

# Oracle VM – Creating & Maintaining a Highly Available Environment for Guest VMs

*An Oracle Technical White Paper  
September 2008*

# Oracle VM – Creating & Maintaining a Highly Available Environment for Guest VMs

Executive Overview .....	3
Advanced Enterprise Virtualization Platform.....	3
Maximize Up-Time: Oracle VM Guest HA Functionality.....	4
How It Works .....	5
Architectural Overview: Oracle VM’s HA-Related Concepts.....	5
Key Concepts: Server Pool And Server Pool Master .....	6
Figure 1. High Level Deployment Architecture.....	6
Guest VM HA Management Infrastructure .....	7
Key Concepts: Quorum And Fencing .....	7
Reliable Guest VM Or Host Server Failure Detection.....	9
Figure 2. Minimize Or Eliminate Downtime .....	10
Eliminate Service Outages With <i>Secure</i> Live Migration.....	12
Figure 3. SSL Encrypt Migration Traffic .....	12
Predictable Start-Up & Availability: Pool Load Balancing.....	12
Figure 4: Avoid Down Servers - Automatic Pool Load Balancing.	13
HA Integration With Secure Migration And Server Restart.....	14
High Availability Management Infrastructure .....	14
Highly Available Enterprise Virtualization.....	14

# Oracle VM – Creating & Maintaining a Highly Available Environment for Guest VMs

## ORACLE'S CERTIFIED VIRTUALIZATION SOLUTION: ORACLE VM

- Complete server virtualization and management with no license costs;
- Speeds application deployment with Oracle VM Templates;
  - Modern, low-overhead architecture for leading price/performance;
- Includes Secure Live Migration, VM High-Availability, and other advanced features.

## EXECUTIVE OVERVIEW

By disassociating workloads from the physical constraints of the underlying hardware, Oracle VM presents the opportunity to significantly reduce service outages associated with server hardware downtime – whether planned or unplanned.

This white paper provides an overview of the functionality provided by Oracle VM to help you dramatically improve guest VM and workload uptime, including features such as Server Pool Load Balancing, Secure Live Migration, and Guest VM High Availability.

## ADVANCED ENTERPRISE VIRTUALIZATION PLATFORM

Oracle VM is a free-license, state-of-the-art server virtualization and management solution that makes enterprise applications easier to deploy, manage, and support. Backed worldwide by affordable enterprise-quality support for both Oracle and non-Oracle environments, Oracle VM facilitates the deployment and operation of your enterprise applications on a fully certified platform to reduce operations and support costs while simultaneously increasing IT efficiency and agility.

Oracle VM has been developed specifically to host enterprise server workloads: your most critical applications, middleware, and databases. In this environment, maximum service availability is critical, but you also have a budget to maintain, and the complexity needs to be manageable: you just can't afford to implement traditional high-availability clusters for each and every component.

Oracle VM helps by providing advanced VM availability management:

- **Guest VM High Availability features minimize unplanned outages –**
  - Included VM High Availability functionality automatically restarts individual failed VMs or entire sets of VMs that were running on a failed server;
  - Flexible deployment: Use with NFS (NAS) or OCFS2 (SAN/iSCSI)
- **Secure Live Migration eliminates outages from planned downtime –**
  - Migrate running VMs from one physical server to another *securely*.

**Secure Live Migration: Oracle VM is the first major virtualization solution to SSL-encrypt migration traffic by default to protect sensitive data from exploitation and eliminate the requirement for dedicated migration networks.**

- Oracle VM is the first major virtualization solution to SSL-encrypt migration traffic by default to protect sensitive data from exploitation. Most migration products don't offer native encryption, creating vulnerabilities, and necessitating dedicated migration networks unless you incur the extra complexity and expense to purchase SSL hardware;
- **Server Pool Load Balancing prevents VM start-up “blocking”–**
  - The physical host for a guest VM is automatically selected from the pool of healthy, available servers by Oracle VM at guest VM power-on based on a pool load balancing and availability algorithm;
  - Oracle VM's server resource pooling and shared-storage architecture assures that a down server does not block guest start-up, to help maintain predictable service levels;
  - Optionally, for each individual guest VM, users can specify a unique list of named servers, called Preferred Servers to be used for hosting that guest VM to further tailor to unique performance and availability needs.
- **A distributed architecture and optional clustering of Oracle VM Manager maximizes management uptime –**
  - Oracle Enterprise Linux management servers with an Oracle Unbreakable Linux support subscription can be clustered with no additional license costs using Oracle Clusterware to permit automatic management service fail-over and recovery to minimize down-time without requiring manual intervention;
  - A distributed management architecture allows most VM operations including Secure Live Migration and Guest HA Auto Restart to succeed even if the management server is temporarily unavailable.

### **MAXIMIZE UP-TIME: ORACLE VM GUEST HA FUNCTIONALITY**

One of the most powerful benefits of having a virtualized environment is that you no longer need to be constrained by the limited level of availability provided by a single server. Traditional HA clustering solutions can provide excellent availability for workloads across multiple servers, but often at a high cost from both a licensing and complexity perspective. These costs are often justified for the most critical and extreme situations, where continuous availability is required, but there is likely a much larger set of servers and services that, while perhaps not absolutely required to be continuously available, are nevertheless required to provide an absolute minimum of downtime. Further, through the sheer volume of these servers alone, an availability solution that is not complex to configure or maintain is essential to keeping operations, training, and support costs low.

Leveraging the expertise that resulted in the first clustered file system to be adopted into the Linux kernel, as well as the first clustered database, Oracle VM's Guest VM HA functionality provides a powerful, easy-to-manage solution for maximizing

up-time for virtually any guest VM workload, without requiring any tailoring inside the VM, making it simple to set-up, use, and maintain. Oracle VM Guest HA functionality provides the following benefits:

- Auto-restart unexpectedly failed individual VMs on other servers in the server pool;
- Auto-restart all the guest VMs on another server in the server pool when an unexpected physical server failure occurs;
- Powerful cluster-based network- and storage heartbeat algorithms quickly and deterministically identify failed and/or isolated servers in the server pool to ensure rapid, accurate recovery;
- Sophisticated distributed lock management functionality for NFS, SAN, and iSCSI storage ensures VMs or entire servers can be rapidly restarted with no risk of data corruption.

### How It Works

To understand how Oracle VM's Guest HA features are some of the most sophisticated in the industry, this white paper takes a closer look at some key aspects of the solution:

- Oracle VM's architecture and concepts as related to HA;
- How the VM or server failure is detected reliably;
- How the VM restart is accomplished while assuring data consistency;
- How the HA features are integrated across the product to assure maximum operational up time.

### Architectural Overview: Oracle VM's HA-related Concepts

Oracle VM includes two major components:

- **Oracle VM Server:** Installs on bare metal Intel and AMD x86 and x86\_64 servers to provide the environment for hosting guest VMs. This incorporates a xen.org open source hypervisor component integrated into the larger, Oracle-developed Virtualization Server (Oracle VM Server);
- **Oracle VM Manager:** Web-based management solution for centrally managing large numbers of Oracle VM Servers and guest VMs. Oracle VM Manager comprises:
  - A Java-based management server;
  - An Oracle Database management repository (Database Express Edition, Standard Edition, Enterprise Edition, or Real Application Clusters). This database can reside on either the management server or on a separate server, depending on scalability and availability requirements;
  - A web-browser based GUI;

- An Oracle VM Server management agent on each Oracle VM Server used to communicate with the management server and to issue management commands to the Oracle VM Server.

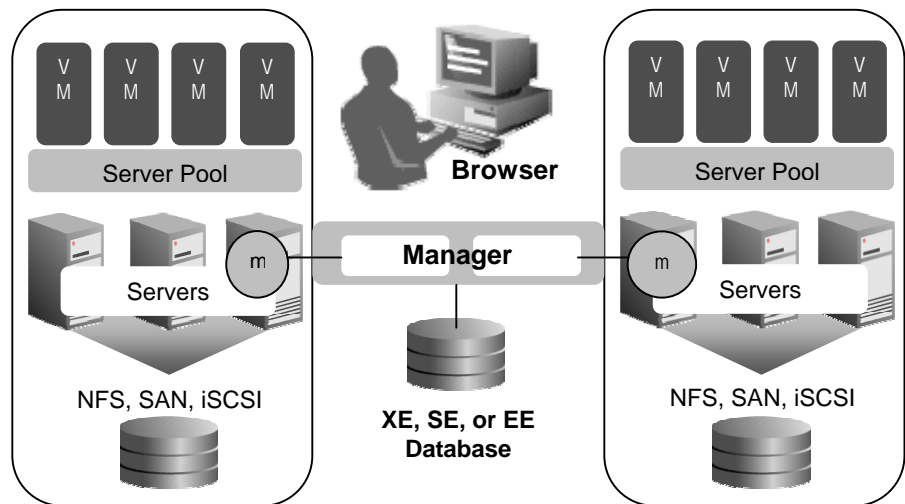
**Key Concepts: Server Pool and Server Pool Master**

From a deployment perspective, multiple Oracle VM Servers are grouped into *Server Pools* in which every server in a given pool has access to shared storage, which can be NFS or SAN/iSCSI storage. This allows VMs associated with the pool to start and run on any physical server within the pool that is available and has the most resources free. Given the uniform access to shared storage, VMs may also be securely Live Migrated or automatically (re-)started across any servers in the pool.

VMs are associated with a given Server in the pool dynamically at power-on based on load balancing algorithms or based on a user-defined list of named servers to be used as a host for that specific VM called the Preferred Server list. When VMs are powered-off and not running, they are not associated with any particular physical server since they are simply residing in a powered-off state on shared pool storage.

As a result of this architecture, VMs can easily start-up, power-off, migrate, and/or restart without being blocked by the failure of any individual server or, by the failure of multiple pool servers as long as the aggregate amount of resources is adequate to support the aggregate requirements of all the VMs running concurrently in the pool.

**Dynamic load balancing: VMs are associated with a given Server in the pool dynamically at power-on based on load balancing algorithms or based on a user-defined list of named servers to be used as a host for that specific VM.**



**Figure 1. High Level Deployment Architecture**

In each Server Pool, there is a *Server Pool Master Agent* that coordinates a number of the activities of the pool, particularly when that action requires coordination across multiple servers. This includes such key activities as coordinating Secure Live Migration actions as well as Guest VM HA auto-restart actions amongst others.

All of the management agents in the pool communicate directly with the master, who, in turn, communicates with the Oracle VM management server. This

architecture provides a number of benefits including high management scalability in large environments, but also higher availability at the Oracle VM Manager instance level by distributing and isolating functionality to minimize the impact of any single failure. For example, a management server outage does not prevent the ability of the pool master(s) to complete Secure Live Migration tasks or to automatically fail over/restart failed VMs. Similarly, a pool master outage on one pool does not affect the operation of another pool.

The Pool Master server can either be a dedicated server or a server that also hosts guest VMs depending on the scalability and availability requirements. For maximum availability, it's recommended that the Pool Master be deployed as a dedicated server.

### **Guest VM HA Management Infrastructure**

In order to assure predictable, reliable, and accurate restarting of failed VMs, it's critical to have a very tightly integrated HA management system to orchestrate everything from the VM failure detection all the way through to the successful restart of the guest VM. It's equally important to insure that there is no opportunity for data corruption anywhere during the process.

Leveraging its deep expertise with OCFS (Oracle Cluster File System), Oracle Real Application Clusters (RAC), and the associated Oracle Clusterware, Oracle has developed an advanced architecture for managing guest availability that goes beyond reliance on simple network "pings" to determine whether a guest is running or not. The result is a solution that is more reliable and dramatically reduces the opportunity for false positives/negatives when determining whether a VM has failed. It also assures that a VM is restarted correctly without any risk of shared data corruption.

Although transparent to the user from an installation perspective, Oracle VM has incorporated Oracle OCFS2 clusterstack into the core product as part of its infrastructure to, in effect, transform server pools into clusters from a high availability perspective. This allows guest VM HA to be managed at a level that is more robust and reliable as compared to competitive HA schemes that rely on simplistic network pings combined with time-outs to determine if a VM or Server has failed. Because those solutions are too basic, they have a greater potential to result in falsely declaring VMs or servers to be failed, or, worse, end up actively shutting-down healthy VMs and/or not restarting them when needed. Oracle VMs architecture essentially eliminates these scenarios as a concern.

*Note: For the remainder of this white paper, the terms "pool" and "cluster" should be considered to be interchangeable terms unless otherwise specified.*

### **Key Concepts: Quorum and Fencing**

Before continuing, it's useful to understand several clustering concepts in order to understand how Oracle VM HA/auto-restart happens.

In HA clustering, it's important that all nodes in the cluster understand both their status and the status of every other node in the cluster in order to assure that all activity is appropriate for the cluster.

Accordingly, each Oracle VM Server in the pool supports the following functions:

- Cluster node management - contains configuration information on every server in the pool;
- Heartbeat management (for both network and storage connections)- for VM/Server failure detection and management as described later in this document;
- Distributed lock management (“DLM”) - to assure VMs can be safely restarted without any risk of corruption caused by attempts at simultaneous modification of shared data as described below.

Since each node contains this functionality, any node can fail at any time and the surviving nodes can coordinate to rapidly execute on recovery in a way that is safe for all the virtual machines and servers in the cluster.

### **Quorum**

When the cluster is healthy and each node has full connectivity to all the other nodes, the cluster is said to have a *quorum*.

This means that all of the servers running in that cluster (pool) have uniform access to all the servers and the shared storage necessary to permit activities such as Secure Live Migration and HA restart. Accordingly, servers that are outside the quorum – because they have lost network and/or storage connectivity with the rest of the nodes in the cluster – do not have full access to the status of the pool and thus need to be dealt with as quickly as possible, e.g. returned to a cluster that has a quorum so they can be managed appropriately or shut down. All the nodes in the cluster are constantly checking-in with each other to assure that they still have a quorum, or to react if it appears that the quorum has been lost due to a failure.

A healthy, valid server pool can only have one quorum at a time to assure full functionality and data consistency. Later in this document, we'll discuss how this is managed in the context of various failure scenarios.

### **Fencing**

If any nodes in the cluster, and thus their hosted VMs, get isolated from the main cluster due to, for instance, a network failure, it is important that they be quickly prevented from trying to access the resources of the cluster, especially storage, because they no longer have a current and consistent view of the state of those resources. As a result, nodes in this situation must be *fenced* to block them accessing the resources of a cluster they no longer belong to. Most other HA clustering solutions, including Oracle Real Application Clusters, perform self-fencing so that they automatically shut themselves down and/or remove themselves from the

cluster under specific, pre-defined circumstances, e.g. if they cannot communicate with the anyone else in their cluster or network. This is often the most practical, reliable method of fencing because, by definition, in this situation the node is not externally reachable and thus cannot be reliably shut down by any outside force.

### **Reliable Guest VM or Host Server Failure Detection**

**Deterministic HA Management: Oracle VM uses not only network heartbeat management but also a storage heartbeat (sometimes known as a “quorum disk”) to obtain an extremely accurate picture of VM status at all times.**

Oracle VM uses a sophisticated, cluster-based heartbeat management scheme to validate that it has a quorum at all times. If the quorum is lost, and not all nodes can contact each other, then the heartbeats can be used to determine which node is failing or has become isolated from the cluster and thus initiate appropriate actions.

Oracle VM uses not only network heartbeat management but, unlike competitive solutions, also a storage heartbeat (sometimes known as a “quorum disk”) to obtain an extremely accurate picture of VM health at all times. With a quorum disk, every node in the quorum reads and writes a tiny amount of status to the same section of shared storage on a regularly scheduled basis. Each node writes its status and reads the status of everyone else. As a result, each node quickly becomes aware if any of the other nodes has failed to update its status. The first node that detects a heartbeat problem (network or storage) with any other node then initiates the process to recover the isolated or failed VMs back into the quorum.

There are two scenarios to consider here:

1. *Failure of an individual VM on a healthy physical Server;*
2. *Failure of a physical server that may contain multiple guest VMs.*

### **Detecting and Recovering from a Single Guest VM Failure on a Healthy Physical Server**

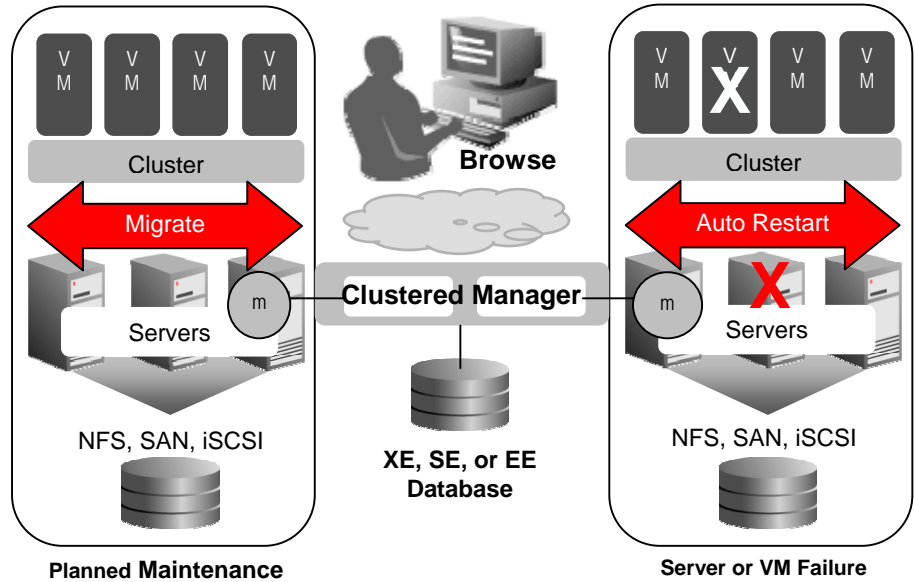
As mentioned previously, the distributed architecture of Oracle VM ensures that each server pool contains a Pool Master Agent to orchestrate some of the activities within the server pool. The master is consistently in contact with every server in the server pool to monitor status and issue commands on behalf of the management server and/or server pool.

If, during a periodic check from the pool master, an individual VM is determined to be in an unexpected state – for example, it’s still in “Running” status but the list of running VMs on that Server does not include that particular VM – the pool master will attempt to restart that specific VM. Because Oracle VM uses a load balancing algorithm to place a VM on the best server in the pool (i.e. with the most resources available), the VM may be restarted the same- or a different node in the pool (or on one of the specific servers you have specified on your Preferred Server list for that VM).

Lock management will assure that no data corruption occurs in shared storage during the process. Even in the extremely unlikely event that the VM is still running and accessing network and storage, and that, due to some other failure, the reported status itself was in error, the distributed lock management functionality

will assure that a duplicate VM instance is not created and thus there is no risk of data corruption.

In the case of a single failed VM on a healthy server, nothing was fenced because it was a failure that was isolated to that specific VM and did not affect the host node itself.



**Figure 2. Minimize or Eliminate Downtime for Planned- or Unplanned Events**

#### ***Detecting and Recovering from Physical Server Failure***

Detecting and recovering from a failed server that may have multiple guest VMs is very similar in concept. In the vast majority of cases, it's a single server or network connection that has gone down, but in the worst case, or in the case of poor network configuration, multiple servers may be isolated from the cluster by the failure of a single network device.

Since all nodes use network and storage heartbeats to check the status of all the cluster members regularly, rather than just one or two IP addresses, the failure of a node will quickly be accurately detected by one of the nodes in the cluster. The detecting node then initiates a process whereby all other nodes are asked to attempt to contact each other so that a connectivity "map" can be created to validate which server or network link might have failed or otherwise become isolated from the cluster.

All of the nodes that have full connectivity to each other form a quorum and any node(s) that are determined to have incomplete communication to the pool, whether through network failure/isolation or server failure, are fenced from the cluster in order to protect the health of the cluster as a whole.

The self-fencing process causes the servers and their VMs to automatically restart but, based on the power-on load-balancing algorithms and lock management, they will only restart on nodes that are a part of the cluster quorum so that they are returned to a healthy pool where Secure Live Migration and further HA actions can occur normally as needed.

***Special Cases: “Split-Brain” Clusters and Node Isolation***

There are a couple of interesting cases that must be accounted-for as part of any enterprise-class HA solution: Split-Brain clusters and node isolation.

**Split-brain clusters** result from a situation whereby a network failure, for example, splits the cluster into two (or more) equal, but smaller clusters. Each surviving cluster would want to declare itself as the governing quorum that should have and control access to the shared storage, etc. With Oracle VM, each individual node contains algorithms to check for exactly this situation when forming a new quorum. In this case, based on the storage heartbeat and other inputs, the nodes will determine whether they should establish a quorum or restart themselves. The restart would cause them to be fenced and cause the VMs to auto-restart, releasing their lock(s) on the storage so that the other surviving cluster could declare itself the quorum and thus begin hosting the re-started VMs.

**Node isolation** occurs when an otherwise healthy server continues to run and even perform storage I/O but where the node has become completely isolated from the rest of the cluster, likely as a result of network failure: it cannot be contacted by anyone nor can it contact any other node. While the node may currently be healthy, this situation is generally undesirable over any period of time since it cannot provide status, cannot be actively managed, and cannot participate in secure live migrations or auto-restarting of its VMs if it does eventually fail unexpectedly.

In this case, again, each node contains algorithms to detect just such a situation and initiate a restart of the Server and thus the VMs hosted on that server. Once the shutdown has completed safely, the storage locks will be released and the VM(s) will be automatically restarted in the quorum. The VM power-on will automatically occur because Oracle VM Manager’s pool master expects the VM(s) from that node to be running and if they are not running in the quorum, it will initiate a power-on. If there are locks on disk preventing that power-on because it is still in the process of shutting-down on the isolated node, then Oracle VM will wait and try again. But if there are no locks, then the VM(s) will start normally in the pool based on the load balancing algorithms.

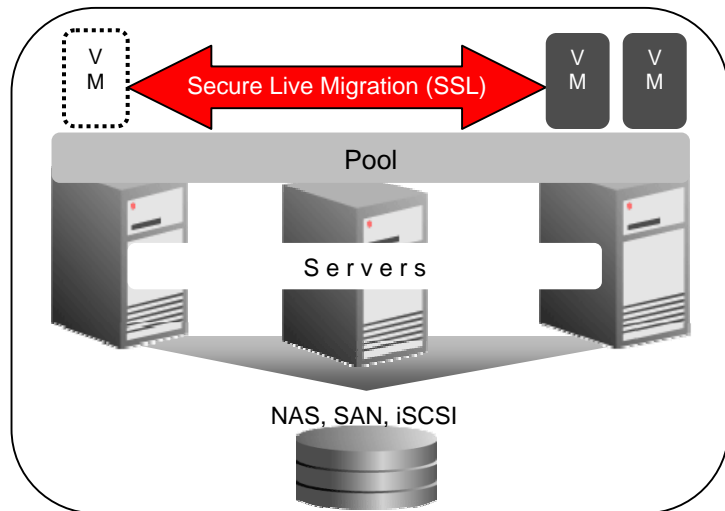
This process is deterministic and, unlike other products, Oracle VM does not rely on guesses or assumptions about the status of the isolated node, which can result in, for example, isolated nodes shutting down without their VMs being restarted in the quorum.

## ELIMINATE SERVICE OUTAGES WITH SECURE LIVE MIGRATION

Oracle VM's guest VM HA features provide the ability to maximize service uptime through unexpected failures at an affordable cost and with low complexity, but how about eliminating outages altogether when you can see them coming?

Oracle VM's Secure Live Migration eliminates outages associated with planned downtime by allowing quick and easy migration of running VMs from one physical server in another. No need to take a service outage and service users won't even notice the change unless they notice how much faster things are once you've migrated to larger hardware.

Most migration products have a problem today: they are not secure. They don't encrypt the migration traffic between servers so all your sensitive data – account numbers, passwords, etc – goes over the wire unencrypted, just as it is in memory, when you migrate.



**Figure 3. SSL Encrypt Migration Traffic To Eliminate Vulnerability**

Oracle VM is designed specifically for the production enterprise running critical workloads and Secure Live Migration uses native SSL encryption by default, and without any requirements for additional hardware, to eliminate the vulnerability that exists today in nearly every other major product on the market. The ability to encrypt the traffic also enables the use of shared networks for migration traffic without fear of exposing sensitive data to exploitation.

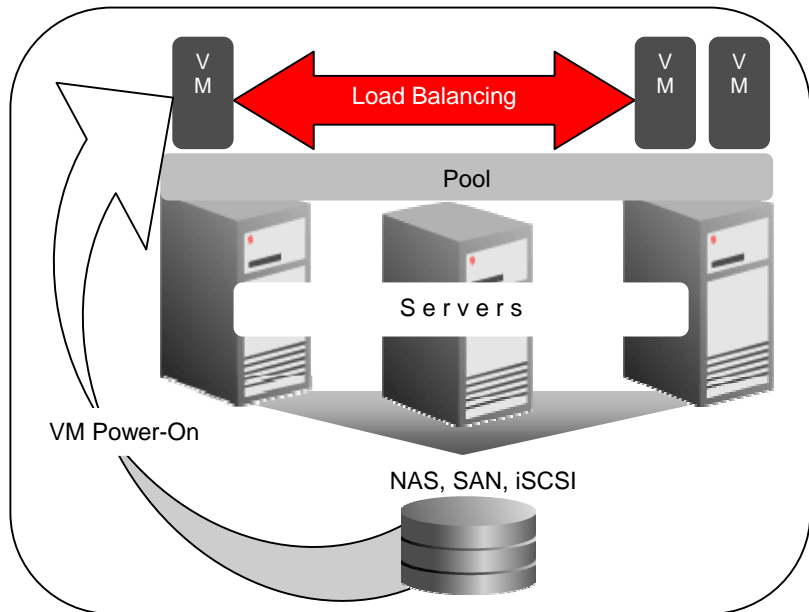
## PREDICTABLE START-UP & AVAILABILITY: POOL LOAD BALANCING

As previously described, Oracle VM's deployment architecture utilizes server pools with shared access to storage across all servers in the pool. Guest VMs are stored on the shared storage and placed on one of the servers in one of two ways:

- Either automatically using Oracle VM's pool load balancing algorithms to automatically select the host server with the most resources available;

- Or from a user-specified Preferred Server list that allows the user to designate a sub pool of named servers where the VM is allowed to start and run.

Since the VMs are not bound to any specific physical server in the pool unless that is explicitly specified via the Preferred Server list, VMs will not be prevented from powering-on simply because an individual server happens to be down for maintenance or otherwise unavailable at that time. Further, since the load-balancing algorithm assures that the VM is placed on the server with the most resources available, it also helps assure the maximum aggregate performance from the pool.



**Figure 4: Avoid Down Servers - Automatic Pool Load Balance at VM Start**

Should you prefer to tailor which server(s) within the pool should be used as the host(s) in order to more finely tune performance, scalability, or simply based on licensing or organizational issues, you can define Preferred Server lists. It can also be used to fine-tune availability: For example, you could use Preferred Server sub pools to assure that multiple components in the same application stack are never on the same physical server at the same time.

Preferred Server lists are created on a per-VM basis that allows, in effect, the creation of a sub pool for that individual VM. The list dictates where the VM can start, run, restart (HA), and which servers can be used for Secure Live Migration. While this obviously allows fine-tuning of the environment, the greatest availability is generally achieved when the VM(s) are permitted access to the largest number of nodes in the pool.

## HA INTEGRATION WITH SECURE MIGRATION AND SERVER RESTART

Oracle VM's guest VM HA features have also been integrated across the product to make it easy to maintain high availability during planned maintenance or other scheduled activities:

- When a physical server is instructed to power-off from Oracle VM Manager, the user will be provided with the option to Securely Live Migrate any or all of the hosted VMs prior to power-off to maintain VM availability easily;
- On server power-off, any VMs that are not migrated, but that still have the HA option enabled will automatically restart on other servers in the pool when the physical server is shutdown, eliminating the need for administrators to go back and manually restart each VM when a server has been powered off;
- HA functionality can be quickly disabled or re-enabled at either the pool level (for all VMs in the pool) or at the individual VM level to prevent unintended restarts during maintenance periods.

For information on how Oracle VM Templates can help you rapidly deploy pre-built VMs containing sophisticated, enterprise software, see the [Oracle VM Templates website](#).

## HIGH AVAILABILITY MANAGEMENT INFRASTRUCTURE

Management operations go on around the clock so you want to maximize your infrastructure up time. Oracle Enterprise Linux-based management servers with an Oracle Unbreakable Linux support subscription at the Basic and Premier levels can be clustered with no additional license costs using Oracle Clusterware to permit automatic management service fail-over and recovery to minimize down time without requiring manual intervention to restart and restore the server.

But even a temporary outage of the management server does not prevent the VMs from running and Oracle VM's distributed management architecture allows most VM operations including Secure Live Migration and Guest HA Auto Restart to succeed even if the management server is temporarily unavailable.

## HIGHLY AVAILABLE ENTERPRISE VIRTUALIZATION

Virtualization should help you solve your problems, not create new ones. Oracle VM has been developed specifically for use in the production enterprise data center to not only make applications easier to deploy, manage, and support but also to increase availability in a secure environment.

By disassociating workloads from the physical constraints of the underlying hardware, Oracle VM helps you dramatically improve guest VM and workload uptime, including features such as Server Pool Load Balancing, Secure Live Migration, and Guest VM High Availability.

### More Information

For more information on Oracle VM and to download a complete, fully featured release for free, visit the Oracle VM web page on Oracle.com.



Oracle VM – Creating & Maintaining a Highly Available Environment for Guest VMs  
September 2008  
Author: A. Hawley

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

Copyright © 2008, Oracle. All rights reserved.

This document is provided for information purposes only and the contents hereof are subject to change without notice.

This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission. Oracle, JD Edwards, PeopleSoft, and Siebel are registered trademarks of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.