive high-availability stack that can be used as the foundation of a database cloud system as well as a shared infrastructure, ensuring high availability, scalability, and agility for any application.

In an Oracle RAC environment, Oracle databases runs on two or more systems in a cluster while concurrently accessing a single shared database. The result is a single database system that spans multiple hardware systems, enabling Oracle RAC to provide high availability and redundancy during failures in the cluster. It enables multiple database instances that are linked by a network interconnect to share access to an Oracle database. Oracle RAC accommodates all system types, from read-only data warehouse systems to update-intensive OLTP systems.
3.6 Oracle Extended RAC

Oracle Extended RAC provides greater availability than local Oracle RAC. It provides extremely fast recovery from a site failure and allows for all servers, in all sites, to actively process transactions as part of a single database cluster. The high impact of latency, and therefore distance, creates some practical limitations as to where this architecture can be deployed. This architecture fits best where the two datacenters are located relatively close and where the costs of setting up direct dedicated channels between the sites have already been taken.

3.7 Oracle Data Guard

Oracle Data Guard provides the management, monitoring, and automation software infrastructure to create and maintain one or more standby databases to protect Oracle data from failures, disasters, errors, and data corruptions. Data Guard is unique among Oracle replication solutions in supporting both synchronous (zero data loss) and asynchronous (near-zero data loss) configurations. Administrators can choose either manual or automatic failover of production to a standby system if the primary system fails to maintain high availability for mission-critical applications.

There are two types of standby databases:

- A physical standby uses Redo Apply to maintain a block for block, exact replica of the primary database. Physical standby databases provide the best DR protection for the Oracle Database. We use this standby database type in this reference architecture.
- The second type of the standby database uses SQL Apply to maintain a logical replica of the primary database. While a logical standby database contains the same data as the primary database, the physical organization and structure of the data can be different.

Oracle Active Data Guard: Oracle Active Data Guard enhances the Quality of Service (QoS) for production databases by off-loading resource-intensive operations to one or more standby databases, which are synchronized copies of the production database. With Oracle Active Data Guard, a physical standby database can be used for real-time reporting, with minimal latency between reporting and production data. Additionally, Oracle Active Data Guard continues to provide the benefit of high availability and disaster protection by failing over to the standby database in a planned or an unplanned outage at the production site.

3.8 Oracle RMAN

Oracle Recovery Manager (RMAN) provides a comprehensive foundation for efficiently backing up and recovering Oracle databases. A complete HA and DR strategy requires dependable data backup, restore, and recovery procedures. RMAN is designed to work intimately with the server, providing block-level corruption detection during database backup and recovery. RMAN optimizes performance and space consumption during backup with file multiplexing and backup set compression, and integrates with third-party media management products for tape backup.
4. Deploying Oracle Extended RAC on vSAN

This section details on deploying Oracle Extended RAC on vSAN.
4.1 Overview

Oracle RAC implementations are primarily designed as scalability and high availability solution that resides in a single data center. By contrast, in an Oracle Extended RAC architecture, the nodes are separated by geographic distance. For example, if a customer has a corporate campus, they might want to place individual Oracle RAC nodes in separate buildings. This configuration provides a higher degree of disaster tolerance in addition to the normal Oracle RAC high availability because a power shutdown or fire in one building would not stop the database from processing if properly setup. Similarly, many customers who have two data centers in reasonable proximity, which are already connected by a high speed link and are often on different power grids and flood plains, can take advantage of this solution.

To deploy this type of architecture, the RAC nodes are physical dispersed across both sites to protect them from local server failures. Similar consideration is required for storage as well.

vSAN Stretched Cluster inherently provides this storage solution suitable for Oracle Extended RAC using the vSAN fault domain concept. This ensures that one copy of the data is placed on one site, a second copy of the data on another site, and the witness components are always placed on the third site (witness site).

A virtual machine deployed on a vSAN Stretched Cluster has one copy of its data in site A, a second copy of its data in site B, and any witness components placed on the witness host in site C. This configuration is achieved through fault domains. In the event of a complete site failure, there is a full copy of the virtual machine data as well as greater than 50 percent of the components available. This enables the virtual machine to remain available on the vSAN datastore. If the virtual machine needs to be restarted on the other data site, VMware vSphere High Availability handles this task.

Oracle Extended RAC is not a different installation scenario comparing to Oracle RAC. Make sure to prepare the infrastructure so the distance between the nodes is transparent to the Oracle RAC Database. Similarly, for underlying storage, the transparency is maintained by data mirroring across distances provided by vSAN Stretched Cluster.

4.2 Deployment Considerations on vSAN Stretched

vSAN Stretched Cluster enables active/active data centers that are separated by metro distance. Oracle Extended RAC with vSAN enables transparent workload sharing between two sites accessing a single database while providing the flexibility of migrating or balancing workloads between sites in anticipation of planned events such as hardware maintenance. Additionally, in case of unplanned event that causes disruption of services in one of the sites, the failed client connections can be automatically redirected using Oracle transparent application failover (TAF) to the oracle nodes running at the surviving site. Furthermore, if configured correctly during a single site failure, VMware vSphere HA can restart the failed Oracle RAC VM on the surviving site providing resiliency.

4.3 Oracle Clusterware and vSAN Stretched

Oracle Clusterware is a software that enables servers to operate together as if they are one server. It is a prerequisite for using Oracle RAC. Oracle Clusterware requires two components: a voting disk to record node membership information and the Oracle Cluster Registry (OCR) to record cluster configuration information. The voting disk and the OCR must reside on a shared storage. Deployments of Oracle Extended RAC emphasize deploying a third site to host the third Oracle RAC cluster voting disk (based on NFS) that acts as the arbitrator in case either site fails or a communication failure occurs between the sites. See Oracle RAC and Oracle RAC One Node on Extended Distance (Stretched) Clusters for more information. However, in Oracle Extended RAC using vSAN Stretched Cluster, the cluster voting disks completely reside on vSAN Stretched Cluster datastore and the NFS-based one is not required. vSAN Stretched Cluster Witness alone is deployed in the third site (or a different floor in a building depending on the fault domain). This provides guaranteed alignment of Oracle voting disk access (Oracle RAC behavior) and vSAN Stretched Cluster behavior during split-brain situation. With Oracle Database 11gR2, Clusterware is merged with Oracle ASM to create the
Oracle grid infrastructure. The first ASM disk group is created at the installation of Oracle grid infrastructure. It is recommended to use Oracle ASM to store the Oracle Clusterware files (OCR and voting disks) and to have a unique ASM disk group exclusively for Oracle Clusterware files. When creating the ASM disk group for Oracle Clusterware files in a vSAN Stretched Cluster storage, use the external redundancy because vSAN takes care of protection by creating replica of vSAN objects.

4.4 Behavior during Network Partition

In the case of Oracle interconnect partitioning alone, Oracle Clusterware reconfigures based on node majority and accesses to voting disks.

In the case of vSAN Stretched Cluster Network partitioning, vSAN continues IO from one of the available site. The Oracle cluster nodes therefore have access to voting disks only where vSAN Stretched Cluster allows IO to continue and Oracle Clusterware reconfigures the cluster accordingly. Although the voting disks are required for Oracle Extended RAC, you do not need to deploy them at an independent third site because vSAN Stretched Cluster Witness provides split-brain protection and guaranteed behavior alignment of vSAN Stretched Cluster and Oracle Clusterware.

4.5 Deployment Benefits

The benefits of using vSAN Stretched Cluster for Oracle Extended RAC are:

- Scale-out architecture with resources (storage and compute) balanced across both sites.
- Cost-effective Server SAN solution for extended distance.
- Simple deployment of Oracle Extended RAC:
  - Reduced consumption of Oracle Cluster node CPU cycles associated with host-based mirroring. Instead, vSAN takes care of replicating the data across sites.
  - Elimination of Oracle Server and Clusterware at the third site.
  - Easy to deploy the preconfigured witness appliance provided by VMware and can be used as vSAN Stretched Cluster Witness at the third site.
  - Simple infrastructure requirements with deployment of Oracle voting disk on vSAN Stretched Cluster datastore. No need for NFS storage at the third site for arbitration.
- vSAN Stretched Cluster offers ease of deployment with easy enablement and without additional software or hardware.
- Integrates well with other VMware features like VMware vSphere vMotion® and vSphere HA.
5. Solution Configuration

This section introduces the resources and configurations for the solution including solution configuration, architecture diagram and hardware & software resources.
5.1 Solution Configuration

This section introduces the resources and configurations for the solution including:

- Architecture diagram
- Hardware resources
- Software resources
- Network configuration
- VMware ESXi Servers
- vSAN configuration
- Oracle RAC VM and database storage configuration

5.2 Architecture Diagram

The key designs for the vSAN Cluster (Hybrid) solution for Oracle RAC are:

- A 4-node vSAN Cluster with two vSAN disk groups in each ESXi host. Each disk group is created from one 800GB SSD and five 1.2TB HDDs.
- Four Oracle Enterprise Linux VMs, each in one ESXi host to form an Oracle RAC Cluster. Each VM has 8 vCPUs and 64GB of memory with 28GB assigned to Oracle system global area (SGA). The database size is 350GB.

Figure 3. vSAN Cluster for Oracle RAC
The key designs for the vSAN Stretched Cluster solution for Oracle Extended RAC are:

- The vSAN Stretched Cluster consists of five (2+2+1) ESXi hosts. Site A and Site B has two ESXi hosts each, each ESXi host has two vSAN disk groups. Each disk group is created from one 800GB SSD and five 1.2TB HDDs. The vSAN Stretched witness site (site C) has a nested ESXi host VM as witness. Both physical ESXi host and virtual appliance (in the form of nested ESXi) can be used as the witness host. In this solution, we used a preconfigured witness appliance provided by VMware as the witness.
- Two Oracle Enterprise Linux VMs each in site A and site B form an Oracle Extended RAC Cluster. Each VM has 8 vCPUs and 64GB of memory with 28GB assigned for Oracle SGA. The databases size is 350GB.

5.3 Hardware Resources

Table 2 shows the hardware resources used in this solution.

Table 2. Hardware Resources

<table>
<thead>
<tr>
<th>SOLUTION</th>
<th>CONFIGURATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>vSAN</td>
<td>4 ESXi Hosts</td>
</tr>
</tbody>
</table>
5.4 Software Resources

Table 3 shows the software resources used in this solution.

Table 3. Software Resources

<table>
<thead>
<tr>
<th>SOFTWARE</th>
<th>VERSION</th>
<th>PURPOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle Enterprise Linux (OEL)</td>
<td>6.6</td>
<td>Oracle Database Server Nodes</td>
</tr>
<tr>
<td>VMware vCenter Server and ESXi</td>
<td>6.0 U1</td>
<td>ESXi cluster to host virtual machines and provide vSAN Cluster. VMware vCenter Server provides a centralized platform for managing VMware vSphere environments.</td>
</tr>
<tr>
<td>VMware vSAN</td>
<td>6.1</td>
<td>Software-defined storage solution for hyper-converged infrastructure</td>
</tr>
<tr>
<td>Oracle 11gR2 Grid Infrastructure</td>
<td>11.2.0.4</td>
<td>Oracle Database and Cluster software</td>
</tr>
<tr>
<td>Oracle Workload Generator</td>
<td>Swingbench 2.5</td>
<td>To generate Oracle workload</td>
</tr>
<tr>
<td>Linux Netem</td>
<td>OEL 6.6</td>
<td>Provides network latency emulation used for Stretched Cluster setup in the lab</td>
</tr>
<tr>
<td>XORP (open source routing platform)</td>
<td>1.8.5</td>
<td>To enable Multicast traffic routing between vSAN Stretched Cluster hosts in different VLANs (between two sites)</td>
</tr>
</tbody>
</table>

5.5 Network Configuration

A VMware vSphere Distributed Switch™ acts as a single virtual switch across all associated hosts in the data cluster. This setup allows virtual machines to maintain a consistent network configuration as they migrate across multiple hosts. The vSphere Distributed Switch uses two 10GbE adapters per host as shown in Figure 5.
Figure 5. vSphere Distributed Switch Port Configuration

Figure 6. Distributed Switch Configuration View from one of the ESXi Host

A port group defines properties regarding security, traffic shaping, and NIC teaming. Default port group setting was used except the uplink failover order which is shown in Table 4. It also shows the
distributed switch port groups created for different functions and the respective active and standby uplink to balance traffic across the available uplinks.

Table 4. Distributed Switch Port Groups

<table>
<thead>
<tr>
<th>DISTRIBUTED SWITCH PORT GROUP NAME</th>
<th>VLAN ID</th>
<th>ACTIVE UPLINK</th>
<th>STANDBY UPLINK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle Private (RAC Interconnect)</td>
<td>4,033</td>
<td>Uplink1</td>
<td>Uplink2</td>
</tr>
<tr>
<td>Oracle Public</td>
<td>1,100</td>
<td>Uplink1</td>
<td>Uplink2</td>
</tr>
<tr>
<td>vSAN</td>
<td>4,030</td>
<td>Uplink2</td>
<td>Uplink1</td>
</tr>
<tr>
<td>vMotion</td>
<td>4,035</td>
<td>Uplink1</td>
<td>Uplink2</td>
</tr>
</tbody>
</table>

5.6 VMware ESXi Servers

Each vSAN ESXi Server in the vSAN Cluster has the following configuration as shown in Figure 7.

Figure 7. vSAN Cluster Node

Storage Controller Mode

The storage controller used in the reference architecture supports pass-through mode. The pass-through mode is the preferred mode for vSAN and it gives vSAN complete control of the local SSDs and HDDs attached to the storage controller.

5.7 vSAN Configuration
vSAN Storage Policy

vSAN can set availability, capacity, and performance policies per virtual machine. Table 5 shows the designed and implemented storage policy.

Table 5. vSAN Storage Setting for Oracle Database

<table>
<thead>
<tr>
<th>STORAGE CAPABILITY</th>
<th>SETTING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of failure to tolerate (FTT)</td>
<td>1</td>
</tr>
<tr>
<td>Number of disk stripes per object</td>
<td>2</td>
</tr>
<tr>
<td>Flash read cache reservation</td>
<td>0%</td>
</tr>
<tr>
<td>Object space reservation</td>
<td>100%</td>
</tr>
</tbody>
</table>

Number of Failures to tolerate:

Defines the number of host, disk, or network failures a virtual machine object can tolerate. For \( n \) failures tolerated, \( n+1 \) copies of the virtual machine object are created and \( 2n+1 \) hosts with storage are required. The settings applied to the virtual machines on the vSAN datastore determines the datastore’s usable capacity.

Object space reservation:

Percentage of the object logical size that should be reserved during the object creation. The default value is 0 percent and the maximum value is 100 percent.

Number of disk stripes per object:

This policy defines how many physical disks across each copy of a storage object are striped. The default value is 1 and the maximum value is 12.

Flash read cache reservation:

Flash capacity reserved as read cache for the virtual machine object. Specified as a percentage of the logical size of the VMDK object. It is set to 0 percent by default and vSAN dynamically allocates read cache to storage objects on demand.
5.8 Oracle RAC VM and Database Storage Configuration

Each Oracle RAC VM is installed with Oracle Enterprise Linux 6.6 (OEL) and configured with 8 vCPUs and 64GB memory with 28GB assigned to Oracle SGA. We used this configuration all through the tests unless specified otherwise.

Oracle ASM data disk group with external redundancy is configured with allocation unit (AU) size of 1M. Data, Fast Recovery Area (FRA), and Redo ASM disk groups are present on different PVSCSI controllers. Archive log destination uses FRA disk group. Table 6 provides Oracle RAC VM disk layout and ASM disk group configuration.

### Table 6. Oracle RAC VM Disk Layout

<table>
<thead>
<tr>
<th>VIRTUAL DISK</th>
<th>SCSI CONTROLLER</th>
<th>SIZE (GB) X NUMBER OF VMDK</th>
<th>TOTAL STORAGE (GB)</th>
<th>ASM DISK GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS and Oracle Binary</td>
<td>SCSI 0</td>
<td>100 x 1</td>
<td>100</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>Database data disks</td>
<td>SCSI 1</td>
<td>60 x 10</td>
<td>600</td>
<td>DATA</td>
</tr>
<tr>
<td>FRA</td>
<td>SCSI 2</td>
<td>40 x 4</td>
<td>160</td>
<td>FRA</td>
</tr>
<tr>
<td>Online redo log</td>
<td>SCSI 3</td>
<td>20 x 3</td>
<td>60</td>
<td>REDO</td>
</tr>
<tr>
<td>CRS, voting disks</td>
<td>SCSI 3</td>
<td>20 x 3</td>
<td>60</td>
<td>CRS</td>
</tr>
</tbody>
</table>

Storage provisioning from vSAN to Oracle:

- Oracle RAC requires attaching the shared disks to one or more VMs, which requires multi-writer support for all applicable virtual machines and VMDKs. To enable in-guest systems leveraging cluster-aware file systems that have distributed write (multi-writer) capability, we must explicitly enable multi-writer support for all applicable virtual machines and VMDKs.
- Create a VM storage policy that is applied to the virtual disks used as the Oracle RAC cluster’s shared storage. See Table 5 for vSAN storage policy used in this reference architecture.
- Create shared virtual disks in the eager-zeroed thick and independent persistent mode.
- Attach the shared disks to one or more VMs.
- Enable the multi-writer mode for the VMs and disks.
- Apply the VM storage policy to the shared disks.

See the VMware Knowledge Base Article 2121181 for detailed steps to enable multi-writer and provisioning storage to Oracle RAC from vSAN.

Each Oracle RAC VM is installed with Oracle Enterprise Linux 6.6 (OEL) and configured with 8 vCPUs and 64GB memory with 28GB assigned to Oracle SGA. We used this configuration all through the tests unless specified otherwise.

Oracle ASM data disk group with external redundancy is configured with allocation unit (AU) size of 1M. Data, Fast Recovery Area (FRA), and Redo ASM disk groups are present on different PVSCSI controllers. Archive log destination uses FRA disk group. Table 6 provides Oracle RAC VM disk layout and ASM disk group configuration.

### Table 6. Oracle RAC VM Disk Layout
6. Solution Validation

In this section, we present the test methodologies and processes used in this solution.
6.1 Solution Validation

The solution designed and deployed Oracle 11gR2 RAC on a vSAN Cluster focusing on ease of use, performance, resiliency, and availability. In this section, we present the test methodology, processes, and results for each test scenario.

6.2 Test Overview

The solution validated the performance and functionality of Oracle RAC instances in a virtualized VMware environment running Oracle 11gR2 RAC backed by the vSAN storage platform.

The solution tests include:

- Oracle workload testing using Swingbench to generate classic-order-entry TPC-C like workload and observe the database and vSAN performance.
- RAC scalability testing to support an enterprise-class Oracle RAC Database with vSAN cost-effective storage.
- Resiliency testing to ensure Oracle RAC on vSAN storage solution is highly available and supports business continuity.
- vSphere vMotion testing on vSAN.
- Oracle Extended RAC performance test on vSAN Stretched Cluster to verify acceptable performance across geographically distributed sites.
- Continuous application availability test during a site failure in vSAN Stretched Cluster.
- Metro and global availability test with vSAN Stretched Cluster and Oracle Data Guard.
- Oracle database backup and recovery test on vSAN using Oracle RMAN.

6.3 Test and Performance Data Collection Tools

Test Tool

We used Swingbench to generate oracle OLTP workload (TPC-C like workload). Swingbench is an Oracle workload generator designed to stress test an Oracle database. Oracle RAC is set up with SCAN (Single Client Access Name) configured in DNS with three IP addresses. SCAN provides load balancing and failover for client connections to the database. SCAN works as a cluster alias for databases in the cluster.

In all tests, Swingbench accesses the Oracle RAC database using EZConnect client and the simple JDBC thin URL. Unless mentioned in the respective test, Swingbench was set up to generate TPC-C like workload using 100-user sessions on Oracle RAC Database.

Performance Data Collection Tools

We used the following testing and monitoring tools in this solution:

- vSAN Observer

  vSAN Observer is designed to capture performance statistics and bandwidth for a VMware vSAN Cluster. It provides an in-depth snapshot of IOPS, bandwidth and latencies at different layers of vSAN, read cache hits and misses ratio, outstanding I/Os, and congestion. This information is provided at different layers in the vSAN stack to help troubleshoot storage performance. For more information about the VMware vSAN Observer, see the Monitoring VMware vSAN with vSAN Observer documentation.

- esxtop utility
esxtop is a command-line utility that provides a detailed view on the ESXi resource usage. Refer to the VMware Knowledge Base Article 1008205 for more information.

- Oracle AWR reports with Automatic Database Diagnostic Monitor (ADDM)

Automatic Workload Repository (AWR) collects, processes, and maintains performance statistics for problem detection and self-tuning purposes for Oracle database. This tool can generate report for analyzing Oracle performance.

The Automatic Database Diagnostic Monitor (ADDM) analyzes data in the Automatic Workload Repository (AWR) to identify potential performance bottlenecks. For each of the identified issues, it locates the root cause and provides recommendations for correcting the problem.

6.4 Oracle RAC Performance on vSAN

Test Overview

This test focused on the Oracle workload behavior on vSAN using Swingbench TPC-C like workload. We configured Swingbench to generate workload using 100 user sessions on a 4-node Oracle RAC. Figure 3 shows the configuration details.

Test Results

Swingbench reported the maximum transactions per minute (TPM) of 331,000 with average of 287,000 as shown in Figure 8. From the storage perspective, vSAN Observer and Oracle AWR reported average IOPS of 28,000 and average throughput of 304 MB/s. The four RAC nodes are spread across the four ESXi host. Therefore, Figure 9 vSAN Observer Client View shows IOPS (~7,000) and throughput (~76MB/s) equally distributed across the four ESXi hosts. Table 7 shows data captured from Oracle AWR report that provides Oracle workload read to write ratio (70:30). We ran multiple tests with similar workload and observed similar results. During this workload, Figure 9 shows the overall IO response time less than 2ms. The test results demonstrated vSAN as a viable storage solution for Oracle RAC.

Figure 8. Swingbench OLTP Workload on 4-Node RAC—TPMs
Figure 9. IO Metrics from vSAN Observer during Swingbench Workload on a 4-Node RAC

Table 7. IO Workload from Oracle AWR Report

<table>
<thead>
<tr>
<th>PHYSICAL IO FROM ORACLE AWR FOR 15-MIN SNAPSHOT (AVERAGE)</th>
<th>4-NODE RAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total IOPS</td>
<td>27,760</td>
</tr>
<tr>
<td>Read IOPS</td>
<td>19,800</td>
</tr>
<tr>
<td>Write IOPS</td>
<td>7,960</td>
</tr>
</tbody>
</table>

6.5 Oracle RAC Scalability on vSAN

Test Overview

Scalability is one of the major benefits of the Oracle RAC databases. Because the performance of a RAC database increases as additional nodes are added, the underlying shared storage should be able to provide increased IOPS and throughput under low latency accordingly. To demonstrate scalability, four different tests were run starting with single instance and gradually scaling up to 4-node RAC, the same number of user sessions was set for all four tests and the average TPM noted for each.

Test Results

As shown in Figure 10, the average TPM increased linearly from the single-instance database to the 4-node RAC database. The TPS shown in Figure 9 are aggregate TPS observed across all Oracle instance in a RAC. We observed linear increase in IOPS and throughput with overall latency less than 2ms throughout. The TPM values shown in Figure 10 are aggregate of all the instances in the RAC database. This demonstrated vSAN as a scalable storage solution for Oracle RAC database.
6.6 vSAN Resiliency

Test Overview

This section validates vSAN resiliency in handling disk, disk group, and host failures. We designed the following scenarios to emulate potential real-world component failures:

- **Disk failure**
  
  This test evaluated the impact on the virtualized Oracle RAC database when encountering one HDD failure. The HDD stored the VMDK component of the oracle database. Hot-remove (or hot-unplug) the HDD to simulate a disk failure on one of the nodes of the vSAN Cluster to observe whether it has functional or performance impact on the production Oracle database.

- **Disk group failure**
  
  This test evaluated the impact on the virtualized Oracle RAC database when encountering a disk group failure. Hot-remove (or hot-unplug) one SSD to simulate a disk group failure to observe whether it has functional or performance impact on the production Oracle database.

- **Storage host failure**
  
  This test evaluated the impact on the virtualized Oracle RAC database when encountering one vSAN host failure. Shut down one host in the vSAN Cluster to simulate host failure to observe whether it has functional or performance impact on the production Oracle database.

Test Scenarios

**Single Disk Failure**

We simulated an HDD disk failure by hot-removing a disk when Swingbench was generating workload on a 4-node RAC. Table 8 shows the failed disk and it had four components that stored the VMDK
(Oracle ASM disks) for Oracle database. The components residing on the disk are absent and inaccessible after the disk failure.

Table 8. HDD Disk Failure

<table>
<thead>
<tr>
<th>FAILED DISK NAA ID</th>
<th>ESXI HOST</th>
<th>NO OF COMPONENTS</th>
<th>TOTAL CAPACITY</th>
<th>USED CAPACITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>naa.5XXXXXXXXXXXXXc077</td>
<td>1xx.xx.28.11</td>
<td>4</td>
<td>1106.62 GB</td>
<td>7.23%</td>
</tr>
</tbody>
</table>

Disk Group Failure

We simulated a disk group failure by hot removing the SSD from the disk group when Swingbench was generating load on a 4-node RAC. The failure of SSD rendered an entire disk group inaccessible.

The failed disk group had the SSD and HDD backing disks as shown in Table 9. The table also shows the number of affected components that stored the VMDK file for Oracle database.

Table 9. Failed vSAN Disk Group—Physical Disks and Components

<table>
<thead>
<tr>
<th>DISK DISPLAY NAME</th>
<th>ESXI HOST</th>
<th>DISK TYPE</th>
<th>NO OF COMPONENTS</th>
<th>TOTAL CAPACITY (GB)</th>
<th>USED CAPACITY (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>naa.5XXXXXXXXXXXXX8935</td>
<td>1xx.xx.28.3</td>
<td>SSD</td>
<td>NA</td>
<td>745.21</td>
<td>NA</td>
</tr>
<tr>
<td>naa.5XXXXXXXXXXXXXd4d7</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>3</td>
<td>1106.62</td>
<td>4.57</td>
</tr>
<tr>
<td>naa.5XXXXXXXXXXXXXc56f</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>4</td>
<td>1106.62</td>
<td>6.51</td>
</tr>
<tr>
<td>naa.5XXXXXXXXXXXXXd7f7</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>3</td>
<td>1106.62</td>
<td>5.42</td>
</tr>
<tr>
<td>naa.5XXXXXXXXXXXXXd12b</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>1</td>
<td>1106.62</td>
<td>1.81</td>
</tr>
<tr>
<td>naa.5XXXXXXXXXXXXXe54b</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>2</td>
<td>1106.62</td>
<td>4.52</td>
</tr>
</tbody>
</table>

Storage Host Failure

One ESXi Server that provided capacity to vSAN was shutdown to simulate host failure when Swingbench was generating workload on a 3-node Oracle RAC. The failed ESXi node did not host Oracle RAC VM. This was consciously set up to understand the impact of losing a storage node alone on vSAN and avoid Oracle instance crash and VMware HA action impact on the workload.

The failed storage node had two disk groups with the following SSD and HDD backing disks as shown in Table 10.

Table 10. Failed ESXi Host vSAN Disk Group—Physical Disks and Components
<table>
<thead>
<tr>
<th>DISK GROUP</th>
<th>DISK DISPLAY NAME</th>
<th>ESXi HOST</th>
<th>ESXi HOST</th>
<th>NO OF COMPONENTS</th>
<th>TOTAL CAPACITY (GB)</th>
<th>USED CAPACITY (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diskgroup - 1</td>
<td>naa.5XXXXXXXXXXX8104</td>
<td>1xx.xx.28.3</td>
<td>SSD</td>
<td>NA</td>
<td>745.21</td>
<td>NA</td>
</tr>
<tr>
<td>Diskgroup - 1</td>
<td>naa.5XXXXXXXXXXXc3e7</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>2</td>
<td>1106.62</td>
<td>8.13</td>
</tr>
<tr>
<td>Diskgroup - 1</td>
<td>naa.XXXXXXXXXXX2153</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>4</td>
<td>1106.62</td>
<td>10.30</td>
</tr>
<tr>
<td>Diskgroup - 1</td>
<td>naa.XXXXXXXXXXXb74f</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>3</td>
<td>1106.62</td>
<td>9.94</td>
</tr>
<tr>
<td>Diskgroup - 1</td>
<td>naa.XXXXXXXXXXX28c3</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>4</td>
<td>1106.62</td>
<td>5.46</td>
</tr>
<tr>
<td>Diskgroup - 1</td>
<td>naa.5XXXXXXXXXXXXb53f</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>4</td>
<td>1106.62</td>
<td>7.26</td>
</tr>
<tr>
<td>Diskgroup - 2</td>
<td>naa.5XXXXXXXXXXX8935</td>
<td>1xx.xx.28.3</td>
<td>SSD</td>
<td>NA</td>
<td>745.21</td>
<td>NA</td>
</tr>
<tr>
<td>Diskgroup - 2</td>
<td>naa.5XXXXXXXXXXXd4d7</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>3</td>
<td>1106.62</td>
<td>6.33</td>
</tr>
<tr>
<td>Diskgroup - 2</td>
<td>naa.5XXXXXXXXXXc56f</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>3</td>
<td>1106.62</td>
<td>7.32</td>
</tr>
<tr>
<td>Diskgroup - 2</td>
<td>naa.5XXXXXXXXXXXd7f7</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>2</td>
<td>1106.62</td>
<td>4.52</td>
</tr>
<tr>
<td>Diskgroup - 2</td>
<td>naa.5XXXXXXXXXXd12b</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>3</td>
<td>1106.62</td>
<td>3.62</td>
</tr>
<tr>
<td>Diskgroup - 2</td>
<td>naa.5XXXXXXXXXXe54b</td>
<td>1xx.xx.28.3</td>
<td>HDD</td>
<td>3</td>
<td>1106.62</td>
<td>8.14</td>
</tr>
</tbody>
</table>

**Test Results**

As shown in Table 11, during all types of failures, performance was affected only by a momentary drop in TPS and increase in IO wait. Time taken for recovery to steady state TPS is also shown in Table 11. In all three failure tests, the steady state TPS after failure value is approximately same as the value before failure. None of the test reported IO error in the Linux VMs or Oracle user-session disconnects, which demonstrated the resiliency of vSAN during the component failures.

Table 11. Impact on Oracle Workload during Failure
### Failure Type

<table>
<thead>
<tr>
<th>Failure Type</th>
<th>Oracle TPS Before Failure</th>
<th>Lowest Oracle TPS After Failure</th>
<th>Time Taken for Recovery to Steady State TPS After Failure (Sec)</th>
<th>Redo Disk (VMDK) Write IOPS Decrease at the Moment of Failure</th>
<th>Redo Disk (VMDK) Write Response Time Increase at the Moment of Failure (MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single disk (HDD)</td>
<td>4,100</td>
<td>1,200</td>
<td>29</td>
<td>250 to 110</td>
<td>1.6 to 2.4</td>
</tr>
<tr>
<td>Disk group</td>
<td>3,900</td>
<td>1,000</td>
<td>51</td>
<td>240 to 110</td>
<td>1.7 to 3.2</td>
</tr>
<tr>
<td>Storage host</td>
<td>3,600</td>
<td>700</td>
<td>165</td>
<td>170 to 48</td>
<td>1.7 to 4.3</td>
</tr>
</tbody>
</table>

After each of the failures, because the VM storage policy was set with “Number of FTT” greater than zero, the virtual machine objects and components remained accessible. In the disk failure tests, because the disks are hot removed and it is not a permanent failure, vSAN knows the disk is just removed so new objects rebuild will not start immediately. It will wait for 60 minutes as the default repair delay time. See the [VMware Knowledge Base Article 2075456](https://kb.vmware.com/kb/2075456) for steps to change the default repair delay time. If the removed disk is not inserted within 60 minutes, the object rebuild operation is triggered if additional capacity is available within the cluster to satisfy the object requirement. However, if the disk fails due to unrecoverable error this is considered as a permanent failure, vSAN immediately responds by rebuilding a disk object. In case of a host failure to avoid unnecessary data resynchronization during the host maintenance, synchronization starts after 60 minutes as the default repair delay value. The synchronization operation time depends on the amount of data that needs to be resynchronized.

### 6.7 vSphere vMotion on vSAN

**Test Overview**

vSphere vMotion live migration allows moving an entire virtual machine from one physical server to another without downtime. We used this feature to distribute and transfer workloads of Oracle RAC configurations seamlessly between the ESXi hosts in a vSphere cluster. In this test, we performed vMotion to migrate one of the 4-node Oracle RAC from one ESXi host to another in the vSAN Cluster as shown in Figure 11.
Figure 11. vSphere vMotion Migration of Oracle RAC VM on vSAN

Test Results and Observations

While Swingbench was generating TPC-C like workload, vMotion was initiated to migrate RAC Node 2 from ESXi 2 to ESXi 3 in vSAN Cluster. When the migration was started, there was approximately 300,000 TPM on the Oracle RAC database. The migration took 182 seconds to finish. TPM reduced for 10 seconds during the last phase of the migration and was back to the normal level afterwards. In the test, ESXi 3 had sufficient resources to accommodate two Oracle RAC VMs so no drop in aggregate transactions was observed after vMotion completed. This test demonstrated mobility of Oracle RAC VMs deployed in a vSAN Cluster.

6.8 Oracle Extended RAC Performance on vSAN

Test Overview and Setup

We tested Oracle Extended RAC on vSAN Stretched Cluster in this solution. A 2-node Oracle Extended RAC was setup as shown in Figure 4.

vSAN Stretched Cluster Setup

A metropolitan area network was simulated in the lab environment. Figure 12 shows a diagrammatic layout of the setup. The three sites were placed in different VLANs. A Linux VM configured with three network interfaces each in one of the three VLANs acting as the gateway for inter-VLAN routing between sites. Static routes were configured on ESXi vSAN VMkernel ports for routing between different VLANs (sites). The Linux VM leverages Netem functionality already built into Linux for simulating network latency between the sites. Furthermore, XORP installed on the Linux VM provided support for a multicast traffic between two vSAN fault domains. While the network latency between data sites was varied to compare performance impact, inter-site round trip latency from the Witness to data site was kept at 200ms.
Latency Setup between Oracle RAC Nodes

Along with the delay introduced between the sites for vSAN kernel ports, latency was also simulated between the 2-node Oracle RAC (public and private interconnects) using the Netem functionality inbuilt in Linux (Oracle RAC VM).

Figure 13 shows the Stretched Cluster fault domain configuration. We configured and tested three inter-site round trip latency 1ms, 2.2ms, and 4.2ms. We used Swingbench to generate the same TPC-C like workload in these tests.
Test Results

During these tests, we recorded Oracle TPS in Swingbench and IOPS metrics on the vSAN Stretched Cluster. With the increase of inter-site latency using Netem functionality discussed in the above section, we observed a reduction in TPS in Oracle RAC. For the test workload, the TPS reduction was proportional to the increase in inter-site round trip latency. As shown in Figure 14, for 1ms, 2.2ms, and 4.2ms round trip latency, the TPS reduced by 12 percent, 27 percent and 47 percent respectively. Similarly, we observed an IOPS reduction in vSAN Observer as shown in Figure 15. We noticed that when inter-site latency increased, IO and Oracle cache-fusion message latency increased as a result. This solution proved that Oracle Extended RAC on vSAN Stretched Cluster provided reasonable performance for OLTP workload. vSAN Stretched Cluster with 1ms inter-site round trip latency (typically 100km distance) is capable of delivering 88 percent of the transaction rate for Oracle Extended RAC when compared to regular vSAN Clusters. As distance and latency between sites increase, performance gets impacted accordingly.

![Figure 14. TPS Comparison for a TPC-C like Workload on Oracle RAC](image)

Figure 14. TPS Comparison for a TPC-C like Workload on Oracle RAC
6.9 vSAN Stretched Cluster Site Failure Tolerance

Site Failure Test Overview

This test demonstrated one of the powerful features of the vSAN Streched Cluster: maintaining data availability even under the impact of complete site failure.

We ran this test on the 2-node Oracle Extended RAC as shown in Figure 4. We enabled vSphere HA and vSphere DRS on the cluster. And we generated Swingbench TPC-C like workload with 100-user sessions on the RAC cluster. After some period, an entire site was down by powering off two ESXi hosts in site A (the vSAN Stretched Cluster preferred site) that is represented as “Not Responding” in Figure 16.

![Figure 15. IOPS Comparison for a TPC-C like Workload on Oracle RAC](image-url)
Site Failure Test Results and Continuous Data Availability

The RAC VM in site A was affected after the site failure. However, transactions continued even though the RAC instance in site A went down and the client connections failed over to the RAC VM on the surviving site B. vSphere HA then restarted the RAC VM on the surviving site B and it was ready to take up connections. The site outage did not affect data availability because a copy of all the data in site A existed in site B. Thus, the affected RAC VM was automatically restarted by vSphere HA without any issues. After site failure, data redundancy could not be maintained because site A was no longer active and all RAC VM are on same site so no Oracle RAC cache-fusion message latency as a result database servered more transactions. However, this performance came at a cost as none of the newly written data since the site failure had a backup. The Streched Cluster is designed to tolerate single site failure at a time.

After some time, site A was available by powering on both ESXi hosts. The vSAN Streched Cluster started the process of site A synchronization, with the changed components in site B after the failure. The results demonstrated how vSAN Streched Cluster provided availability during site failure by automating the failover and failback process leveraging vSphere HA and vSphere DRS. This proves vSAN Streched Cluster’s ability to survive a complete site failure offering zero RPO and RTO for Oracle RAC database.

Best Practices for Site Failure and Recovery

Performance after site failure depends on adequate resources such as CPU and memory to be available to accommodate the virtual machines that are restarted by vSphere HA on the surviving site. If the surviving RAC VM cannot support the additional load from the failed site RAC VM and there is no extra resource on the ESXi hosts in the surviving site, disable vSphere HA and Oracle RAC will take care of HA at the database level by failing over client connections to the surviving RAC VM. In the event of a site failure, instead of bringing up the failed vSAN hosts one by one, it is recommended to bring all hosts online approximately at the same time within a span of 10 minutes. The vSAN Streched Cluster waits for 10 minutes before starting the recovery traffic to a remote site. This avoids repeatedly resynchronizing a large amount of data across the sites. After the site is recovered, it is recommended to wait for the recovery traffic to complete before migrating virtual machines to the recovered site. Similarly, it is recommended to change vSphere DRS policy from fully automated to partially automated in the event of a site failure.
6.10 Metro and Global DR with vSAN Stretched

Solution and Setup Overview

Mission-critical Oracle RAC applications require continuous data availability to prevent planned and unplanned outages or to replicate data to anywhere in the world for disaster recovery. This zero data loss disaster recovery solution is achieved using the combination of vSAN Stretched Cluster and Oracle Data Guard. In this solution, vSAN Stretched Cluster delivered active/active continuous availability at a metro distance and Oracle Data Guard provided replication and recovery at a global distance.

Figure 17 shows the setup environment:

- Site A and site B were the data sites of vSAN Stretched Cluster, which had the Oracle Extended RAC nodes distributed across them.
- Site C hosted the Witness for vSAN Stretched Cluster.
- Site A, B, and C together form Oracle Extended RAC Cluster that hosts the production database that functioned as the primary role in Oracle Data Guard configuration.
- Site D had a 2-node Oracle RAC on a regular vSAN Cluster for disaster recovery, which was ideally located at a global distance (however, this was set up on the same data center in the lab for demonstration) for disaster recovery. The 2-node Oracle RAC in site D was a physical standby database in Oracle Data Guard configuration. Oracle Active Data Guard was setup between the primary and physical standby databases with the protection mode in Maximum Performance.

Figure 17. Metro Global DR with vSAN Stretched Cluster and Oracle Data Guard

Solution Validation

Oracle TPC-C like workload was generated on the primary database using Swingbench. With the Data Guard setup in the Maximum Performance mode, the transactions were committed as soon as all the redo data generated by the transactions has been written to the online log. The redo data was also written to the standby databases asynchronously with respect to the transaction committed. This
protection mode ensures the primary database performance is unaffected by delays in writing redo data to the standby database. Ideally suited when the latency between the production and the DR site is more and bandwidth is limited.

With Oracle Active Data Guard, a physical standby database can be used for real-time reporting, with minimal latency between reporting and production data. Furthermore, it also allows backup operations to be off-loaded to the standby database. As shown in Figure 17, RMAN backup is taken from standby database in Site D which is registered and managed from RMAN catalog Database. This enables efficient utilization of vSAN storage resource at the DR site increasing overall performance and return on the storage investment.

Another protection mode that can be considered is Maximum Availability. This mode provides the highest level of data protection that is possible without compromising the availability of a primary database. Transactions do not commit until all redo data needed to recover those transactions has been written to the online redo log and to the standby redo log at the standby database, so it is synchronized. If the primary database cannot write its redo stream to at least one synchronized standby database, it operates as if it were in the maximum performance mode to preserve the primary database availability until it is able to write its redo stream again to a synchronized standby database.

This solution demonstrated how vSAN Stretched Cluster hosting a primary database and vSAN hosting a standby database together with Oracle Data Guard provided a cost-effective solution with the highest level of availability across three data centers, offering near-zero data loss during a disaster at any one of the sites.

6.11 Oracle RAC Database Backup and Recovery on vSAN

Backup Solution Overview

RMAN is an Oracle utility that can back up, restore, and recover database files. It is a feature of the Oracle database server and does not require separate installation. RMAN takes care of all underlying database procedures before and after backup or recovery, freeing dependency on operating system and SQL*Plus scripts. In this solution, we used Oracle RMAN for database backup and recovery and we demonstrated the following backup scenarios:

- Oracle RAC database backup from production site when there is workload on the database. In this scenario, there is no DR site or oracle data guard configuration.
- Oracle RAC database backup from physical standby database (DR site) when there is workload on the primary site. We implemented the solution as shown in Figure 17. This method offloads backup to the standby database.

RMAN accesses backup and recovery information from either the control file or the optional recovery catalog. Having a separate recovery catalog is a preferred method in the production environment as it serves as a secondary metadata repository and can centralize metadata for all the target database. A recovery catalog is required when you use RMAN in a Data Guard environment. By storing backup metadata for all the primary and standby databases, the catalog enables you to offload backup tasks to one standby database while enabling you to restore backups on other databases in the environment.

In the lab environment, RMAN recovery catalog database is installed on a separate virtual machine as shown in Figure 17. An NFS mount point is used to store the actual backup (disk-based backup). The RMAN catalog database and NFS mount point are stored in a separate vSAN datastore where all the infrastructure components are hosted. A separate backup network interface and VLAN is used for backup traffic, so we can isolate the backup traffic from the Oracle public and RAC interconnect traffic.

Backup Solution Validation

When we ran the BACKUP command in RMAN, the output is always either backup sets or image copies. A backup set is an RMAN-specific proprietary format, whereas an image copy is a bit-for-bit copy of a file. By default, RMAN creates backup sets that are used in the following tests.
Oracle RAC database backup from the production site

We configured Swingbench to generate TPC-C like workload using 100-user sessions on a 4-node Oracle RAC. Oracle TPS was between 4,500 and 4,800. We initiated the RMAN full backup from one of the production RAC VM (vmorarac1) when there was a workload. On this RAC VM, the read throughput increased from 50 MB/s to 115 MB/s after RMAN backup was started. This read throughput increase was caused by the backup workload. Oracle transactions continued during the backup and the TPS was the same as before starting the backup. Although the backup workload did not affect transaction performance, since RMAN backups used some of the CPU and storage resources, it is recommended to initiate them during off-peak hours or offload backup to the standby database if applicable. Further RMAN’s has incremental backup ability which only backup those database blocks that have changed since a previous backup. This can be used to improve efficiency by reducing the time and resources required for backup and recovery.

Oracle RAC database backup from the physical standby database (DR site)

While Swingbench was generating workload on the 2-node Oracle Extended RAC as shown in Figure 17. The RMAN backup was initiated from the standby database. The primary and the standby databases are installed in different vSAN Clusters. Since the backup is initiated from the standby database, there is no impact to resources on the primary site hence no impact to Oracle transactions on the primary database. The primary and the standby databases are registered with the RMAN catalog database. This allows you to offload backup tasks to the standby database while enabling you to restore backups to the primary database when required.

The tests demonstrate Oracle RMAN backup as a feasible solution for Oracle RAC Database deployed on vSAN.
7. Best Practices of Virtualized Oracle RAC on vSAN

This section provides the best practices to be followed for Virtualized Oracle RAC on vSAN.
7.1 Best Practices of Virtualized Oracle RAC on vSAN

vSAN is key to a successful implementation of enterprise mission-critical Oracle RAC Database. The focus of this reference architecture is VMware vSAN storage best practices for Oracle RAC Database. For detailed information about CPU, memory, and networking configurations refer to the Oracle Databases on VMware Best Practices Guide along with vSphere Performance Best Practices guide for specific version of vSphere.

7.2 vSAN Storage Configuration Guidelines

vSAN is distributed object-store datastore formed from locally attached devices from ESXi host. It uses disk groups to pool together flash devices and magnetic disks as single management constructs. It is recommended to use similarly configured and sized ESXi hosts for the vSAN Cluster:

- Design for growth: consider initial deployment with capacity in the cluster for future growth and enough flash cache to accommodate future requirements. Use multiple vSAN disk groups per server with enough magnetic spindles and SSD capacity behind each disk group. For future capacity addition, create disk groups with similar configuration and sizing. This ensures an even balance of virtual machine storage components across the cluster of disks and hosts.
- Design for availability: consider design with more than three hosts and additional capacity that enable the cluster to automatically remediate in the event of a failure.
- vSAN SPBM can set availability, capacity, and performance policies per virtual machine. Number of disk stripes per object and object space reservation are the storage policies that were changed from the default values for Oracle RAC VMs in this reference architecture:
  - When using the multi-writer mode, the virtual disk must be eager-zeroed thick. Since the disk is thick provisioned, 100 percent capacity is reserved automatically, so the object space reservation is set to 100 percent.
  - Number of disk stripes per object—with the increase in stripe width, you may notice I/O performance improvement because objects spread across more vSAN disk groups and disks. However, in a solution like Oracle RAC where we recommend multiple VMDKs for database, the objects are spread across vSAN Cluster components even with the default stripe width of 1. So increasing the vSAN stripe width might not provide tangible benefits. Moreover, there is an additional ASM striping at the host level as well. Hence it is recommended to use the default stripe width of 1 unless there are performance issues observed during read cache misses or during destaging. In our Oracle RAC tests, stripe width of 2 provided marginal increase in Oracle TPS when compared to the stripe width of 1. However, further increase in stripe width did not provide any benefits, so stripe width of 2 was used in this testing.
  - Create separate Oracle ASM disk groups for Data, Online Redo and Archive (FRA).
  - Depending on the database size, it is recommended to have multiple VMDKs in each Oracle ASM disk group. Multiple VMDKs in each ASM disk group spread database components across disk groups in vSAN Cluster.
  - Oracle ASM data distribution policy is capacity based. Make sure that Oracle ASM disks (VMDKs) in an ASM disk group have the same capacity to maintain balance.
  - Use Paravirtualized SCSI adapter for disks that store Oracle database and separate PVSCSI controllers for Data, REDO and FRA VMDKs. In this solution, Table 6 shows how separate PVSCSI controllers are used for Oracle ASM disk group and multiple VMDKs in each ASM Disk group.
  - The large-scale workloads with intensive I/O patterns require adapter queue depths greater than the Paravirtual SCSI (PVSCSI) default values. In the testing environment, we appended vmw_pvscsi.cmd_per_lun=254 and vmw_pvscsi.ring_pages=32 to default Oracle Linux kernel boot arguments. This modification increased the number of pages used by the PVSCSI adapter for the request ring to 32. It also increased the queue depth of the devices attached to the PVSCSI adapter to 254. See the VMware Knowledge Base Article 2053145 for more information.
  - Use the NOOP value as the I/O scheduler for Linux kernel. This configuration value performs better for virtualized Linux guests and is recommended for Oracle database. In the testing environment, default I/O scheduler cfq was changed to "elevator=noop" by
appending it to default Oracle Linux kernel boot argument. See the [VMware Knowledge Base Article 2011861](https://v2xknowledgecenter.vmware.com/article/2011861) for more information.

### 7.3 CPU Configuration Guidelines

This section provides guidelines for CPU configuration for Oracle database virtual machines.

Database server utilization varies widely by the application:

- If the application is commercial, follow published guidelines by vendor.
- If the application is custom-written, work with the application developers to determine the resource requirements.

If the exact workload is unknown, start with fewer virtual CPUs and increase the number later if necessary. Allocate multiple vCPUs to a virtual machine only if the anticipated database workload can take advantage of all the vCPUs.

When consolidating multiple virtual machines on a single ESXi host, proper hardware sizing is critical for optimal performance. Confirm that cumulative physical CPU resources on a host are adequate to meet the needs of the virtual machines by testing your workload in the planned virtualized environment. CPU over-commitment should be based upon actual performance data to avoid adversely affecting virtual machine performance.

VMware recommends keeping Oracle Non-Uniform Memory Architecture (NUMA) enabled in server hardware BIOS and at the guest-operating system level. Although Oracle NUMA support is disabled by default for 11g and previous versions, in some circumstances, enabling Oracle NUMA support may improve performance and the Oracle doc (Doc ID: 864633.1) suggests testing it before deciding to use it with the production system.

Per the [Oracle Databases on VMware Best Practices Guide](https://v2xknowledgecenter.vmware.com/article/2011861), it is recommended to size the virtual machine to fit within a NUMA node as far as possible. However, there may be cases where an Oracle VM may cross NUMA boundaries such as data warehousing VM, which would be deemed as a wide VM. In that case, make sure that the vCPU number is a multiple of the number of physical cores per socket.

### 7.4 Memory Configuration Guidelines

This section provides guidelines for memory allocation for Oracle database virtual machines:

- In the production environment, Oracle databases can be memory intensive, so set the memory reservation equal to the aggregate size of SGA, program global area (PGA), and operating system background process. The memory reservation should be large enough to avoid kernel swapping between ESXi and the guest operating system.
- In development, test or QA memory over-commitment can be considered to take advantage of VMware memory reclamation features and techniques. However, the type and number of databases that can be deployed using over-commitment largely depend on their actual workload.
- Do not disable the balloon driver.
- The guest operating system within the virtual machine still needs its own separate swap or page file. Follow the same swap space guidelines given for physical environments.
- Configure HugePages in the Linux guest operating system to improve the performance of Oracle databases on vSphere. The Automatic Memory Management (AMM) and HugePages are not compatible, so disable AMM on 11gR2 and enable Automatic Shared Memory Management (ASMM). See Oracle metalink notes for more details on configuring HugePages on Linux:
  - Note 361323.1—HugePages on Linux

Copyright © 2017 VMware, Inc. All rights reserved.
• Note 361468.1—HugePages on 64-bit Linux
• Note 401749.1—Shell Script to Calculate Values Recommended HugePages/Huge TLB Configuration
• Set the USE_LARGE_PAGES initialization parameter to "only" so the database fails to start if it is not backed by HugePages.
• Setting memory reservations might limit vSphere vMotion. A virtual machine can only be migrated if the target ESXi host has free physical memory equal to or greater than the size of the reservation.

7.5 Network Configuration Guidelines

This section covers design guidelines for the virtual networking environment. The standard VMware networking best practices apply to running Oracle databases on vSAN.

• Separate VLANs are recommended per traffic type. Use separate infrastructure traffic from VM traffic for security and isolation.
• Use the VMXNET3 network adapter. The VMXNET3 adapter is optimized for virtual environments and is designed to provide high performance.
• For Oracle RAC interconnect:
  • Use jumbo frames. To enable jumbo frames, refer to the VMware Knowledge Base Article 1003712 for enablement.
  • Disable interrupt coalescing to reduce latency on vNICs (ethernetX.coalescingScheme=disabled).
• Take advantage of Network I/O Control to converge network and storage traffic (vSAN) onto 10GbE. Network I/O Control in vSphere enables you to guarantee service levels (bandwidth) for particular vSphere traffic types: VM traffic, RAC Interconnect, vSAN, management, backup, and vSphere vMotion.
• Use NIC teaming for availability and load balancing. NIC teaming occurs when multiple uplink adapters are associated with a single vSwitch to form a team.

Refer to the VMware vSAN Network Design Guide for more information.
This section summarizes on the validation of vSAN as a storage platform supporting a scalable, resilient, highly available, and high performing Oracle RAC Cluster.
8.1 Conclusion

vSAN is a cost-effective and high-performance storage platform that is rapidly deployed, easy to manage, and fully integrated into the industry-leading VMware vSphere platform.

This solution validates vSAN as a storage platform supporting a scalable, resilient, highly available, and high performing Oracle RAC Cluster.

vSAN’s integration with vSphere enables ease of deployment and management from the single-management interface. vSAN Stretched Cluster provides an excellent storage platform for Oracle Extended RAC Cluster solution providing zero RPO and RTO at the metro distance. It also demonstrates how vSAN Stretched Cluster along with Oracle Data Guard and RMAN backup provides disaster recovery and business continuity required for mission-critical application at the same time efficiently using the available resources.
9. Reference

This section lists the relevant references used for this document.
9.1 White Paper

For additional information, see the following white papers:

- VMware vSAN Stretched Cluster Performance and Best Practices
- VMware vSAN Stretched Cluster Bandwidth Sizing Guidance
- Oracle Databases on VMware Best Practices Guide
- Oracle Databases on VMware RAC Deployment Guide
- Oracle RAC and Oracle RAC One Node on Extended Distance (Stretched) Clusters

9.2 Product Documentation

For additional information, see the following product documentation:

- vSAN 6.1 Stretched Cluster Guide
- Oracle Real Application Clusters Administration and Deployment Guide
- Performance Best Practices for VMware vSphere 6.0

9.3 Other Documentation

For additional information, see the following document:

- Swingbench Installation and Usage
10. About the Author and Contributors

This section provides a brief background on the author and contributors of this document.
10.1 About the Author and Contributors

Palanivenkatesan Murugan, Solution Architect in the Storage and Availability, Product Enablement team wrote the original version of this paper. Catherine Xu, Technical Writer in the Product Enablement team, edited this paper to ensure that the contents conform to the VMware writing style. Sudhir Balasubramanian, Senior Solution Architect in the Global Technical and Professional Services team, also contributed to this document.