Oracle Real Application Clusters in Oracle VM Environments

An Oracle Technical White Paper
March 2012

ORACLE

Oracle Real Application Clusters in Oracle VM Environments

Oracle Real Application Clusters in Oracle VM Environments

EXECUTIVE SUMMARY

Running today's non-critical business applications in virtualized environments has shown to be efficient and cost saving. More sophisticated or highly available applications on the other hand were most likely incompatible with commonly used software based virtualization solutions.

The availability of Oracle VM overcomes this obstacle. Providing software based virtualization infrastructure (Oracle VM) and the market leading high availability solution Oracle Real Application Clusters (RAC), Oracle now offers a highly available, grid-ready virtualization solution for your data center, combining all the benefits of a fully virtualized environment.

Oracle VM *and* Oracle Real Application Clusters (RAC) enable the benefits of a virtualized data center infrastructure for highly available applications.

The combination of Oracle VM and Oracle RAC enables a better server consolidation (RAC databases with underutilized CPU resources or peaky CPU utilization can often benefit from consolidation with other workloads using server virtualization) sub-capacity licensing, and rapid provisioning.

RAC on OVM also supports the creation of non-production virtual clusters on a single physical server for product demos, educational settings and test environments. This deployment combination permits dynamic changes to pre-configured database resources for agile responses to changing service level requirements common in consolidated environments.

Oracle VM is the only software based virtualization solution that is fully supported and certified for Oracle Real Application Clusters.

This paper discusses various Oracle RAC deployment scenarios and provides best practices for an optimized Oracle RAC deployment in Oracle VM environments for Oracle VM versions 2.1.2 through Oracle VM 3.0.3. Version specific information is noted accordingly.

Note: Oracle constantly tests additional, and advanced, Oracle VM features with Oracle RAC. This paper is updated regularly as new test results are available.

INTRODUCTION

Oracle Real Application Clusters (RAC)[1] is an option to the award-winning Oracle Database Enterprise Edition. Oracle RAC is a cluster database with a shared cache architecture that overcomes the limitations of traditional shared-nothing and shared-disk approaches to provide highly scalable and available database solutions for all your business applications. Oracle RAC is a key component of Oracle's enterprise grid architecture and cloud foundation.

Oracle RAC utilizes **Oracle Clusterware**[2] for cluster node and group membership, high availability, and application resource management. Oracle Clusterware is the technology that transforms a server farm into a cluster. A cluster in general is a group of independent servers that cooperate as a single system. Oracle Clusterware is the intelligence in this system that ensures the required cooperation and is a key component of Oracle's enterprise grid architecture as well as cloud foundation.

In a typical Oracle RAC installation, **Oracle Automatic Storage Management (ASM)**[3] acts as the underlying, clustered volume manager. It provides the database administrator with a simple storage management interface that is consistent across all server and storage platforms. As a vertically integrated file system and volume manager, purpose-built for Oracle database files, ASM provides the performance of raw I/O with the easy management of a file system. Oracle Automatic Storage Management provides the basis for a shared storage pool in Oracle's enterprise grid architecture and cloud foundation.

Oracle VM[4] is a platform that provides a fully equipped environment to better leverage the benefits of virtualization technology. Oracle VM enables deployment of operating systems and application software within a supported virtualization environment. Oracle VM completes the Oracle's enterprise grid offering by providing an Oracle RAC certified virtualization environment.

---

[1] Oracle Real Application Clusters (RAC) homepage: http://otn.oracle.com/rac
[2] For more information on Oracle Clusterware, visit http://otn.oracle.com/clusterware
[3] For more information on Oracle Automatic Storage Management (ASM), visit http://otn.oracle.com/asm
[4] Oracle VM homepage: http://www.oracle.com/virtualization

WHAT IS ORACLE REAL APPLICATION CLUSTERS (RAC)?

A RAC Database is a clustered database. A cluster is a group of independent servers that cooperate as a single system. Clusters provide improved fault resilience and modular incremental system growth over single symmetric multi-processor (SMP) systems. In the event of a system failure, clustering ensures the highest availability to users and access to mission critical data is not lost. Redundant hardware components, such as additional nodes, interconnects, and disks, allow the cluster to provide high availability. Such redundant hardware architectures avoid single points-of-failure and provide exceptional fault resilience.

Oracle RAC enables the Oracle Database to run mainstream business applications of all kinds on clusters, including popular packaged products (such as Oracle Ebusiness Suite, Peoplesoft, Siebel, SAP) and in-house developed applications generating OLTP, DSS, or mixed workload.

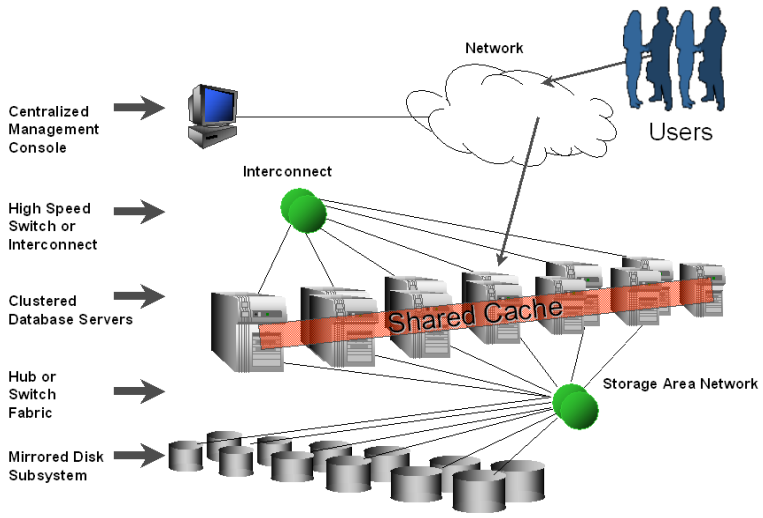

Figure 1: Oracle Real Application Clusters Overview

With Oracle Real Application Clusters (as with any other Oracle database) the Oracle Instance (the processes and memory structures allocated on a server to allow access to the data) is de-coupled from the Oracle Database (the physical structures residing on the storage, which actually hold the data. These structures are commonly known as 'datafiles').

However, a clustered database (using more than one instance) differs from a single instance database in a way that the database can be accessed by multiple instances concurrently. Each instance runs on a separate server in the cluster (formed by Oracle Clusterware).

When additional resources are required, additional nodes and instances can easily be added to the cluster with no downtime. Once a new instance has been started, applications using services can immediately take advantage of it with no changes to the application or application server.

Oracle Real Application Clusters is an extension to the Oracle Database and therefore benefits from the manageability, reliability, and security features built into the Oracle Database.

WHAT IS ORACLE RAC ONE NODE?

Oracle Real Application Clusters (RAC) One Node[5] is a new option to the Oracle Database 11g Release 2 Enterprise Edition. It improves upon many of the benefits of server virtualization and extends them to databases running in physical server environments. Oracle RAC One Node enables:

- Better server consolidation
- Enhanced protection from failures
- Greater flexibility and workload management
- Better online maintenance

In addition, it allows customers to virtualize database storage, standardize their database environment, and, should the need arise, upgrade to a full multi-node Oracle RAC database without downtime or disruption. Further, it is fully compatible with and complementary to Oracle Virtual Machine (VM), allowing customers to build environments that leverage the strengths of both Oracle RAC One Node and server virtualization.

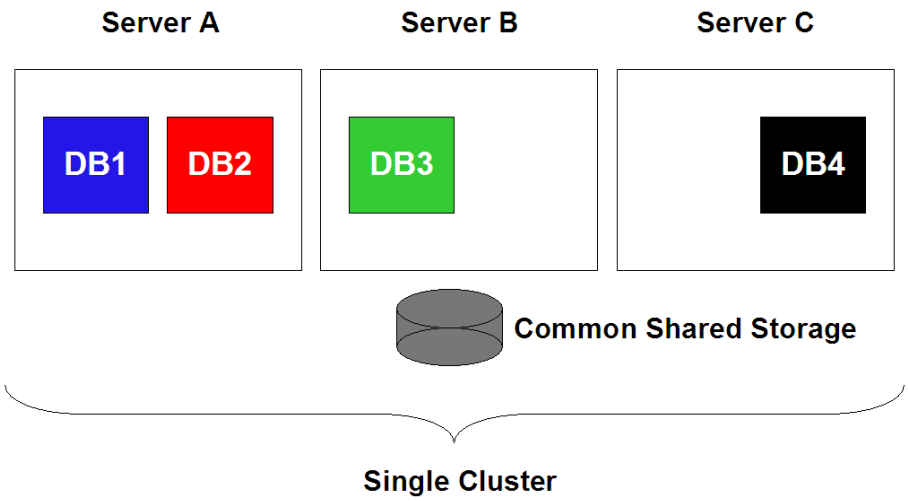

Figure 2: Oracle RAC One Node - typical architecture overview

In the configuration shown in figure 2, five single-instance Oracle RAC One Node databases are running in a cluster of three servers. Server A is hosting Oracle RAC One Node databases DB1 and DB2, server B is hosting database DB3 and server C is hosting databases DB4 and DB5. Each server runs one OS. In servers A and C above, multiple databases are consolidated onto a single OS.

[5] For more information on Oracle RAC One Node see http://otn.oracle.com/rac

WHAT IS ORACLE VM?

Oracle VM[6] is a platform that provides a fully equipped environment for better leveraging the benefits of virtualization technology. Oracle VM enables you to deploy operating systems and application software within a supported virtualization environment. The components of Oracle VM are:

- Oracle VM Manager: Provides the user interface, which is a standard ADF (Application Development Framework) web application, to manage Oracle VM Servers[7]. Manages virtual machine lifecycle, including creating virtual machines from installation media or from a virtual machine template, deleting, powering off, uploading, deployment and live migration of virtual machines. Manages resources, including ISO files, virtual machine templates and sharable hard disks.
- Oracle VM Server: A self-contained virtualization environment designed to provide a lightweight, secure, server-based platform for running virtual machines. Oracle VM Server is based upon an updated version of the underlying Xen hypervisor technology, and includes Oracle VM Agent.
- Oracle VM Agent: Installed with Oracle VM Server. It communicates with Oracle VM Manager for management of virtual machines.

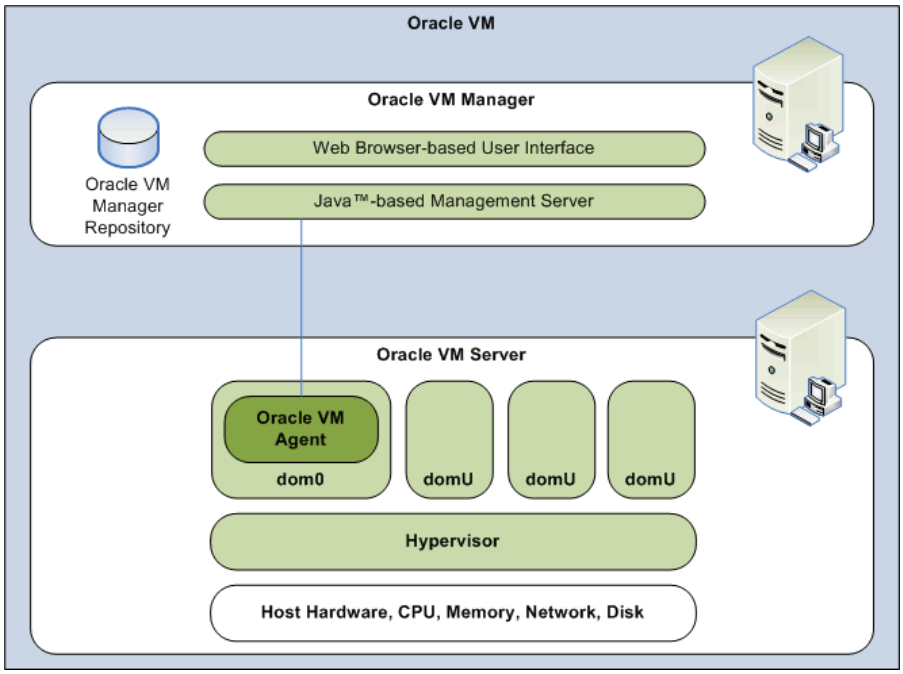

Figure 3: Oracle VM overview

[6] Oracle VM homepage: http://www.oracle.com/virtualization
[7] Oracle VM is also fully integrated into Oracle Enterprise Manager 10.2.0.5: www.oracle.com/technology/software/products/oem/index.html

WHY RUN ORACLE RAC ON ORACLE VM?

There are several reasons why customers may want to run Oracle RAC in an Oracle VM environment. Some of the more common ones are:

1. **Server Consolidation:** Oracle RAC databases or Oracle RAC One Node databases with underutilized CPU resources or variable CPU utilization can often benefit from consolidation with other workloads using server virtualization. A typical use case for this scenario would be the consolidation of several Oracle databases (Oracle RAC, Oracle RAC One Node or Oracle single instance databases) into a single Oracle RAC database or multiple Oracle RAC databases where the hosting Oracle VM guests have pre-defined resource limits configured for each VM guest.
2. **Sub-capacity licensing:** The current Oracle licensing model requires the Oracle RAC database to be licensed for all CPUs on each server in the cluster. Sometimes customers wish to use only a subset of the CPUs on the server for a particular Oracle RAC database. Oracle VM can be configured in such way that it is recognized as a hard partition. Hard partitions allow customers to only license those CPUs used by the partition instead of licensing all CPUs on the physical server. More information on sub-capacity licensing **using hard partitioning** can be found in the Oracle partitioning paper[8]. For more information on using hard partitioning with Oracle VM refer to the "Hard Partitioning with Oracle VM" white paper[9].
3. **Create a virtual cluster:** Oracle VM enables the creation of a virtual cluster on a single physical server. This use case is particularly interesting for product demos, educational settings, and test environments. This configuration should never be used to run production Oracle RAC environments. The following are valid deployments for this use case:
    a. Test / development cluster
    b. Demo cluster
    c. Education cluster
4. **Rapid Provisioning:** The provisioning time of a new application consists of the server (physical or virtual) deployment time, and the software install and configuration time. Oracle VM can help reduce the deployment time for both of these components. Oracle VM supports the ability to create deployment templates. These templates can then be used to rapidly provision new (Oracle RAC) systems.

---

[8] Oracle Licensing – Partitioning: www.oracle.com/corporate/pricing/partitioning.pdf
[9] Hard Partitioning with Oracle VM - www.oracle.com/technology/tech/virtualization/pdf/ovm-hardpart.pdf

BUSINESS CONTINUITY, HIGH AVAILABILITY AND SCALABILITY

Business continuity is a key element in today's business and while Oracle RAC is still the ultimate solution when it comes to Oracle Databases, there are more options to choose from when considering virtualization technologies.

Two different kinds of High Availability (HA) mechanisms can be distinguished in an Oracle virtual environment: Oracle VM HA (external HA) and Oracle Clusterware based, internal HA (used for Oracle RAC) as illustrated in figure 3.

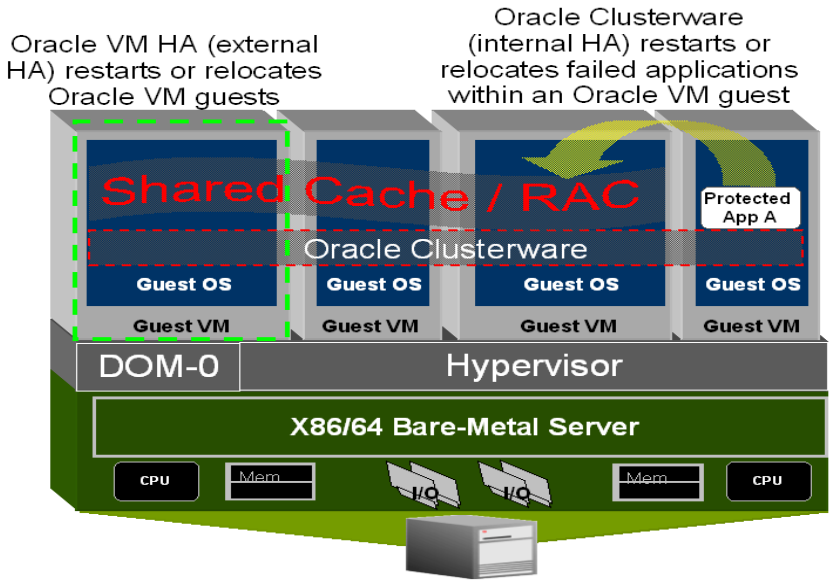

Figure 4: Oracle VM HA and Oracle RAC

In general, Oracle VM HA enables the restart of an Oracle VM guest on either the same physical machine or a different machine, if more than one physical machine is available in the server pool. However, in any case, Oracle VM would operate on the VM guest as a whole, restarting whatever is running within it.

The reason is that the Oracle VM guest is usually unaware of the applications running in the virtualized environment as much as the application is typically unaware of the Oracle VM guest it is running in.

For a fast, finer grained recovery from a process or application failure within an Oracle VM guest, an internal, cluster based HA solution like Oracle Clusterware (used for Oracle RAC) should be used. Utilizing application specific agents, these solutions will perform corrective actions particular to the failure without the overhead of restarting the whole Oracle VM guest

While a combination of both an internal application HA and the external VM HA serves different HA functions, it needs to be noted that the Oracle VM HA and Oracle VM policy-based resource management, Distributed Resource Scheduling, (DRS) must not interfere with inherent Oracle RAC instance placement rules.

Specifically, for mission-critical, production deployments it is unsupported to co-locate Oracle VM guests hosting instances of the same Oracle RAC database on a single OVS physical server as a result of Oracle VM guest failover or automated DRS placement policy. Any Oracle VM guest failover or DRS placement policies must respect this fundamental RAC instance placement rules.

Using Oracle VM HA for Oracle VM guests hosting an Oracle RAC database is therefore not supported in OVM 2.2 and earlier versions. Starting with Oracle VM 3.0.3 the following HA combinations are possible:

1. Oracle Real Application Clusters (RAC) High Availability and Scalability
    a. Without additional Oracle VM guest HA
    b. In conjunction with Oracle VM guest HA,
    where Oracle RAC instance placement policies are respected.
2. Oracle Clusterware (failover cluster) provided High Availability
    a. Without additional Oracle VM guest HA
    b. In conjunction with Oracle VM guest HA,
    where Oracle RAC instance placement policies are respected.
3. Standalone Oracle VM guest High Availability
    for Oracle single instance databases for example.

Lately, some vendors of software based virtualization have announced that they will provide better business continuity with "continuous availability" or "fault tolerant" solutions that would be solely based on their virtualization software.

Those solutions might be suitable for lightweight, stateless applications, however, they are not suitable for mission critical production databases. This means that for production environments the choice is limited to the solutions listed above.

For Oracle RAC production environments, solution 1a (Oracle RAC HA and scalability without additional Oracle VM guest HA) **must be used**, which still provides the full Oracle RAC HA and scalability benefits. Please see Live Migration support noted later in this paper.

However, for test systems and small development systems, i.e. non-production environments, solution 1b (Oracle RAC HA and scalability in conjunction with Oracle VM guest HA) can be used.

On the other end of the scale, solution 3 (Standalone Oracle VM guest High Availability) **is fully supported by Oracle**. However, it is limited in its capability to protect the actual application within the virtual environment. It will nevertheless provide optimal protection against physical hardware failures.

Solution 2 Oracle Clusterware [failover cluster] together with Oracle RAC One Node might be an alternative when it comes to single instance databases or other applications that need to be protected against hardware and process failures.

For Oracle Clusterware based solutions regardless of Oracle VM, the current support status can be found in **Oracle Metalink Note 790189.1 - Oracle Clusterware and Application Failover Management.** More information can be found on http://www.oracle.com/goto/clusterware.

LIVE MIGRATION AND ONLINE DATABASE RELOCATION SUPPORT

Additional solutions that improve business continuity and availability for Oracle RAC in an Oracle Virtual Machine environment include Oracle VM *Live Migration* and Oracle RAC One Node *Online Database Relocation*

Oracle VM *Live Migration* moves a virtual machine from one physical node to another, within the same pool of servers. Oracle RAC One Node *Online Database Relocation* moves an Oracle database instance from one server to another within the same cluster. In Oracle VM environments, these servers are virtual machines, which host an Oracle Clusterware based cluster.

As of now, Oracle does not support mixed cluster environments, in which some cluster members are based on virtual machines while others are based on physical servers. Oracle RAC One Node Online Database Relocation can therefore not be used to migrate between physical and virtual environments.

As pointed out above, Oracle VM HA cannot be applied to Oracle VM guests that host an Oracle RAC cluster due to possible violations of Oracle RAC instance placement rules and CPU allocation rules when automatically re-placing an Oracle VM instance after a failure.

However, Oracle VM *Live Migration* offers a controlled mechanism for manually moving Oracle VM guests between physical nodes in the server pool. Since Oracle VM *Live Migration* allows for the relocation of an Oracle VM guest while respecting the placement and CPU allocation rules for Oracle RAC, it can be applied to Oracle VM guests that host an active Oracle RAC instance or instances.

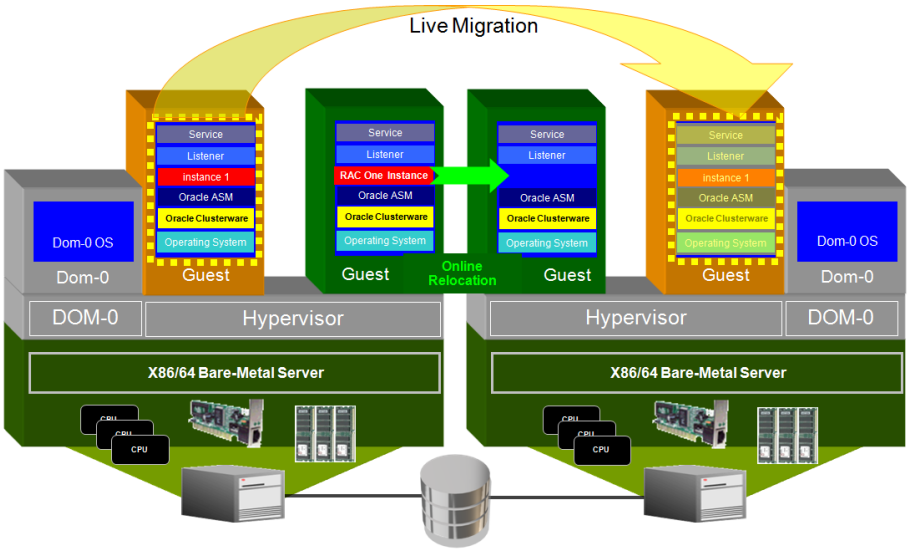

Figure 5: Live Migration and RAC One Node Database Relocation illustrated

When to use *Live Migration* or Oracle RAC One Node *Relocation*?

Oracle VM Live Migration and Oracle RAC One Node Online Database Relocation, together, offer complimentary solutions which support continuous, non-interrupted High Availability for the Oracle database.

Live Migration of Oracle VMs offers operating system (OS) level migration to facilitate workload management, fault management, and system maintenance operations, for example, Live Migration permits server level maintenance while not losing access to the Oracle RAC database VM.

Oracle Live Migration basically consists of iteratively pre-copying the contents of a source VM's memory from one physical host to another target VM on a remote host with minimal interruption to service.

Oracle RAC One Node Online Database Relocation offers instance level migration between properly configured VMs to facilitate maintenance and workload management within VMs. Using Oracle RAC One Node Online Database Relocation, the Oracle RAC One Node instance can be relocated to another server, if the current server is resource constrained or requires maintenance operations such as OS patches for example.

Oracle RAC One Node Online Database Relocation will also migrate Oracle RAC One Node databases to target VMs to facilitate database maintenance on the source VM without interruption of service to other database instances that may be running in the source or target VM. A typical example for such an operation would be a rolling patch upgrade for a database.

DEPLOYMENT METHODS

Historically and typically, virtualized environments were **ideal for development or test environments**. Easy to re-install by re-deploying a formerly saved base image or template, once the current environment is 'worn out' by various test runs.

Oracle VM is the ideal infrastructure for Oracle RAC used in virtual development, test, demo, or education clusters, based on only one Oracle VM host.

Oracle VM provides those benefits as part of the virtual cluster creation and the rapid provisioning approach, even when used together with Oracle RAC, as described in the previous section.

Figure 5 shows a typical Oracle RAC deployment for a development or test environment based on Oracle VM. It should be noted that **for virtual test or development clusters** Oracle allows having the 2 Oracle VM Guest Domains, each hosting 1 Oracle RAC database instance, running on only 1 Oracle VM host.

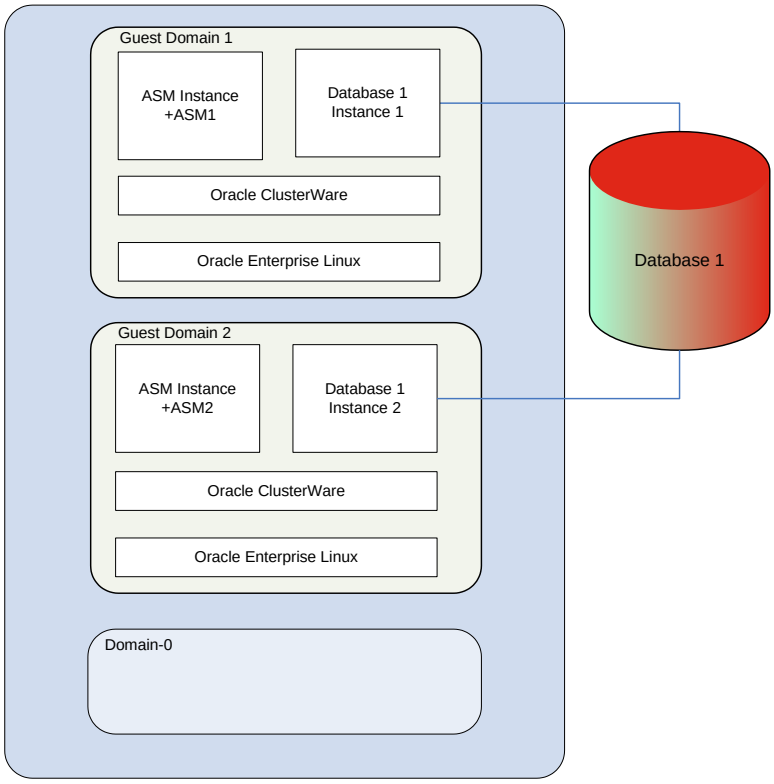

Figure 6: Oracle RAC on Oracle VM: Development Deployment example

Based on rigorous certification tests, Oracle has **certified Oracle RAC on Oracle VM** for **production environments from Oracle VM 2.1.2 through Oracle VM 3.0.3**. This is the only certified software based virtualization technology for Oracle Real Application Clusters and thereby enables these technologies to be used beyond the scope of pure development or test environments.

Oracle VM together with Oracle RAC enables fully virtualized environments as the bases for server consolidation and highly available applications. In order to avoid a single point of failure, a minimum of two Oracle VM hosts must be used for production systems.

Unlike test and development environments, typical production environments would use more than one Oracle VM host for each Oracle VM Guest Domain (minimum 2), each, again, hosting an Oracle RAC database instance, as illustrated in figure 6. These kinds of configurations are as well most likely used in server consolidation environments.

Having more than one Oracle VM host as the underlying hardware platform to host the Oracle VM Guest domains with the Oracle RAC database instances eliminates the host hardware to be the single point of failure.

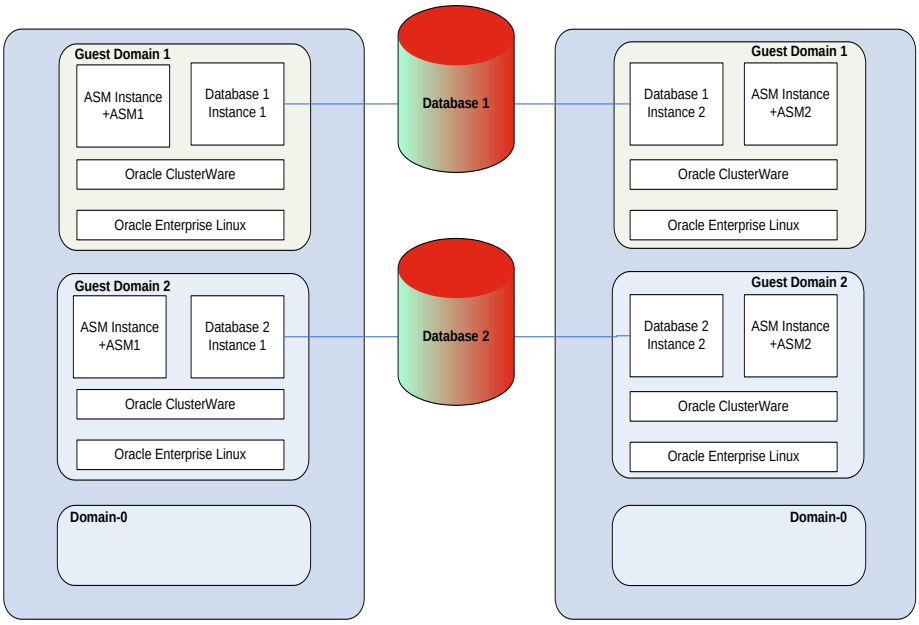

Figure 7: Oracle RAC on Oracle VM: Production Deployment example

HARDWARE AND SOFTWARE REQUIREMENTS

Oracle has certified Oracle RAC in Oracle VM environments based on the following hardware and software requirements. Configurations that do not fulfill these requirements are currently not supported.

Note: As a minimum, the hardware and software requirements listed in the Oracle RAC documentation must be met to deploy Oracle RAC on Oracle VM in production environments.[10]

Note: Oracle is constantly in the process of testing additional and advanced Oracle VM features with Oracle RAC. This paper will be updated regularly as new test results are available.

Hardware requirements for Production Environments:

- Minimum of 2 Oracle VM hosts are strongly recommended
    - Each host must provide a minimum of 2 Ethernet NICs for Oracle VM management and the Oracle RAC public and private (interconnect) communication. A minimum of 4 Ethernet NICs are required, if network based storage connectivity is used.
    - 10GbE network interfaces are supported. It is recommended that these NICs support IEEE Data Center Bridging standards.
    - For the interconnect 1 Gbit Ethernet or 10GbE is required
- Network redundancy requirements will double the number or NICs given above accordingly. (4-6 NICs strongly recommended.)
- Storage redundancy requirements will require a minimum of two HBAs or SCSI controllers.

Hardware requirements for Development Environments:

- For Development or non-production environments on a single Oracle VM host, external storage is recommended, but not required.
- Virtual network connectivity can be established within the Oracle VM host, if no external client connects are required.
- Oracle recommends using identical hardware configurations for development and pre-production Q&A environments accurately modeling production environments.

Oracle RAC Software

- Oracle RAC and Oracle Clusterware 10.2.0.4, 11g Rel. 1, 11g Rel. 2 in 32-bit and 64-bit software versions are currently supported.

[10] http://www.oracle.com/technology/documentation/index.html

General Oracle VM

Oracle VM release: 2.1.2 through Oracle VM release 3.0.3 are supported. Older releases are not supported with Oracle RAC or Oracle Clusterware.

- When using Oracle RAC:
    - Dynamic vCPU and memory changes are not supported.
    - Live Migration of an Oracle RAC VM is not supported with Oracle VM releases earlier than 2.2.1
    - Oracle RAC One Node (Online Database Relocation) is supported with Oracle VM 2.1.5 and later and Oracle Database 11g Rel. 2 (11.2.0.2) and later.

Oracle VM Guest Configuration

- The guest must be a para-virtualized guest
- Oracle Linux 5.1 or higher (OL 5.1 or higher)
- 32-bit and 64-bit Linux is currently supported for the Oracle VM Guest
- Oracle VM Server release 3.0 requires 64-bit x86 hardware

Oracle VM vCPU configuration for Oracle RAC

Testing has shown that a small amount of incidental over-committing of physical CPUs will not greatly diminish the overall stability of the system or the cluster stack where the competing workloads for oversubscribed resources are well understood and concurrent demand does not exceed physical capacity.

It is still recommended, however, **not to over-commit CPUs**. Multi-core CPUs will treat every core as a virtual CPU (vCPU) and multi-core CPUs with multi-thread support **will treat each thread** as a vCPU. **vCPU allocation** in Oracle RAC / VM environments **must therefore adhere to the following rules**:

- Maintain Oracle VM's default vCPU allocation for dom-0: Oracle VM will allocate 1 vCPU for each real CPU or core to dom-0.
- The total amount of vCPUs allocated to guest domains (running Oracle RAC guests), should not exceed two times (2x) the amount of real CPUs / cores in the Oracle VM server.
- The amount of vCPUs allocated to a single guest domain should not exceed the amount of real CPUs / cores in the Oracle VM server.
- CPU pinning is only recommended for hard partitioning. If no hard partitioning is required, CPU pinning should not be used.

Note: vCPUs are based on cores, not threads. Single core CPUs are counted as 1 CPU.

vCPU allocation examples

Example 1: A server with 8 cores – 2 CPUs, quad core.

Valid configuration (CPUs are over-committed in accordance to the rules)

- Dom-0 has 8 vCPUs allocated (default allocation)
- Guest domain 1 (running RAC) has 8 vCPUs allocated
- Guest domain 2 (running RAC) has 4 vCPUs allocated
- Guest domain 3 (running RAC) has 2 vCPUs allocated
- Guest domain 4 (running RAC) has 2 vCPUs allocated

Invalid configuration

- Dom-0 has 4 vCPUs allocated (non-default allocation)
    - Violation 1: non-default dom-0 vCPU allocation
- Guest domain 1 (running RAC) has 10 vCPUs allocated
    - Violation 2: allocated vCPUs (10) > real CPUs (8)
- Guest domain 2 (running RAC) has 6 vCPUs allocated
- Guest domain 3 (running RAC) has 8 vCPUs allocated
    - Violation 3: Total allocated vCPUs (24) > 2x real CPUs (16)

Example 2: A server with 24 cores – 4 CPUs, 6-core

Valid configuration (CPUs are not over committed)

- Dom-0 has 24 vCPUs allocated (default allocation)
- Guest domain 1 (running RAC) has 12 vCPUs allocated
- Guest domain 2 (running RAC) has 2 vCPUs allocated
- Guest domain 3 (running RAC) has 4 vCPUs allocated
- Guest domain 4 (running RAC) has 6 vCPUs allocated

Invalid configuration

- Dom-0 has 32 vCPUs allocated (non-default allocation)
    - Violation 1: non-default dom-0 vCPU allocation
- Guest domain 1 (running RAC) has 28 vCPUs allocated
    - Violation 2: allocated vCPUs (28) > real CPUs (24)
- Guest domain 2 (not running RAC) has 8 vCPUs allocated
- Guest domain 3 (running RAC) has 12 vCPUs allocated

Oracle VM supported storage configurations for Oracle RAC

Oracle VM itself allows configuring storage in many different ways. Oracle RAC and Oracle Clusterware are not supported on all possible storage configurations that Oracle VM offers.

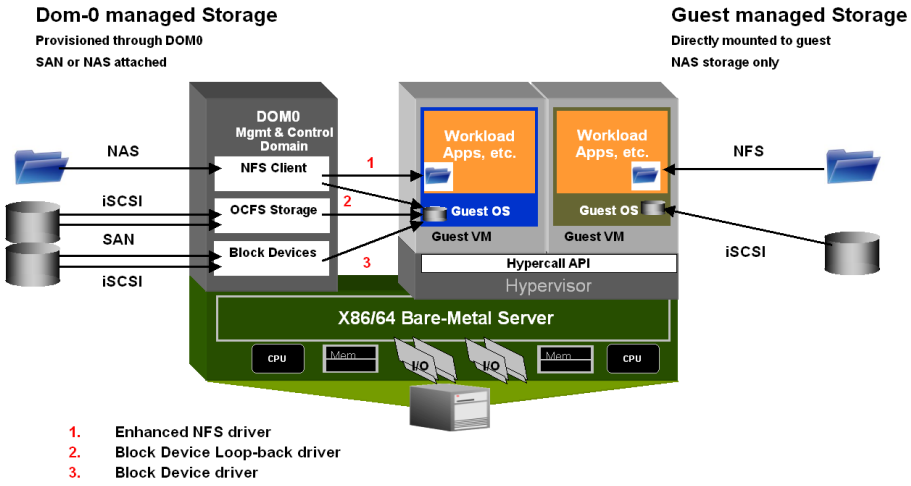

Figure 8: Storage configuration options in Oracle VM

Generally, storage configurations can be divided into two groups:

Dom-0 managed storage

In this method of storage configuration, the storage is made available in dom-0 and then made available in the guest domains using the guest's configuration file.

Guest managed storage

In this method of storage configuration, the storage is made available directly in the guest domain. It is not visible in dom-0.

Storage configuration method	Supported with Oracle RAC
Dom-0 managed storage	
• SAN / block device	Yes
• iSCSI / block device	Yes
• SAN / OCFS	No
• iSCSI / OCFS	No
• NFS	No
Guest managed storage	
• iSCSI	Yes
• NFS	Yes

Table 1: Storage options supported with Oracle RAC

ORACLE RAC ON ORACLE VM – BEST PRACTICES

Based on extensive tests in the course of the Oracle RAC for Oracle VM certification, Oracle has developed some best practices recommendations in order to run Oracle RAC in an optimized virtual environment. Below you will find some of the recommendations as a result of those tests.

Note: Oracle continually tests additional and advanced Oracle VM features with Oracle RAC. This paper is updated regularly as new test results are available.

Oracle Installation Recommendations

Setting diagwait in Oracle Clusterware

When running Oracle RAC in highly stressed clusters (CPU and memory bound), Oracle recommends setting the diagwait to 13 seconds. This prevents false evictions and enables a better diagnosis in case of false evictions.

When running Oracle RAC 10.2.0.4 or 11g Rel. 1 in Oracle VM environments, it is strongly recommended to set diagwait to 13 seconds for all installations, regardless of utilization. Before changing the diagwait value, please, review Metalink Note 559365.1 (Using Diagwait as a diagnostic to get more information for diagnosing Oracle Clusterware Node evictions).

Note that with 11g Rel. 2 the diagwait setting has been internalized. Setting the diagwait value is not required anymore.

Generic Configuration Recommendations

Time synchronization

By default, the Oracle VM Guest time is automatically synchronized with Domain-0. This time synchronization can allow some time drift in the guest domains running Oracle RAC. Therefore, the following recommendations should be implemented when running Oracle RAC in Oracle VM:

- Configure `xen.independent_wallclock=1`
  in /etc/sysctl.conf in the guest domains only.
  This allows the para-virtualized guests to manage their own system clocks.
- Configure NTPD in dom-0
- Configure NTPD with the –x flag in all the guest domains per Metalink Note 1056693.1. See also Metalink Note 580296.1 (Clock drift issue between Dom-0 and Dom-U on OVM Server) for more context.

Note that **Oracle Grid Infrastructure 11g Rel. 2** provides the Oracle Cluster Time Synchronization Service (**CTSS**) daemon, which can be used to synchronize the time between cluster members if NTPD is not used in the VM Guests to synchronize the time against an external time server.

Storage Configuration

Multipathing and Device Persistency in Domain-0

For Oracle RAC in Oracle VM environments, multipathed access to the (SAN) storage is highly recommended. Multipathing should be configured in Domain-0 and not in the guest domains. All supported multipathing tools can be used, e.g. Device Mapper, QLogic multipathing, etc.

Device persistency, if not already configured as part of the multipathing solution mentioned above, should also be set up in Domain-0. If the multipathing software doesn't offer device persistency, the Linux inherent udev configuration should be used. There is no need to set up device persistency in the Guest VMs. The mapping of Oracle VM disks to guest disks is done statically in the guest configuration file (see Appendix A for an example).

Storage Configuration for the Oracle VM Guests hosting the Oracle RAC database

The only configuration that is required to be set up in the Oracle RAC guest domains is the setup of the permissions on the block devices. In addition, when running pre-Oracle RAC 11g Release 2, it may be required to configure RAW devices. It is strongly recommended to use the Linux native udev tool to create the RAW devices with the right permissions on the disks in the Oracle VM guest.

Oracle recommends using Oracle Automatic Storage Management (ASM) to manage these RAW devices and as the underlying volume manager for the Oracle RAC database files. Optionally, ASMLib can be used in the guest domains in order to further simplify OS disk administration.

Note, as of 11gR2 raw devices are no longer supported for new RAC installs.

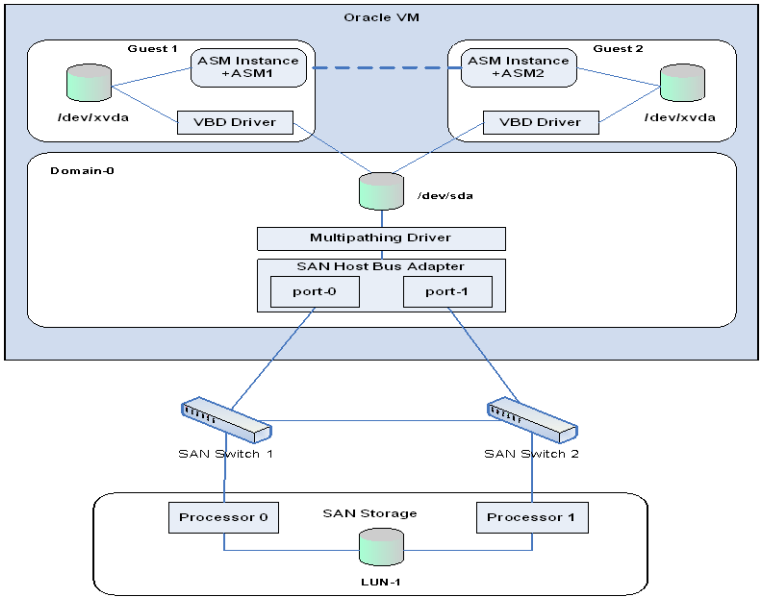

Figure 9: SAN Storage on Oracle VM

Cluster file system requirements

Some applications require running the Oracle RAC database on a cluster file system or relying on some files that need to be available on every node. For these applications the Oracle ASM Cluster File System (ACFS – 11g Rel. 2 or later) should be used in the guest domains.

Network Configuration

Required Networks

The Oracle VM host should have a sufficient number of NICs to provide at least two distinct bridges and virtual NICs to the guest domains. For production systems the number of NICs per Oracle VM host is specified in the "Hardware and Software Requirements" section of this paper. The actual number of NICs depends on the overall configuration and level of redundancy desired.

As for any other Oracle RAC configuration, a separation of networks for the

1. Public network
2. Private network (cluster interconnect)
3. Storage network (if applicable)

is required for production environments. Therefore, the Oracle VM host must ensure that the respective communication would not interfere with each other.

These networks can be shared with multiple Oracle RAC VM Guests on an Oracle VM host. However, network bandwidth and capacity must be considered in this case (see "Sizing the Oracle VM environment for Oracle RAC" for recommendations).

Network Bonding

For production environments it is strongly recommended to have two Network Interface Controllers (NICs) for each network and use Linux bonding configured in Dom0 to make them highly available. This results in an Oracle RAC network HA requirement of a minimum of four NICs (for public and private network) per Oracle VM host running Oracle RAC, as illustrated in figure 8. Appendix B explains setting up bonding devices step-by-step.

For the Oracle Clusterware private interconnect, in place of Linux bonding in Dom0, you may present two to four vNICs for the private interconnect to each of the configured Oracle VM guests. Each vNIC would map to a distinct physical network interface, either on a single subnet or multiple subnets for Oracle clusterware interconnect high availability.

This feature, Redundant Interconnect Usage[11] or HAIP is a layer 3 NIC aggregation model managed by the Oracle Clusterware in each of the Oracle VM guests and uses automatically assigned, non-routed link local addresses for the Oracle RAC interconnect communication. An acceptable configuration would be a combination of Linux bonding of the clusterware interconnects in Dom0 with HAIP configured in the Oracle VM guests.

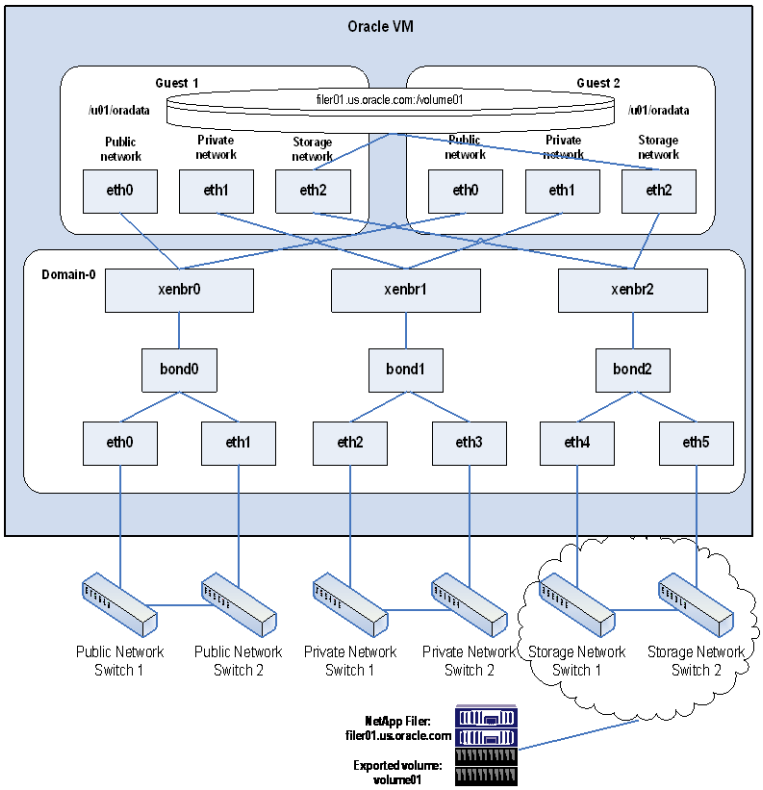

Figure 10: Full network bonding in Oracle VM

[11] See **Oracle® Clusterware Administration and Deployment Guide 11*g* Release 2 (11.2)**, Chapter 2, Administering Oracle Clusterware

Sizing the Oracle VM Environment for Oracle RAC

As a general rule of thumb, you should "size your Oracle VM environment in a similar way as you would size a non-VM environment." Following this rule when sizing an Oracle VM system Oracle RAC should cover most of the optimization. Particular attention should be paid to the following parameters:

CPUs and cores

Follow the requirements as stated in this document for assigning vCPUs to dom-0 and guest domains. It is supported, but not recommended, to over-commit by a factor of 2.

Memory

Memory is not shared between domains in Oracle VM. When sizing memory for your Oracle RAC VM system, make sure your system memory is large enough to accommodate the memory needed for Domain-0, the Linux Operating systems, Oracle Clusterware, and the Oracle RAC databases in each of the guest domains.

Network

For an Oracle RAC in Oracle VM environments the general recommendation is to install a minimum amount of 2 NICs per Oracle VM host machine. For a production environment the recommendation is to have at least 4 network cards installed in the host and use bonding.

Using a shared network bridge, bandwidth is shared between the different guest domains installed on one Oracle VM server. In production environments, these network interfaces must be at a minimum 1 Gbit Ethernet, for example 1000BASE, however, 10GbE, 10GBASE is supported.

When configuring a larger number of guest VMs on one host, or when deploying guest VMs that require a high amount of network bandwidth, conscientious monitoring for oversubscription events resulting in packet loss, either on the NIC or the switch, and/or increased latencies and bandwidth degradation must be implemented.

Storage

When using SAN storage for your Oracle RAC in Oracle VM deployment, more than one HBA can be used to use a higher bandwidth to the storage in the Oracle VM host. The multipathing drivers usually combine the bandwidth of each path automatically.

When using NAS storage (preferably using iSCSI), it is recommended to use a dedicated network interface to access the NAS storage. The dedicated storage network should be configured in the same way as the public or interconnect network. Again, multiple NICs can be used to eliminate a single NIC as a Single-Point-Of-Failure and appropriate bonding configurations can be used to increase bandwidth.

INSTALLING ORACLE RAC IN ORACLE VM ENVIRONMENTS

The following steps describe the complete installation of Oracle RAC in an Oracle VM environment. Oracle recommends to use the Oracle RAC Oracle VM templates for a rapid provisioning of Oracle RAC on Oracle VM.

1. Plan your installation carefully. Consider the requirements for your new environment and plan accordingly using the guidelines and best practices in this white paper and the official Oracle documentation.
2. Install the appropriate version of Oracle VM that you want to use (Oracle VM 2.1.5 through 3.0.3 are supported) on the machine(s) on which you want to perform the Oracle RAC installation. Follow the Oracle VM Server Installation Guide for the details on how to do this.
3. From the Oracle VM Manager GUI, discover the servers and configure storage and networking. When using the certified Oracle RAC templates, time synchronization in Domain-0 of your Oracle VM host(s) is correctly pre-configured. Otherwise, you will need to verify and correct it on each server as required.
4. Create the para-virtualized guests domains for the Oracle RAC installation.
5. Install and configure the guest operating system (Oracle Linux 5.1 or higher). Make all necessary configuration changes to the guest operating system as described in the requirements for installing Oracle Clusterware and Oracle RAC. Use udev to configure the raw devices required for installing Oracle RAC. Optionally, ASMLib can be used in the guest domains in order to further simplify OS disk administration.
6. Install Oracle Clusterware and Oracle RAC 10.2.0.4, 11g Rel. 1 or 11g Rel. 2 as for non-virtualized environments, following the installation guides for these products. Consult My Oracle Support for more and up-to-date information regarding the installation and certification of Oracle RAC on specific Operating System versions.

USE THE ORACLE RAC ORACLE VM TEMPLATES

Using Oracle VM Oracle RAC templates provide an easy and fast way of deploying an Oracle RAC cluster in Oracle VM **for test** as well as **production environments**.

The 11g Rel. 1 based templates are designed to build a 2-node Oracle 11*g* Release 1 (11.1.0.7) RAC cluster, which includes an Oracle 11g Database, Oracle 11g Clusterware, and Oracle 11g Automatic Storage Management (ASM). More nodes can be added easily. The 11g Release 2 based templates allow for flexibility, including the creation of *N*-node clusters and installations using role separation.

The entire installation is automated, requesting minimal information (e.g. node names, IP addresses, etc.) from the user when the first guest is booted at creation time. The final cluster configuration adheres to the best practices for production systems as described in this paper.

The current Oracle VM Oracle RAC templates are available supporting a variety of Oracle Linux 5.x based system images for either 32-bit or 64-bit. The OS image contains a minimal installation of Oracle Linux 5.x. Only basic RPM packages are installed. Configuration and benefits at a glance:

- The Oracle VM templates build a 2-node Oracle RAC 11*g* Release 1 (11.1.0.7) or *N*-node Oracle RAC 11g Release 2 (11.2.0.x) cluster, which includes Oracle 11g Clusterware, Oracle 11g Database, and Oracle Automatic Storage Management (ASM), patched to the latest recommended patches. More nodes can be added easily.
- The environment comes loaded with Swingbench, Oracle Cluster Health Monitor, OS Watcher, and more tools.
- During the installation process an Oracle RAC database instance is created on both (all) nodes.
- Find more information on how to configure an Oracle 11*g* RAC
    - on Oracle VM in **test** or **production** systems.
- Download the Oracle VM Template for Oracle RAC 11*g* Release 1
- Download the Oracle VM Template for Oracle RAC 11g Release 2

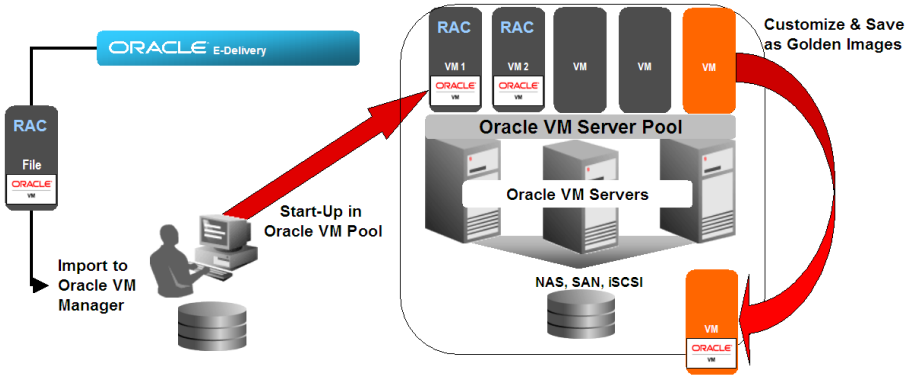

Figure 11: Oracle Template Deployment

LIVE MIGRATION AND DATABASE RELOCATION BEST PRACTICES

Oracle VM Live Migration Best Practices

Live Migration of VMs offers operating system (OS) level migration to facilitate workload management, fault management and system maintenance operations. Oracle Live Migration basically consists of iteratively pre-copying the contents of a source VM's memory from one physical host to another target VM on a remote host with minimal interruption to running services.

This iterative pre-copy of the memory continues until it is determined that the dirty page rate on the source machine is consistently higher than the memory transfer rate to the target machine. At this point, the VM is stopped and all remaining dirty pages are copied from the source VM to the target VM.

The stop-and-copy operation to complete the Live Migration is a suspend time blackout of the VM. For most applications, the dirty page rate will be low and the suspend time will measure in milliseconds to seconds. But for large highly active applications with potential dirty page sets of greater than 2.7GB, the suspend time of the VM could be considerably longer, potentially violating clusterware heartbeat thresholds which would trigger node fencing at the Oracle Clusterware level.

To avoid these potential failures during a Live Migration of VMs the following best practices should be observed:

1. Prior to initiating a Live Migration in an Oracle RAC production environment redefine the Oracle Clusterware misscount from the default of 30 seconds to 60. Issue the following command as root:

   `crsctl set css misscount 60`

   This will allow a Live Migration for an application with a potential dirty page set of greater than 2.7GB. Assuming a conservative maximum throughput for a single 1GbE shared link provides 90MB/sec., 60 seconds misscount would allow a suspend period of 5.4GB (90MB/sec * MC) where MC= misscount set to 60 seconds).

2. When the Live Migration is complete, set the Oracle Clusterware misscount back to the default. Issue the following command as root:

   `crsctl unset css misscount`

Only one Live Migration should be executed at a time. Multiple, simultaneous Live Migrations within the clustered environment may violate the suspend time thresholds described above.

If possible, a dedicated, high bandwidth network should be used for Live Migrations and the VM guests should be on the same network and IP subnet. The default network used by Live Migration is the OVM management network that may be the public network. If Live Migration is used as a frequent maintenance operation, it is recommended to deploy bonded interfaces used to separate the Dom0/ovs-agent/OCFS2/Live Migration traffic from other application traffic.

While Secure Live Migration is enabled by default, it should be considered that a secure connection to the remote guest (using –ssl) adds overhead to the Live Migration operation. It is therefore recommended to avoid such secure connections, if permitted. In general, a secure connection can be avoided, if the Live Migration network is inherently secure.

NOTE: The pause, suspend and resume OVM management functionality for Guest domains should be avoided when the Guest Domain contains an actively running production RAC instance. Any pause or suspend operation could trigger the clusterware to fence the suspended Oracle VM.

Oracle RAC One Node Online Database Relocation Best Practices

The best practices for Oracle RAC One Node Online Database Relocation remain the same when performing a Database Relocation operation within an Oracle VM based environment as opposed to executing this procedure on physical machines. More information on the Oracle RAC One Node Online Database Relocation can therefore be found in the Oracle RAC Administration and Deployment guide. When performing an Online Database Relocation between two Oracle VM guests, consider the following:

1. The target VM for a Database Relocation event must have adequate OS/system resources to allow for an Oracle RAC One Node instance to run.
2. Only one Database Relocation event at a time should be executed.
3. The target VM must be a configured member of the cluster.

SUMMARY

Having certified Oracle RAC in Oracle VM through release 3.0.3 and following the recommendations made in this paper; there are no limitations in utilizing these standard technologies in (Oracle) enterprise grid and cloud infrastructures like the one illustrated in fig. 9 or even more sophisticated configurations in the future.

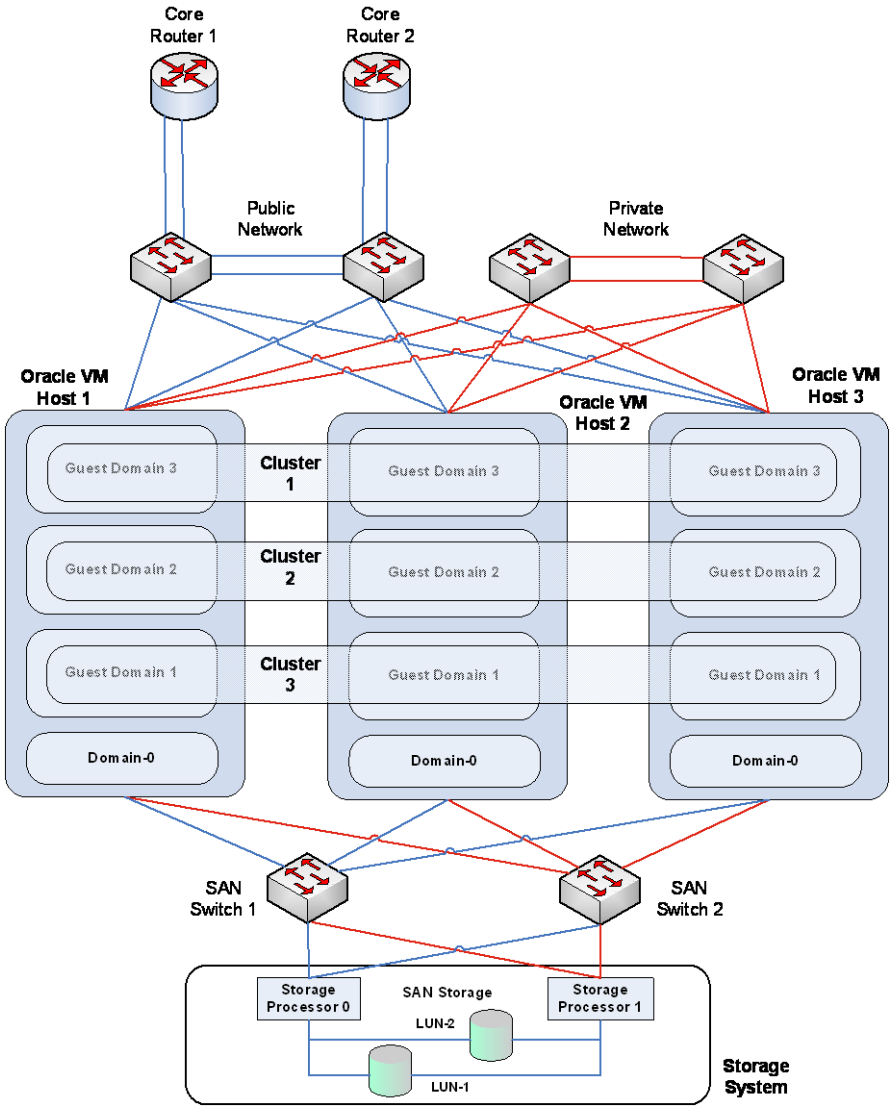

Figure 12: Production Architecture Example

APPENDIX A – GUEST CONFIGURATION FILE EXAMPLE

```
# xen config file example for RAC Guest Domain
name = "vmhost01-g01"
memory = "16384"
disk = [
'phy:/dev/shared/sdsk-a1-l20-spb,xvda,w!',
'phy:/dev/shared/sdsk-a1-l21-spb,xvdb,w!',
'phy:/dev/shared/sdsk-a1-l22-spb,xvdc,w!',
 ]
vif = [
'mac=00:16:3E:00:00:08, bridge=xenbr0',
'mac=00:16:3E:10:A5:96, bridge=xenbr1',
 ]
vfb = ["type=vnc,vncunused=1"]
uuid = "3d6f1de4-626c-e02a-42a1-458c9c17e728"
bootloader="/usr/bin/pygrub"
vCPUs=8
on_reboot   = 'restart'
on_crash    = 'restart'
```

APPENDIX B – BONDING SETUP EXAMPLE

Disclaimer: Making changes to the network configuration of an Oracle VM machine to the following extend must be performed using remote or direct console access to the machine. In the course of this configuration, network connectivity to dom-0 will be lost. Direct or remote console access will be required to finish the configuration.

Introduction

The following example shows how to create two bonding devices and two bridges to be used by the guests, based on 4 Ethernet Network Interface Cards (NIC), configured in dom-0. The network devices used in this example are eth0, eth1, eth2 and eth3. This setup is meant to support an Oracle RAC running in VM guest domains. A separate public and private network is required.

The eth0 and eth1 interfaces are dedicated to "public" connectivity. These interfaces will be part of the bond0 bonding device. The bond0 bonding device will be attached to the xenbr0 bridge. The dom-0 IP address will also be configured on bridge xenbr0 in order to ensure connectivity to dom-0.

The eth2 and eth3 interfaces are dedicated to "private" connectivity. These interfaces will be part of the bond1 bonding device. The bond1 bonding device will be attached to the xenbr1 bridge. These interfaces are not used for external connectivity and therefore do not get any other IP addresses assigned.

The bonding mode used in this example is mode 1, active-backup. This is the easiest mode and will work on most switches without any problems. Any other modes might require specific switches and a certain switch configuration.

Step 1: Restoring and disabling the xend Network Configuration

Execute the following commands as root to restore the network configuration changed by xend:

```
# cd /etc/xen/scripts/
# ./network-bridges stop
```

Create a script that only returns "true" in the /etc/xen/scripts directory.
For example: **/etc/xen/scripts/network-bridge-dummy**

```
#!/bin/sh

/bin/true
```

In the /etc/xen/xend-config.sxp change the following line from:

```
(network-script network-bridges)
```

to

```
(network-script network-bridge-dummy)
```

Step 2: Loading the Bonding Modules

Add the following lines to the /etc/modprobe.conf:

```
alias bond0 bonding
alias bond1 bonding
options bonding max_bonds=2
```

Add more alias lines and increase the max_bonds parameter as required.

Step 3: Configuring the network devices and bridges for public connectivity

To configure the eth0 and eth1 interfaces, edit **/etc/sysconfig/network-scripts/ifcfg-eth0** file as follows:

```
DEVICE=eth0
BOOTPROTO=none
ONBOOT=yes
MASTER=bond0
SLAVE=yes
USERCTL=no
HWADDR=00:14:22:10:A5:F7
```

For **/etc/sysconfig/network-scripts/ifcfg-eth1:**

```
DEVICE=eth1
BOOTPROTO=none
ONBOOT=yes
MASTER=bond0
SLAVE=yes
USERCTL=no
HWADDR=00:14:22:10:A5:F8
```

The HWADDR is the pointer to the network card (MAC address) being used. The SLAVE parameter defines this network card as a slave of a bond-device. The MASTER parameter points to the actual bonding device, which this network interface will be part of. Then configure the bonding device bond0 as follows:

In **/etc/sysconfig/network-scripts/ifcfg-bond0** set:

```
DEVICE=bond0
ONBOOT=yes
BRIDGE=xenbr0
```

```
BONDING_OPTS="mode=active-backup miimon=100
downdelay=200 updelay=200 use_carrier=1"
```

Note that the parameter BONDING_OPTS used in the configuration file illustrated above allows for setting the bonding options for each individual bonding interface. The BRIDGE parameter specifies that this bonding interface does not have an IP address configured to it, but will be connected to bridge xenbr0.

In **/etc/sysconfig/network-scripts/ifcfg-xenbr0** set:

```
DEVICE=xenbr0
ONBOOT=yes
STP=off
IPADDR=130.35.166.150
NETMASK=255.255.252.0
```

An IP address is assigned to xenbr0 using the IPADDR and NETMASK in order to ensure connectivity to dom-0. The parameters NETWORK and BROADCAST are deprecated. These will be automatically calculated with ipcalc.

Step 4: Configuring the network devices and bridges for private connectivity

To configure the eth2 and eth3 interfaces, edit **/etc/sysconfig/network-scripts/ifcfg-eth2** as follows:

```
DEVICE=eth2
BOOTPROTO=none
ONBOOT=yes
MASTER=bond1
SLAVE=yes
USERCTL=no
HWADDR=00:04:23:BB:54:66
```

For: **/etc/sysconfig/network-scripts/ifcfg-eth3**

```
DEVICE=eth3
BOOTPROTO=none
ONBOOT=yes
MASTER=bond1
SLAVE=yes
USERCTL=no
HWADDR=00:04:23:BB:54:67
```

Then configure the bonding device bond1 by editing **/etc/sysconfig/network-scripts/ifcfg-bond1** as follows:

```
DEVICE=bond1
ONBOOT=yes
BRIDGE=xenbr1
BONDING_OPTS="mode=active-backup miimon=50 downdelay=200
updelay=200 use_carrier=1"
```

In **/etc/sysconfig/network-scripts/ifcfg-xenbr1** set:

```
DEVICE=xenbr1
ONBOOT=yes
STP=off
```

Step 5: Restart Oracle VM

To activate the changes made to the network configuration, its is recommended to restart the machine, by executing the reboot command in dom-0.

If required: Configuring bridges on interfaces without bonding

For the purpose of completeness and since this guide explains how to use bonding, which requires to disable the xend based network and bridge configuration, in the following it is explained how to set up bridges on a non-bonded interfaces, which otherwise would be performed by xend automatically.

This example assumes a configuration of a bridge configured directly on top of an Ethernet interface, but without the use of bonding. This example uses the network interface eth4 and bridge xenbr3.

Edit **/etc/sysconfig/network-scripts/ifcfg-eth4** as follows:

```
DEVICE=eth4
BOOTPROTO=none
ONBOOT=yes
USERCTL=no
BRIDGE=xenbr3
HWADDR=00:04:23:AA:51:23
```

In **/etc/sysconfig/network-scripts/ifcfg-xenbr3** set:

```
DEVICE=xenbr3
ONBOOT=yes
STP=off
IPADDR=192.168.20.34
NETMASK=255.255.255.0
```

The IPADDR and NETMASK parameters are optional and only required if connectivity to dom-0 needs to be established through this interface.

APPENDIX C – REFERENCES

Oracle VM on oracle.com : http://www.oracle.com/virtualization

Oracle VM on the Wiki : http://wiki.oracle.com/Oracle+VM

Oracle RAC on oracle.com: http://otn.oracle.com/rac

Oracle Clusterware on oracle.com: http://otn.oracle.com/clusterware

Oracle ASM on oracle.com: http://otn.oracle.com/asm

Oracle ASMLib:

http://www.oracle.com/technology/tech/linux/asmlib/index.html

Oracle RAC documentation:

http://download.oracle.com/docs/cd/B19306_01/install.102/b14203/toc.htm

http://www.oracle.com/pls/db111/to_toc?pathname=install.111/b28263/toc.htm
http://www.oracle.com/pls/db111/to_toc?pathname=install.111/b28264/toc.htm

http://www.oracle.com/pls/db112/to_toc?pathname=install.112/e10812/toc.htm
http://www.oracle.com/pls/db112/to_toc?pathname=install.112/e10813/toc.htm

Metalink Notes

Note 414163.1 : 10gR2 RAC Install issues on Oracle EL5 or RHEL5 or SLES10

Note 738269.1 : Enabling network bonding in Oracle VM

Note 735975.1 : Oracle VM: Configuring Quality of Service (QoS) for Guest Virtual Machines

Note 580296.1 : Clock drift issue between Dom-0 and Dom-U on OVM Server

Note 564580.1 : Configuring raw devices (multipath) for Oracle Clusterware 10g Release 2 (10.2.0) on RHEL5/OL5

Note 465001.1 : Configuring raw devices (singlepath) for Oracle Clusterware 10g Release 2 (10.2.0) on RHEL5/OL5

Note 790189.1 : Oracle Clusterware and Application Failover Management

APPENDIX D – TABLE OF FIGURES

APPENDIX E – KNOWN LIMITATIONS AND OPEN ISSUES

- Oracle VM Manager 3.0 must be used for all guest management functions. Command Line Interface support is currently not available.

- Oracle VM Manager 2.1.2 cannot be used to create Guest VMs running Oracle RAC. This is due to the requirement for Oracle RAC to use physical devices as Virtual Block Devices (VBDs).
    - Oracle VM Manager 2.2 and higher does not have this limitation.
    - File backed VBDs are not supported in an Oracle RAC on Oracle VM for mission critical, production deployments

- Physical storage for Oracle ASM disks must be used (Bug13700449)
- Unpublished BUG 13767911 requires the use of a common, dedicated Oracle VM management network between the Oracle VM manager node and the OVS nodes.
- CVU check for "sharedness" of disks fails due to BUG 12965027
- When physically removing a public NIC cable on the server (no bonding) the Oracle Clusterware managed VIP in the OVM guest does not failover as the failure of the physical connection is not reported (BUG 13104709)

- No support for HugePages in OVM 2.2.1 and earlier versions.
- Guest OS resides in the storage repository of server pool.

- Oracle ACFS was not tested with 11.2.0.3 on UEL 5.7
    - There is currently no Oracle ACFS support for UEL5.7.

Oracle Real Application Clusters in Oracle VM Environments
March 2012, Version 2.1a
Authors: John McHugh, Markus Michalewicz
Contributing Authors: Saar Maoz, Nitin Vengurlekar, Billy He, Yunzhou Zheng

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

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