

## Design Document

### Open HA Cluster agent for x86-64 xVM Guest Domains

*[HA-xVM]*

*Neil Garthwaite*

*14<sup>th</sup> February 2008*

### Revision History

<i>Version</i>	<i>Comments</i>	<i>Date</i>	<i>Author</i>
1.0	Initial draft	14/02/08	Neil Garthwaite
1.1	Review comments	19/03/08	Neil Garthwaite

## 1 Introduction

The OpenSolaris x86-64 xVM Server is a hypervisor based on Xen and Solaris on x86-64 systems. Running the xVM hypervisor allows the physical server to be virtualized, essentially virtualizing the system's hardware.

Once the xVM hypervisor is running, it is possible to create guest domains that utilize this virtualized environment. In this regard, with the OpenSolaris x86-64 xVM Server it is possible to host several heterogeneous x86-64 guest domains on top of a physical server.

While the benefit of platform virtualization encourages server consolidation it introduces a problem that the physical platform now becomes a huge single point of failure. With today's x86-64 servers providing several multi-core cpus and several GBs of memory, a physical server failure hosting several heterogeneous x86-64 guest domains soon becomes a recovery headache.

In this regard, the Open HA Cluster (OHAC) agent for x86-64 guest domains aims to mitigate against a physical server failure by managing the guest domains. In particular the agent will perform the following:

- Manage the start/stop and restart of x86-64 xVM guest domains within an OHAC/xVM (dom0) environment.
- Failover a x86-64 xVM guest domain between OHAC/xVM (dom0) nodes.
- Allow for strong positive and negative affinities between x86-64 xVM guest domains across OHAC/xVM (dom0) nodes.
- Allow for different failover techniques, i.e. Stop/Failover/Start, Migration or Live Migration.

At this point as the terms "Migration" and "Live Migration" are being introduced, it is important to understand that these xVM failover techniques do not mitigate against a physical server failure. In particular, Migration and Live Migration require that both the source and target node remain active until the migration has completed. Migration is therefore not high availability.

Nevertheless, Migration and Live Migration are extremely powerful and useful failover techniques when one wants to move a x86-64 xVM guest domain between OHAC/xVM (dom0) nodes. In this regard, failover that uses Migration or Live Migration implies a move.

The Open HA Cluster (OHAC) agent for x86-64 xVM guest domains will be designed as a failover agent and will be based on the Generic Data Service (GDS) which is an existing Solaris Cluster Resource Type that provides a robust and well tested environment.

As the design of this agent will use GDS, we need to provide several method scripts and functions that essentially provide the "glue" between OHAC and xVM (dom0). These method scripts and functions control the management of the x86-64 xVM guest domains. The remaining part of this document will discuss the expected behavior of these method scripts.

## 2 Control script

As the name suggests, this script provides the high level control and interaction with OHAC/xVM (dom0) and the x86-64 xVM guest domains.

This script will be responsible for the following:

- Retrieve any required resource properties.

- Determine the method to execute, i.e. Validate, Start, Stop or Probe.
- Execute the appropriate function for the method.

It is proposed that this control script will be located and named as `/opt/SUNWscxvm/bin/control_xvm`.

### **3 Validate function**

The validate function is called from the control script and determines if the runtime environment is acceptable in which to start a x86-64 xVM guest domain. The runtime environment is determined by values that are supplied by the administrator when the resource is registered. Please see “7. Configuration File” for more information.

The validate function will determine the following:

- The OHAC/xVM node is running as Domain-0.
- The guest domain pathname is valid.
- The failover type specified is either normal, migrate or migrate -live.

While the term failover implies the action to perform after a physical server failure, it is possible to failover a x86-64 xVM guest domain when all the nodes are active. In this regard, one may want to move x86-64 xVM guest domains from one node to another to either spread the load or evacuate workload from a server in preparation for some server maintenance.

When moving a x86-64 xVM guest domain from one node, it is possible to either gracefully shutdown the guest domain, migrate or live migrate the guest domain between nodes. These specific failover types are described below:

- Normal
  - Stop the resource (gracefully shutdown the domain).
  - Failover the resource group from the source node to the target node.
  - Start the resource (start the domain).
- Migrate
  - Suspend the domain on the source node.
  - Copy the domain's memory pages from the source node to the target node.
  - Resume the domain on the target node.
- Live Migrate
  - Iteratively copy the domain's memory pages from the source node to the target node.
  - When this “pre-copy” is no longer beneficial, suspend the domain on the source node.
  - Copy the domain's remaining memory pages from the source node to the target node.
  - Resume the domain on the target node.

The validate function is called whenever the OHAC agent for x86-64 xVM guest domains is registered, updated or whenever the agent is about to start. Failure from the validate function will result in either the OHAC xVM resource not being registered, updated or a start failure to occur.

It is proposed that the validate function will be located within the functions file `/opt/SUNWscxvm/bin/functions` as `validate()`.

### **4 Start function**

The start function is called from the control script and is responsible for starting a x86-64 xVM guest domain. This function is called whenever the OHAC

xVM resource is started or restarted.

The start function will perform the following:

- Ensure that the OHAC Process Monitor Facility (PMF) is turned off. PMF is inherently built into GDS and by default PMF will expect at least one long running process that PMF then monitors. However, the semantics of starting a x86-64 xVM guest domain are different which simply means that a long running process is not running on the OHAC/xVM (dom0) node. Therefore, the start function needs to turn of PMF monitoring within the GDS resource.
- Ensure that the x86-64 xVM guest domain to be started has been defined to the OHAC/xVM (dom0) node.
- If the x86-64 xVM guest domain does not exist, the start function should attempt to define it if the domain definitions were previously exported to an XML file. The scenario here is where a domain has been defined on one node and is now being failed over to another node where the target node is not aware of the domain, i.e the domain has not been defined on the target node. However, for this to happen then domain's definitions must have been previously exported.
- The start function should tolerate a manually started x86-64 xVM guest domain. In this scenario, if the domain was manually started and afterwards the administrator wants to place the domain under the control of the OHAC agent for x86-64 guest domains, then the start function should tolerate this and not require the domain to be stopped and then started again.
- The start function should adhere to a NO-OP request.

At this point as the term "NO-OP" is being introduced, it is important to understand that a NO-OP start is only performed after a NO-OP stop. As such a NO-OP stop is performed whenever a Migration or Live Migration has been successfully performed. In this scenario, the OHAC xVM resource managing the x86-64 xVM guest domain has requested a failover type of Migrate or Live Migrate and that whenever possible this failover technique will be used. As such, if the migration is successful a NO-OP stop and start should be performed. The purpose of this is simply to ensure that the OHAC Resource Group Manager (RGM) can action any dependencies and reflect the correct state and status of the resource groups and resources.

- If the x86-64 xVM guest domain is successfully started, the start function should dump the domain's definitions to an XML file. As discussed earlier this XML file can then be used to define the domain on a new OHAC/xVM (dom0) node, but only if the domain does not exist on that node.
- Start the x86-64 xVM guest domain using OpenSolaris xm or virsh commands.

It is proposed that the start function will be located within the functions file /opt/SUNWscxvm/bin/functions as start\_domain().

## ***5 Stop function***

The stop function is called from the control script and is responsible for stopping the x86-64 xVM guest domain. This function is called whenever the OHAC xVM resource is stopped or restarted.

The stop function will perform the following:

- Determine if the OHAC xVM resource is being disabled or switched to another node. If a resource is being disabled, the stop function must

perform a shutdown regardless of the failover type requested. In this scenario the administrator has requested that the x86-64 xVM guest domain is to be shutdown.

- Adopt a phased shutdown of the x86-64 xVM guest domain. If domain is being shutdown, the stop function needs to adopt a phased approach. In this regard, 70% of the Stop\_timeout (600 seconds) will be allocated to a graceful shutdown of the domain, after which 25% of the Stop\_timeout will be allocated to an immediate shutdown of the domain. Using the default allocated Stop\_timeout of 600 seconds this equates to 450 and 150 seconds respectively. Please note that the Stop\_timeout value can be adjusted. The remaining 5% of Stop\_timeout is set aside for OHAC processing.
- The stop function should perform a migration or live migration if the failover type has requested this. In this scenario, the administrator has requested a switch of the resource group using OHAC commands. It is important to understand that the administrator can issue a switch command on any OHAC node and has the choice of two commands – clrg or scswitch. In this regard the stop function needs to determine the target host and validate it is an appropriate target host for the migration.
- The stop function needs to tolerate a node evacuation. In this scenario the OHAC/xVM (dom0) node is either being rebooted or the administrator is requesting that OHAC should evacuate all resource groups from the node. It is important to understand here that the administrator can issue an evacuate command on any OHAC node and has the choice of two commands – clrg or scswitch. In this regard the target host is not explicitly known, as such the stop function will always perform a phased shutdown regardless of the failover type requested. This ensures that RGM determines the failover target node for a node evacuate.
- If the stop function is performing a migration or live migration of the the x86-64 xVM guest domain and that migration is not successful, then the stop function needs to perform a phased shutdown as described earlier.
- If the stop function successfully performs a migration then a NO-OP stop should be performed. Furthermore, the stop function should request that the subsequent start function performs a NO-OP start as the x86-64 xVm guest domain will already be running.
- Stop the x86-64 xVM guest domain using OpenSolaris xm or virsh commands.
- If the stop function successfully stops a x86-64 guest domain, either via a graceful shutdown or via a migration/live migration, then the domain will be deleted from the OHAC/xVM (dom0) node. This is done to ensure that the domain is only ever defined on one OHAC/xVM (dom0) at any time which means that the x86-64 guest domain can only be started on one node at any time.

It is proposed that the stop function will be located within the functions file /opt/SUNWscxvm/bin/functions as stop\_domain().

## **6 Probe function**

The probe function is called from the control script and is responsible for periodically checking the x86-64 xVM guest domain state.

The probe function will perform the following:

- Check the x86-64 xVM guest domain state. The “virsh domstate” command will be issue against the x86-64 xVM guest domain but from Domain-0. It is therefore important to understand that the state of x86-64 xVM guest

domain is determined by xVM and it is that state that is then used by the OHAC xVM resource.

- The OHAC xVM resource should tolerate “acceptable” run states as reported by the “virsh domstate” command, namely running, blocked, paused and in shutdown. If any of these states exist for the x86-64 xVM guest domain the probe should indicate a healthy domain back to RGM. It is therefore important to understand that as far as the OHAC xVM resource is aware the domain has been booted and is available. However, this maybe somewhat misleading as the domain itself may require some additional administration such as a manual check and repair of a file system etc.
- The OHAC xVM resource should react to “restartable” run states as reported by the “virsh domstate” command, namely shut off or crashed. If any of these states exist for the x86-64 xVM guest domain the probe should indicate an unhealthy domain back to RGM. In this regard the domain will be restarted.
- The OHAC agent for x86-64 xVM guest domains should provide the ability for plugin probes. A plugin probe is a user defined probe that can perform specific checks against a x86-64 guest domain. The agent probe will call the plugin probe. A plugin probe will be supplied named ppkssh. This plugin probe performs a private/public key ssh authentication from Domain-0 to the x86-64 guest domain and will check either a Linux runlevel or an OpenSolaris/Solaris SMF service.

It is proposed that the probe function will be located within the functions file /opt/SUNWscxvm/bin/functions as check\_domain().

## ***7 Administrative file system***

As the name suggest, the OHAC agent for x86-64 xVM guest domains agent requires an administrative file system. This file system must be available on all OHAC/xVM (dom0) nodes and can be either a cluster file system or NFS. The administrative file system will contain NO-OP flags for STOP and START as well as a dumped XML file for each OHAC agent for x86-64 xVM guest domain.

## ***8 Configuration file***

The configuration file provides variable values to define the environment for the OHAC xVM resource. This file will be used by the registration script to register the OHAC xVM resource.

It is proposed that the configuration file will be located and named as /opt/SUNWscxvm/util/xvm\_config. Of particular interest the administrator will be able to specify the runtime environment for the xVM resource via the following variables:

```
DOMAIN=  
PLUGIN_PROBE=  
FAILOVER=  
ADMIN=
```

The value of FAILOVER must be either “normal”, “migrate” or “migrate -live” as discussed earlier on.

## ***9 Registration script***

The registration script uses as input the configuration file to register the OHAC xVM resource.

It is proposed that the registration script will be located and named as /opt/SUNWscxvm/util/xvm\_register.

### **10 Debug file**

If debug of the OHAC agent for x86-64 xVM guest domains is required a debug file will be provided to simply turn on debug for a OHAC xVM resource.

It is proposed that the debug file will be located and named as /opt/SUNWscxvm/etc/debug.

### **11 OHAC registration file**

The OHAC registration file is required by OHAC so that the OHAC agent for x86-64 xVM guest domains can be registered and upgraded by OHAC.

It is proposed that the OHAC registration file will be located and named as /opt/SUNWscxvm/etc/SUNW.xvm.

### **12 Packaging**

The following package name and files are proposed,

- Package name SUNWscxvm
- /opt/SUNWscxvm/bin/control\_xvm
- /opt/SUNWscxvm/bin/functions
- /opt/SUNWscxvm/bin/ppkssh
- /opt/SUNWscxvm/etc/config
- /opt/SUNWscxvm/etc/SUNW.xvm
- /opt/SUNWscxvm/util/xvm\_config
- /opt/SUNWscxvm/util/xvm\_register

### **13 Further Information**

- General information on OpenSolaris xVM  
<http://opensolaris.org/os/community/ha-clusters>
- General information on OpenSolaris xVM  
<http://opensolaris.org/os/community/xen>
- General information on RGM  
<http://docs.sun.com/app/docs/doc/819-2969/6n57kl13s?a=view>
- General Information on GDS  
<http://docs.sun.com/app/docs/doc/819-2972/6n57ngit6?a=view>
- START, STOP, PROBE and VALIDATE methods of GDS  
SUNW.gds(5) man page (/usr/cluster/man)
- General information on PMF  
pmfadm(1m) man page (/usr/cluster/man)