JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

VDCF - Virtual Datacenter Cloud Framework

for the Solaris™ Operating System

HA / High Availability

Version 4.0
12 November 2024

JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

Table of Contents

JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

1 Introduction

This documentation describes the HA features of the Virtual Datacenter Cloud Framework (VDCF) for the Solaris Operating System and explains how to use this features.

See these documents for more information about the related VDCF components:

VDCF – Administration Guide	for information about VDCF usage
VDCF – Monitoring Guide	for information about VDCF Monitoring

1.1 Overview

VDCF HA is a VDCF Enterprise extension available to VDCF Enterprise customers.

The VDCF High Availability feature is used to monitor the health of Nodes. If a failed Node is discovered the Node may be stopped and/or the Node evacuation logic is called to failover all vServers or Guest Domains to other Nodes. This evacuation is based on resource usage information to avoid overloading the remaining Nodes.

This solution is positioned between manual failover initiated by a System Administrator and a full-featured failover solution using Cluster software. This VDCF HA feature is able to handle the typical Node failures, like boot disk issues, network outages, platform errors like CPU, memory problems or power supply failures. The goal is to keep this solution as simple and usable as possible, therefore it doesn't require cluster interconnects between the Nodes and it doesn't check and handle issues with SAN connections like a Cluster software does.

1.1.1 Components

The HA monitor is built from several components:

Each Node participating has a daemon (SMF service vdcf_keep_alive) installed that calls periodically into the management server. These keep-alive messages are stored within the `/var/opt/jomasoft/vdcf/keepalive` directory.

The second component is the monitoring daemon (hamon_watchd) on the VDCF management server. This daemon consists of two processes. One (hamon_monitord) is used to monitor for keep-alive messages at the interval of HAMON_KEEP_ALIVE_INTERVAL seconds from all participating Nodes. The second process (hamon_checkd) is used to check and act upon a failed Node was detected.

JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

1.1.2 Node failure detection

A Node is considered as failed if the following rules are met:
- no keep-alive messages are received within the defined threshold `(HAMON_KEEP_ALIVE_ACTION_THOLD)`
- a ssh connection from VDCF to the Node fails
- Node's system controller / console does not respond or Node is at the OK prompt or powered off

An optional network probing rule may be activated by setting `HAMON_CHECK_NETWORK_PROBES="true"`. If the Node system controller is not reachable, the reason may be network-related or the Node has no power at all. If this setting is true, VDCF tries to connect to configured intermediate network equipment. If the network equipment is reachable, VDCF considers its network connection as good and therefor the Node as failed.

For more details about this failure detection consult the Appendix 5.1 Node failure detection details.

Based on the description above, the VDCF HA monitor is able to detect the following failures:
- complete hardware failure of the Node
- accidental shutdown of a Node by a System Administrator
- failure of network interfaces on the Node

The following failures are detected if network probing is activated and proper configured:
- complete power-failure of the Node (system controller not reachable)
- complete data center failure, as long as the network is still reachable (depends on configuration)

The following failures are **NOT** detected:
- failure or configuration issues of SAN components
- complete data center failure, if the network is affected (depends on configuration)
- accidental misconfiguration of a network interface by a System Administrator

For setting up and configuring your HA environment, consulting services from JomaSoft are available.

1.1.3 Node Evacuation sequence

For customers using LDoms and Zones: It must be defined which objects should be evacuated if a Control domain fails: The vServers or the Guest Domains.

Use this setting to decide which Objects should be failed over.

```
export HAMON_CDOM_FAILOVER=VSERVER | GDOM
```

JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

1.1.4 vServer Evacuation from Node

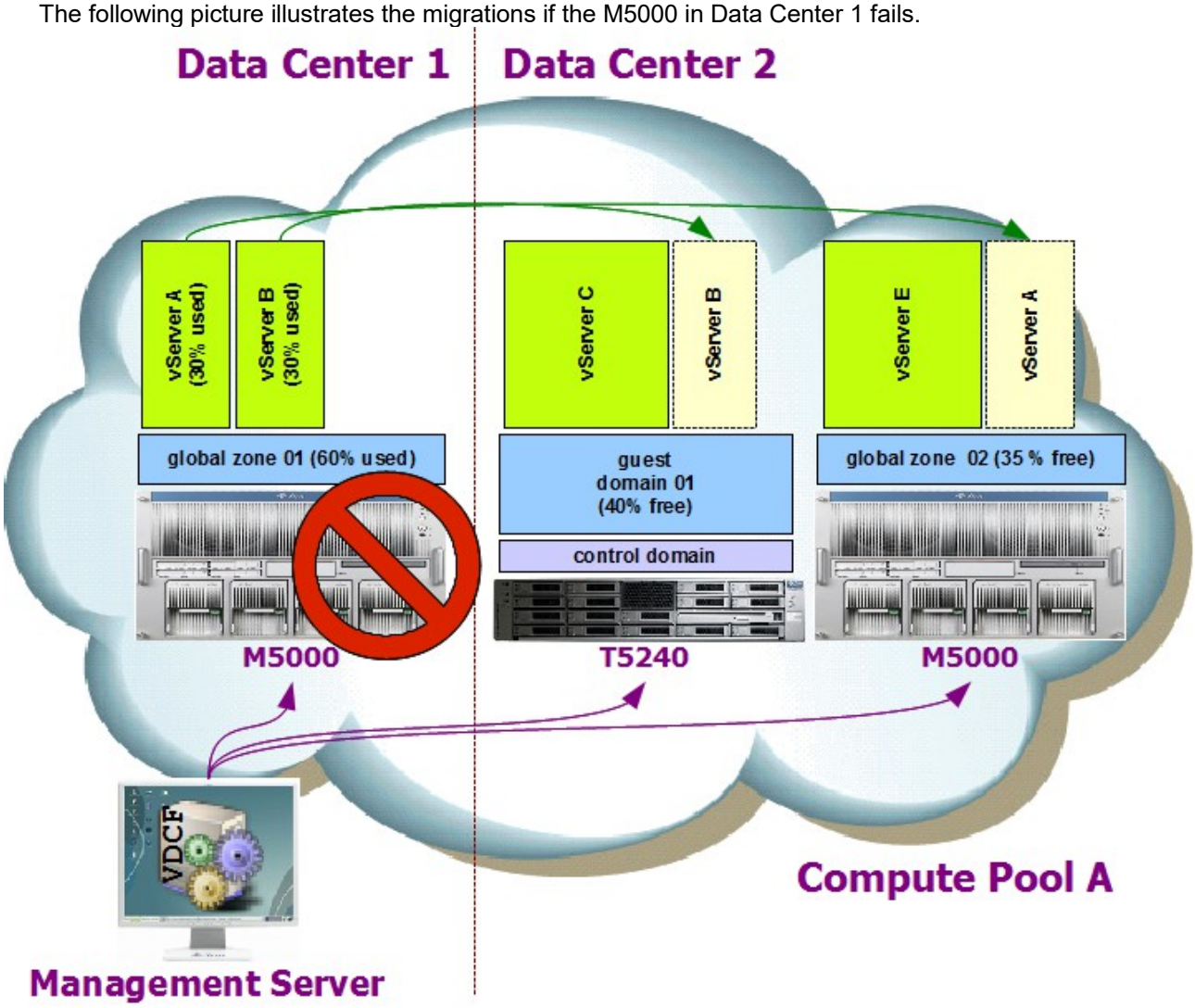
If Node Evacuation is configured, all vServers of a faulted Node are evacuated (failed over) to other active Nodes in the same computepool. The procedure to detect the possible target Nodes looks as follows:

1. For each vServer we get a list of candidate Nodes (using `vserver -c show candidates`).
2. Based on the resource usage data reported from resource monitoring we select a possible target Node for each vServer.
3. Because the source Node isn't reachable anymore we do a vserver detach force.
4. Then we try to attach and boot the vServer on the new Node.
5. If attach has failed we try the same procedure on the next possible target Node until all vServers are evacuated or no more target Nodes are left.

Upgrade on attach is supported by setting the value `HAMON_EVACUATE_UPGRADE` to true in the customize.cfg file.

The sequence of the vServer migration is ordered by the vServer category and/or priority. See configuration items for more details.

The following picture illustrates the migrations if the M5000 in Data Center 1 fails.



The Node Evacuation can be started manually using the command `node -c evacuate`.

1.1.5 GDom Evacuation from CDoms

If Node Evacuation is configured, all GDoms of a faulted CDom are evacuated (failed over) to other active CDoms in the same computepool. The procedure to detect the possible target CDoms looks as follows:

1. For each GDom we get a list of candidate CDoms (using `gdom -c show candidates`).
2. CDoms with the same hardware are preferred
3. Because the source CDom isn't reachable anymore we do a gdom detach force.
4. Then we try to attach and boot the GDom on the new CDom.
5. If attach has failed we try the same procedure on the next possible target CDom until all GDoms are evacuated or no more target CDoms are left.

The CDom Evacuation can be started manually using the command `cdom -c evacuate`.

This evacuation operation offers a `trial-run` flag to allow a verification run to show what would happen without executing the evacuation.

1.1.6 Requirements

Because the HA monitor is monitoring the console and is trying to shutdown a failed Node through the system controller, it's required to configure a 'console' for each Node within VDCF.

The Node evacuation logic for vServers is based on resource information from Resource Monitoring. Activated VDCF Resource Monitoring on all participating Nodes is therefore required.

JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

2 Installation

2.1 Prerequisites

The JSvdcf-ha package requires the following VDCF packages to be installed on the VDCF Management Server:

- JSvdcf-base 9.0.0 or later

- JSvdcf-monitor 4.0.0 or later

2.2 Installation

a) sparc platform

```
cd <download-dir>
pkgadd -d ./JSvdcf-ha_<version>_sparc.pkg
```

b) i386 platform

```
cd <download-dir>
pkgadd -d ./JSvdcf-ha_<version>_i386.pkg
```

JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

3 Configuration

3.1 Granting User Access

The VDCF HA package introduces a new RBAC profile

“VDCF hamonitor Module” for the Automated Failover

Assign this RBAC profile to your admin users.

3.2 Customizing HAMON

3.2.1 Keep Alive Interval

At each HAMON_KEEP_ALIVE_INTERVAL (default: 60 seconds) the Node is posting a keep-alive message to the Management Server.

3.2.2 Warning Threshold

After a number of missing keep-alive messages (HAMON_KEEP_ALIVE_WARN_THOLD (default 10) an e-Mail is sent if requested. Define your e-Mail addresses as follows:

```
export HAMON_EVENT_EMAIL_LIST="user1@company.ch user2@company.ch"
```

3.2.3 Action Threshold

A Node is considered as suspect if during HAMON_KEEP_ALIVE_ACTION_THOLD (default 20) intervals no keep-alive message has been posted.

You may display the current setting with the status command:

```
$ hamon -c status
             HA Monitor Information
                    Interval: 60s
           Warning Threshold: 10
            Action Threshold: 20
                Watch Daemon: disabled

                     VDCF Configuration Variables
              MONITOR_EVENT_EMAIL_FROM support@jomasoft.ch
                HAMON_EVENT_EMAIL_LIST support@jomasoft.ch
             HAMON_EVACUATE_ON_FAILURE false
             HAMON_POWEROFF_ON_FAILURE false
                   HAMON_CDOM_FAILOVER VSERVER
    VIRTUAL_EVACUATION_CATEGORY_ORDER
  VIRTUAL_EVACUATION_IGNORE_CATEGORIES
VIRTUAL_EVACUATION_SHUTDOWN_CATEGORIES
```

3.2.4 Actions on failure

Set HAMON_POWEROFF_ON_FAILURE to 'true' for a Node poweroff after failure detection.
This setting is highly recommended. If this setting is false, you risk to corrupt your data if the filesystems are mounted twice ...

Set also HAMON_EVACUATE_ON_FAILURE to 'true' if all vServers or GDoms of failed Nodes must be migrated to other running Nodes. If the failed Node is a Control Domain, all vServers running on dependent Guest domains are migrated to other Nodes if HAMON_CDOM_FAILOVER is set to 'VSERVER'. If HAMON_CDOM_FAILOVER is set to 'GDOM' the Guest Domains are failed over to other Cdoms.

3.2.5 Node evacuation

A Node is set to INACTIVE after an evacuate by default. Set HAMON_EVACUATE_INACTIVATE to 'false' to leave the Node in ACTIVE state.

vServer do not upgrade on attach by default. Therefore Nodes with a higher patch-levels aren't potential targets for the evacuated vServers. Set HAMON_EVACUATE_UPGRADE to 'true' to enable the upgrade on attach feature.

3.2.6 vServer target detection

First of all you have to categorize/prioritize your vServer using the `vserver -c modify` command. You may use categories to identify important or less important vServers and the priority to order within a category. vServers with Priority 1 are evacuated first, then Priority 2, ...

Then customize the evacuation variables in your customize.cfg.
Use VIRTUAL_EVACUATION_CATEGORY_ORDER to identify the most important categories to be migrated first. Identify categories which you don't want to evacuate at all in VIRTUAL_EVACUATION_IGNORE_CATEGORIES.

By default CPU_Share resource definitions aren't used for target Node detection. Set the NODE_EVACUATION_USE_CPUSHARES to 'true' to enable a check if the target Node has enough free CPU_Shares available.

3.2.7 vServer shutdown on target Nodes

Your target Nodes may not have enough free resources for the evacuated vServers. In such environments you can defined the Categories for less important vServers, which VDCF can shutdown to free resources. The vServers are shutdown only when required and ordered by the vServer Priority.

Define the Categories in VIRTUAL_EVACUATION_SHUTDOWN_CATEGORIES

3.2.8 Network reachability check

To enable the network reachability check you have to configure the HAMON_CHECK_NETWORK_PROBES to true and the HAMON_KEEP_ALIVE_NET_PROBE variable. The monitor selects the target probe address based on the Nodes MNGT interface and derives the network number from it. With this network number a search is done in HAMON_KEEP_ALIVE_NET_PROBE to find an associated probe address. If no match is found the default address is used if it is not set to 0.0.0.0. The variable HAMON_KEEP_ALIVE_NET_PROBE has the following format: `"net_number:probe_ip default:probe_ip net_number:probe_ip"`

3.2.9 Other recommended settings

The following are recommended settings. Please set these in the customize.cfg file:

```
export HAMON_EVENT_EMAIL_LIST="user1@company.ch user2@company.ch"
export HAMON_POWEROFF_ON_FAILURE="true"
export HAMON_EVACUATE_ON_FAILURE="true"
export HAMON_EVACUATE_UPGRADE="true"

# migration category order (comma separated categories)
export VIRTUAL_EVACUATION_CATEGORY_ORDER="PROD,ACC,BANK1"

# migration ignore categories (comma separated categories)
export VIRTUAL_EVACUATION_IGNORE_CATEGORIES="TEST,MAINT"
```

Optional settings

1. To lower the reaction times (Warn after 5 Mins, instead of 10 / Action 20 → 10)

```
export HAMON_KEEP_ALIVE_WARN_THOLD="5"
export HAMON_KEEP_ALIVE_ACTION_THOLD="10"
```

2. To take CPU_Shares into account for the check of free resources on target Nodes.

```
export HAMON_EVACUATE_USE_CPUSHARES="true"
```

3. To enable Network Probing (depends on your network infrastructure)

```
export HAMON_CHECK_NETWORK_PROBES=”true”
export HAMON_KEEP_ALIVE_NET_PROBE="192.168.0.0:192.168.0.1 10.1.1.0:10.1.1.1"
```

4. Define Shutdown Categories

```
# shutdown categories (comma separated categories)
export VIRTUAL_EVACUATION_SHUTDOWN_CATEGORIES="DEV,TEST,MAINT"
```

If High Availability monitoring was already enabled before, you have to re-enable the daemon to activate the new settings:

```
$ hamon -c disable daemon
$ hamon -c enable daemon
```

JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

4 Usage

4.1 Enabling / Disabling

The HA monitoring feature can be enabled/disabled globally.

```
$ hamon -c enable daemon
$ hamon -c disable daemon
```

Then each participating Node has to be enabled too:

```
$ hamon -c enable node=<node name>
$ hamon -c disable node=<node name>
```

Please notice that only non-cluster Nodes may be enabled for HA monitoring.

To display the status of HA monitoring use this command:

```
$ hamon -c status

             HA Monitor Information
                    Interval: 60s
           Warning Threshold: 10
            Action Threshold: 20
                Watch Daemon: online

                      VDCF Configuration Variables
                 MONITOR_EVENT_EMAIL_FROM support@jomasoft.ch
                   HAMON_EVENT_EMAIL_LIST support@jomasoft.ch
                HAMON_EVACUATE_ON_FAILURE true
                HAMON_POWEROFF_ON_FAILURE true
                      HAMON_CDOM_FAILOVER GDOM
     VIRTUAL_EVACUATION_CATEGORY_ORDER
   VIRTUAL_EVACUATION_IGNORE_CATEGORIES
VIRTUAL_EVACUATION_SHUTDOWN_CATEGORIES
```

JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

4.2 Display Node State

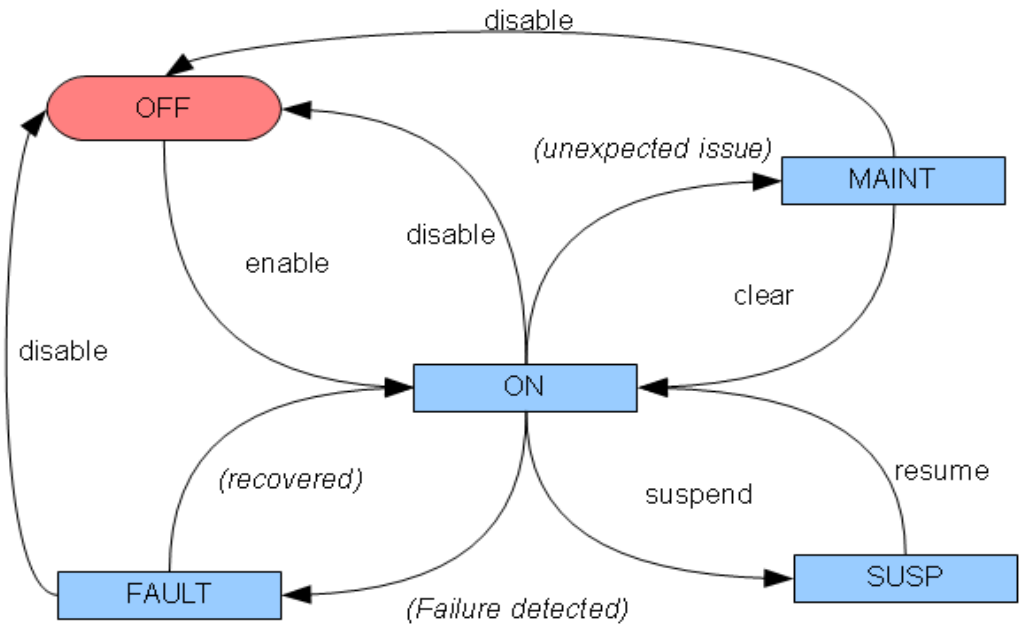
Using the show operation an overview of all Nodes is displayed.

```
$ hamon -c show

 Node  Mon State  Ops State  Date                Details
s0003  ON         PROBING    2011-02-16 16:37:48  normal operation
s0009  ON         PROBING    2011-02-16 16:32:43  normal operation
s0010  ON         PROBING    2011-02-16 16:38:19  normal operation
s0004  FAULT      -          2011-02-16 16:45:22  console did not respond /
not powered off
```

Each Node has a Mon(itoring) State, which is influenced by the System Administrator using hamon operations and by the VDCF HA monitor.

The following diagram explains the possible states and actions:



JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

4.3 Suspending Nodes

To avoid unnecessary failovers, it is required to suspend the Node from Monitoring if Maintenance is done on the Node. Suspend the Node before you shutdown the Node, for example to add more Memory.

```
$ hamon -c suspend node=s0003
HA monitor suspended on Node s0003

$ hamon -c show node=s0003
  Node  Mon State     Ops State     Date                  Details
 s0003  SUSP          -             2011-02-16 16:57:19   -

$ hamon -c resume node=s0003
HA monitor resumed for Node s0003

$ hamon -c show node=s0003
  Node  Mon State     Ops State     Date                  Details
 s0003  ON            PROBING       2011-02-16 16:57:33   normal operation
```

4.4 Fallback after Evacuation

Using the VDCF recommended settings, if a Node fails, the vServers/GDoms are evacuated and the Node is set to state INACTIVE. This is done to avoid usage of that Node for new objects.

You boot the Node when the issues, that caused the Node to fail, are solved, The HA Monitoring is then re-activated automatically. To use the Node again, you need to activate the Node:

```
$ node -c activate name=mynode
```

The vServers/GDoms are NOT automatically migrated back to the Node. You need to migrate the vServers manually back to your Node using the migrate operation.

```
$ vserver -c migrate name=myvserver node=mynode shutdown
```

The VDCF keep_alive SMF does remove evacuated LDoms from the System, if the VDCF management Server is reachable from the CDom. Messages are displayed on the CDom Console.

If is recommended to check the logfile of the VDCF keep_alive SMF on the Control Domain.

```
tail -50 /var/svc/log/site-vdcf_keep_alive:default.log
```

If the Control Domain looks ok, you can migrate the evacuated Guest Domains back.

```
$ gdom -c migrate name=mygom1 cdom=mycdomA shutdown | live
```

JomaSoft GmbH
Falkensteinstr. 54a
9000 St. Gallen
Switzerland

5 Appendixes

5.1 Node failover detection details

A Node is considered as failed if for a defined number of intervals no probe message has been posted from a Node. The monitor will kick off an action after (HAMON_KEEP_ALIVE_ACTION_THOLD+1) * HAMON_KEEP_ALIVE_INTERVAL seconds after a Node is no longer submitting its keep alive messages.

The action part of the hamon_check goes through several steps until it considers a Node as failed:

1. First of all network connectivity is verified by trying to check the status of the vdcf_keep_alive service on the suspect Node. If the Node can be reached and the check returns a service state other then enabled, the monitor tries to reestablish the vdcf_keep_alive service. If this succeeds, the monitor returns to normal operation and awaits the keep alive probe for this Node. If the service state already was enabled and the monitor was able to query its state, it also returns to normal operation, assuming the probe failure was of temporary nature.

2. If network reachability of the suspect Node is not given, the monitor tries to access the Nodes system controller. If we successfully reach the system controller the monitor checks the Node's console for a running operating system. In this case the monitor resumes normal operation, assuming a healthy Node with keep-alive failures due to temporary network problems. If the console check returns no signs of live the Node will be powered off, if configured so and its workload will be evacuated.

3. If the monitor is not able to reach the system controller and HAMON_CHECK_NETWORK_PROBES is true, the network will be checked. This is done by trying to reach intermediate network equipment as defined in HAMON_KEEP_ALIVE_NET_PROBE. If, based on this check, the network is considered as healthy, the suspect Node will be assumed as failed and the workload is evacuated. If the network is considered as failed, the monitor resumes normal operation without acting on the suspect Node.