Best Practices for SGeRAC and Oracle RAC on HP-UX 11i

March 2009

Introduction

This paper describes common best practices for configuring clusters with Serviceguard (SG), Serviceguard Extension for RAC (SGeRAC), Oracle® Clusterware, and Oracle® Real Application Cluster (RAC). A basic configuration is shown followed by specific preferred configurations with regards to the following topics:

- Availability for the public network
- Network for cluster communication
- Storage configuration
- Serviceguard packages

The information in this paper is applicable to Oracle 10g and Oracle 11g Release 1.

Audience

The intended audience is assumed to be familiar with HP-UX 11i, Oracle RAC, and SGeRAC.

Terms and definitions

- APA – Auto Port Aggregation provides bonding of multiple networking interface cards where traffic is distributed to all interface cards.
- APA/Hot Standby – Auto Port Aggregation Hot Standby mode provides high availability through bonding of a primary and a standby interface card. Traffic is not distributed.
- CFS – Cluster File System allows multi-system shared access to common file system.
- CSS – Cluster Synchronization Service is a component of Oracle Clusterware that maintains Oracle cluster membership and heartbeat.
- CSS-HB – Cluster Synchronization Service heartbeat traffic
- CVM – Cluster Volume Manager allows multi-system shared access to volumes.
- GAB – Group Membership Service/Atomic Broadcast manages cluster membership and cluster communication for Symantec Veritas CFS 4.1/5.0 and CVM 4.1/5.0.
- GMS – Group Membership Service refers to HP's implementation of the NMAPI2 API on HP-UX with SGeRAC that provides group membership notification and process monitoring facility to monitor group status.
- HA – High Availability refers to configurations that are resilient to single failure.
- LLT – Low Latency Transport provides kernel-to-kernel communications at link level and monitors network connections for Symantec Veritas CFS 4.1/5.0 and CVM 4.1/5.0. Distributes Symantec Veritas traffic amount network connections and maintains Symantec Veritas heartbeat.
- MNP – Multi-node package, a Serviceguard package that runs on multiple nodes at the same time and can be independently started and halted on individual nodes.
- OC – Oracle Clusterware can run in conjunction with Serviceguard Extension for RAC and provides Oracle cluster membership and resource management services.
- OCR – Oracle Cluster Registry is shared storage used to keep Oracle cluster and configuration information.
- ODM – Oracle Disk Manager is a standard API specified by Oracle for database I/O.
- RAC – Real Application Cluster enables a multi-instances concurrent shared access database.
- RAC-DB-IC – Real Application Cluster Interconnect traffic for both Global Cache Service and Global Enqueue Service.

- RIP – Serviceguard Relocatable IP Address user for client application access and failovers with package
- SG-HB – Serviceguard Heartbeat traffic.
- SGeRAC – Serviceguard Extension for RAC extends Serviceguard to support Oracle RAC.
- SLVM – Shared Logical Volume Manger allows multi-system shared access to LVM volumes for RAC.
- VIP – Virtual IP address is used by OC to configure access to Oracle clients and for remote failover to reject client connections.
- Voting Disk – Shared storage used by Oracle Clusterware as vote tie breaker and for disk based heartbeat.

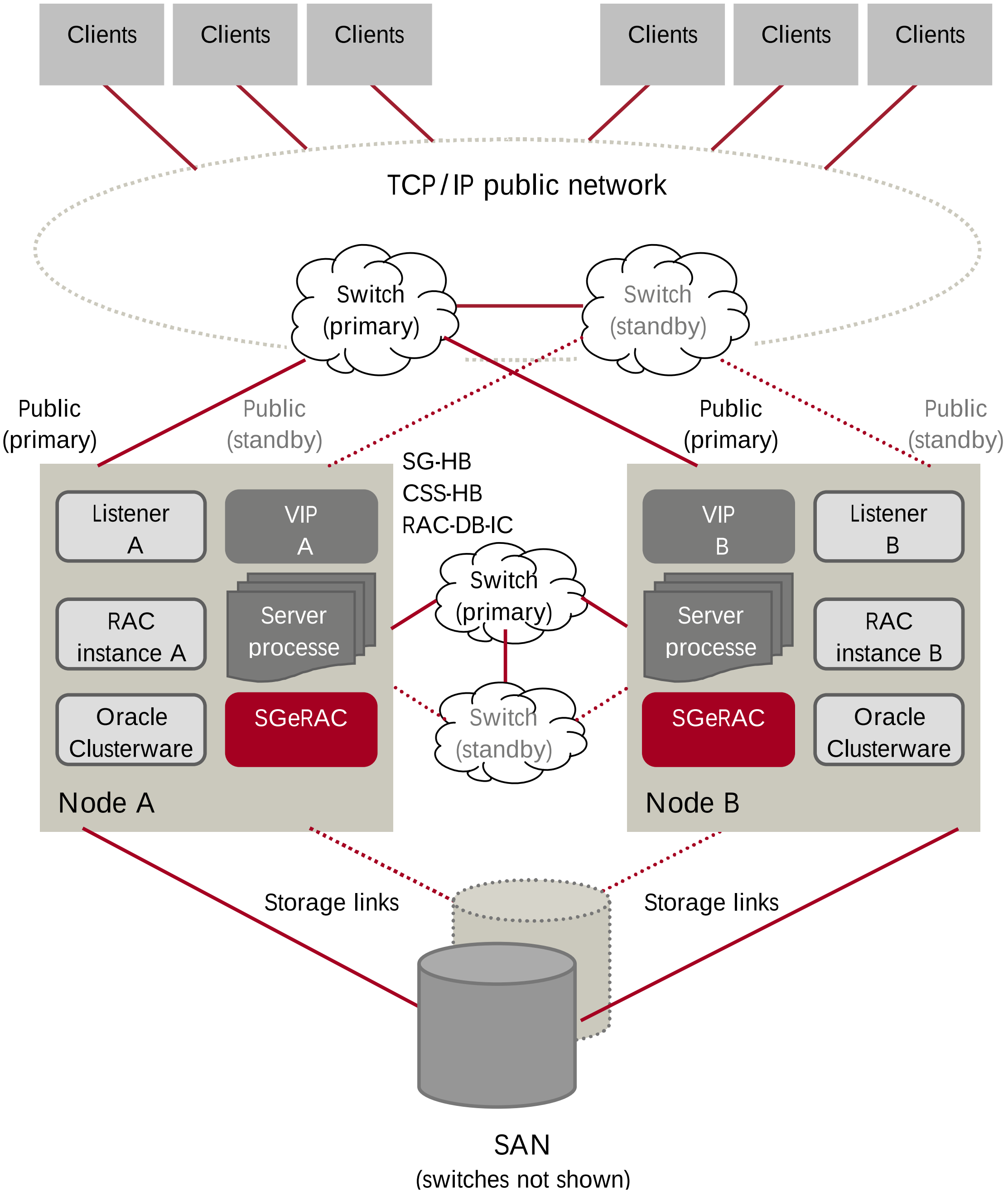
Basic configuration

Figure 1 illustrates a basic configuration that includes common high availability (HA) components in SGeRAC clusters.

There are redundant nodes in the cluster to protect against node failures. You can increase availability by configuring more nodes; the maximum number supported is documented at the HP Technical Documentation site[1]. Each node runs the same set of processes. All network and storage are protected by redundant components. SG-HB refers to Serviceguard heartbeat and cluster traffic. CSS-HB refers to Oracle Clusterware heartbeat and cluster traffic. RAC-DB-IC refers to Oracle RAC database cluster interconnect traffic.

[1] http://docs.hp.com/ à High Availability à Serviceguard Extension for Real Application Cluster à Support Matrixes (http://docs.hp.com/en/6257/SGeRAC-SLVM-CVM_Support.pdf)

Figure 1. SGeRAC and Oracle RAC basic configuration



Configurations for Oracle RAC

Public network

Client public network high availability (HA) involves two levels: redundant components and client failover.

Redundant network interfaces and switches with local LAN failover managed by Serviceguard or bonding via Auto Port Aggregation (APA), protect against single point network failures.

Client failover is needed when failures affect existing and new client sessions. These failures include node failures (e.g. power failures) and network failures (e.g. all redundant network interface or links failed). Protection is available at three levels:

1. Oracle Fast Application Notification (FAN)
2. Remote virtual IP address (VIP) fail over
3. Client connect timeout

Clients that are FAN integrated or using the FAN API may interrupt existing sessions and failover. Remote VIP failover is useful for non-FAN clients attempting to connect to the local node to avoid the TCP connect timeout. The client connect timeout is useful when client connect takes a long time for whatever reason.

VIP high availability

This section describes high availability for VIP.

Previously, Oracle virtual IP address (VIP) and Serviceguard relocatable IP address (RIP) should not exist on the same subnet on the same because of potential collisions on IP address configuration.

Note:
This issue has been addressed in Oracle 10.2.0.2 and later for the HP Integrity platform and is addressed in Oracle 10.2.0.3 and later for the HP 9000 platform.

Serviceguard local LAN failover – preferred choice

For client public network HA in a SGeRAC configuration, the preferred method for network HA is to use Serviceguard primary and standby interfaces. Serviceguard monitors the redundant network and additional APA software is not required.

When the client network is configured with Serviceguard local LAN failover, Serviceguard performs the local LAN failover and Oracle Clusterware (OC) configures the VIP after Serviceguard local LAN failover. Since OC performs monitoring and manages the VIP address, client connectivity maybe unavailable until OC detects the outage and configures the VIP address on the local node.[2]

Local LAN failover using APA

When APA is used to bond the network interface cards, APA provides traffic distribution and load balancing capability among multiple physical network interface cards (NIC) or links. Load balancing may be useful in configurations where a single interface is insufficient to handle the network traffic. When a physical NIC or link fails, APA provides HA by distributing traffic among the remaining NIC or links. One virtual link is presented to OC and APA network load balancing is transparent to OC. APA requires that all the NICs in the link aggregate be of the same type. Since APA network connections go to the same switch, a switch failure causes outage of the client network.

When APA/Hot Standby is used, APA/Hot Standby provides the primary-to-hot-standby failover by rerouting traffic from failed primary link to hot standby link. APA/Hot Standby does not load balance. Serviceguard does not monitor this network. One virtual network link is presented to OC and the physical failover is transparent to OC since the same virtual network link remains available. Both NICs must be the same type as with other types of APA.

[2] See Doc ID: Note:296874.1 Configuring the HP-UX Operating System for the Oracle 10g VIP at https://metalink.oracle.com/ (Oracle MetaLink account required)

Remote failover

In the case of a catastrophic failure such as node failure or network failure, OC fails over the VIP address to a surviving node.

Network for cluster communication

The network for cluster communication is used for private internal cluster communications. Although one network adapter per node is sufficient for the private network, HP recommends that a minimum of two network adapters for each network to be used for higher availability.

Serviceguard, OC, and each RAC instance maintain communication with peers on other nodes. When communication is broken, either through network partition or node failure, each of these components needs to reform its membership and eject non-members as needed.

The categories of traffic between nodes are distinguished as follows:

- SG-HB – Serviceguard heartbeat and communications traffic. Supported over single or multiple subnet networks.
- CSS-HB – Cluster Synchronization Service (CSS) heartbeat traffic and communications traffic for Oracle Clusterware. CSS-HB uses a single logical connection over a single subnet network.
- RAC-DB-IC – RAC instance peer to peer traffic and communications for Global Cache Service (GCS) and Global Enqueue Service (GES), formerly Cache Fusion (CF) and Distributed Lock Manager (DLM). Per RAC database. Network HA is provided by the HP-UX 11i platform (Serviceguard or APA bonding).
- ASM-IC – Applicable only when using Automatic Storage Management (ASM). ASM instance peer to peer traffic. When it exists, ASM-IC should be on the same network as CSS-HB. Network HA is required either through Serviceguard failover or APA bonding.
- GAB/LLT – Applicable only when using CFS/CVM, Cluster File System and Cluster Volume Manager. Symantec cluster heartbeat and communications traffic. GAB/LLT communicates over link level protocol (DLPI) and supported over Serviceguard heartbeat subnet networks, including primary and standby links. GAB/LLT is not supported over APA or virtual LANs (VLAN).

Note that each category maintains its own timeout for which members may be evicted from its respective membership.

The interconnect network requires HA configurations. When a single network failure occurs, for example LAN card or switch failures, all the cluster nodes continue to operate. Without HA, a single network failure results in a network partition between the nodes and evicted nodes are halted.

Using Serviceguard primary and standby links is the preferred HA model to provide HA for the cluster communications interconnect network HA. With redundancy through Serviceguard primary and standby, Serviceguard monitors the network and performs local failover if the primary network becomes unavailable.

General principles

It is preferred to have all interconnect traffic for cluster communications to go on a single heartbeat network that is redundant so that Serviceguard will monitor the network and resolve interconnect failures by cluster reconfiguration. This preferred configuration is the recommended common configuration.

The following examples are instances when it is not possible to place all interconnect traffic on the same network:

- RAC GCS (cache fusion) traffic may be very high, so a separate network for RAC-DB-IC may be needed.[3] One RAC-DB-IC may interfere with another RAC-DB-IC on the same cluster. RAC-DB-IC may also interfere with heartbeat traffic.
- Some networks are not supported by CFS/CVM, so the RAC-DB-IC traffic may be on a separate network.
- Certain kinds of fast re-configurations require a minimum of two Serviceguard heartbeat networks, whereas CSS-HB and RAC-DB-IC do not support multiple network for HA purposes.

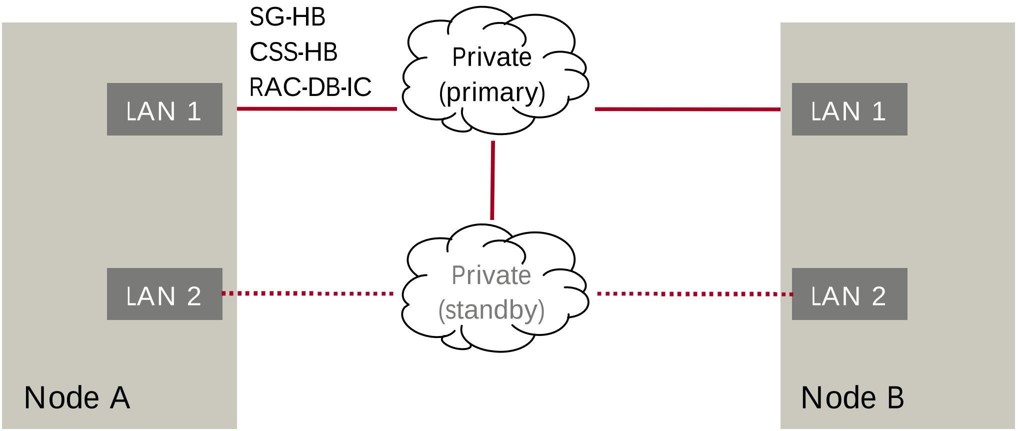
Note:
Starting with A.11.19, the faster failover capability is integrated as part of the base Serviceguard product.

In these cases, you may see longer recovery times for some kinds of network failure (other than those managed by Serviceguard's primary-standby mechanism), unless you develop special logic to handle them."

Common configuration

When a single network is sufficient to address the internal interconnect bandwidth requirements, the recommended choice is to use a single network for SG-HB, CSS-HB, RAC-DB-IC, for ASM-IC when ASM is used and GAB/LLT when CFS/CVM is used.

Figure 2. Common configuration



This configuration is the most common configuration. There is one network and the network has sufficient bandwidth. If there are multiple databases, then they will also use the same interconnect.

The primary-and standby-pair protects against single failure. Serviceguard monitors the network and performs local LAN failover if the primary fails. The local LAN failover is transparent to CSS and RAC. When ASM is used, ASM-IC traffic is on the same network as CSS-HB. When CFS/CVM is used, GAB/LLT traffic is on the same network as SG-HB. Here, both CSS-HB and SG-HB are on the same network.

[3] See CLUSTER_INTERCONNECTS, page 5-11, Oracle Clusterware and Oracle Real Application Clusters Administration and Deployment Guide version 10g Release 2 (10.2) (http://download-west.oracle.com/docs/cd/B19306_01/rac.102/b14197.pdf); See also Administering Multiple Cluster Interconnects on Linux and UNIX Platforms, page 3-16, Oracle Real Application Clusters Administration and Deployment Guide 11g Release 1 (11.1) B28254-04, November 2007 (http://download.oracle.com/docs/cd/B28359_01/rac.111/b28254.pdf)

When both the primary and standby interfaces fail, Serviceguard resolves the interconnect failure by performing a cluster reconfiguration. After Serviceguard completes its reconfiguration, SGeRAC notifies CSS and CSS updates RAC.

Serviceguard allows addition and deletion of subnets. When subnets are changed, be sure to keep the SG-HB on the same network as CSS-HB and RAC-DB-IC to maintain the monitoring behavior.

Timeouts

The SG-HB timeout (MEMBER_TIMEOUT)[4] should be set based on service availability requirements. The usual determining factor is how soon the service must be available when a node failure or a complete heartbeat network failure has occurred. On installation, CSS-HB and RAC-DB-IC use default timeouts of 10 minutes and 17 minutes respectively. ASM-IC default timeout is predefined on installation and should not be changed. GAB/LLT timeout is set automatically at CVM/CFS configuration time to coordinate with Serviceguard MEMBER_TIMEOUT and should not require manual change.

Serviceguard MEMBER_TIMEOUT is in the Serviceguard cluster configuration file. The CSS-HB timeout is the CSS MISSCOUNT. The RAC-DB-IC timeout is the timeout for Instance Membership Recovery (IMR) and is specified by the Oracle parameter `_dlm_send_timeout.`

Note:
The `_dlm_send_timeout` parameter is unavailable in 11g.

Since Serviceguard resolves the interconnect failure, the CSS-HB timeout should be greater than the Serviceguard reconfiguration time. The default values for CSS-HB timeout and RAC-DB-IC timeout should be sufficient. If there is a need to tune any timeouts, the CSS-HB timeout should be tuned to provide an opportunity for Serviceguard to complete reconfiguration and update CSS through group membership service (GMS) prior to CSS timeout. The RAC-DB-IC timeout should be 15 seconds above CSS-HB timeout. For maximum availability, tune Serviceguard MEMBER_TIMEOUT, CSS-HB timeout, and RAC-DB-IC timeout together.

Alternate configuration – multiple RAC databases

When RAC-DB-IC traffic is very high, there is a possibility that it may interfere with other traffic.

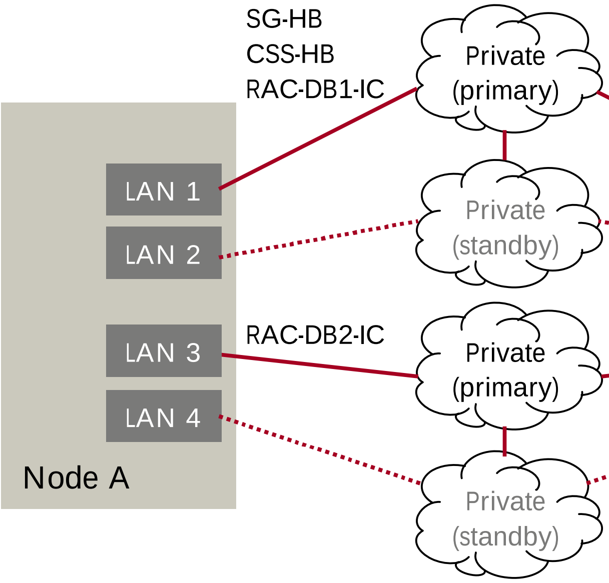
When there are multiple independent RAC databases in the same cluster, if there is insufficient bandwidth over a single network, a second network can be used for other database interconnect traffic.[5] If ASM is used, ASM-IC traffic will be on the same network as CSS-HB (LAN1/lan2). If CFS/CVM is used, GAB/LLT traffic will be one the same network as SG-HB (LAN1/lan2).

Each primary and standby pair protects against single failure. If the subnet with SG-HB (LAN1/lan2) fails, Serviceguard will resolve the subnet interconnect failure with a Serviceguard cluster reconfiguration. If the subnet with RAC-DB2-IC (LAN3/4) fails, unless subnet monitoring is used, IMR will resolve the subnet interconnect failure.

[4] Serviceguard A.11.19 introduced includes a new cluster manager protocol and a new parameter MEMBER_TIMEOUT to specify heartbeat timeout.

[5] See Oracle RAC Administration documentation on how to specify additional RAC-DB-IC.

Figure 3. Alternative configuration – multiple RAC databases



Timeouts

The SG-HB timeout (MEMBER_TIMEOUT) should be set based on service availability requirements. Optionally if CSS-HB is changed, the CSS-HB timeout should be tuned to provide an opportunity for Serviceguard to complete reconfiguration and update CSS through group membership service (GMS) prior to CSS timeout. Optionally if RAC-DB-IC timeout is changed, RAC-DB-IC timeout should be 15 seconds above CSS-HB timeout. Note that the timeout relations are slightly different when using Cluster Interconnect Subnet monitoring.

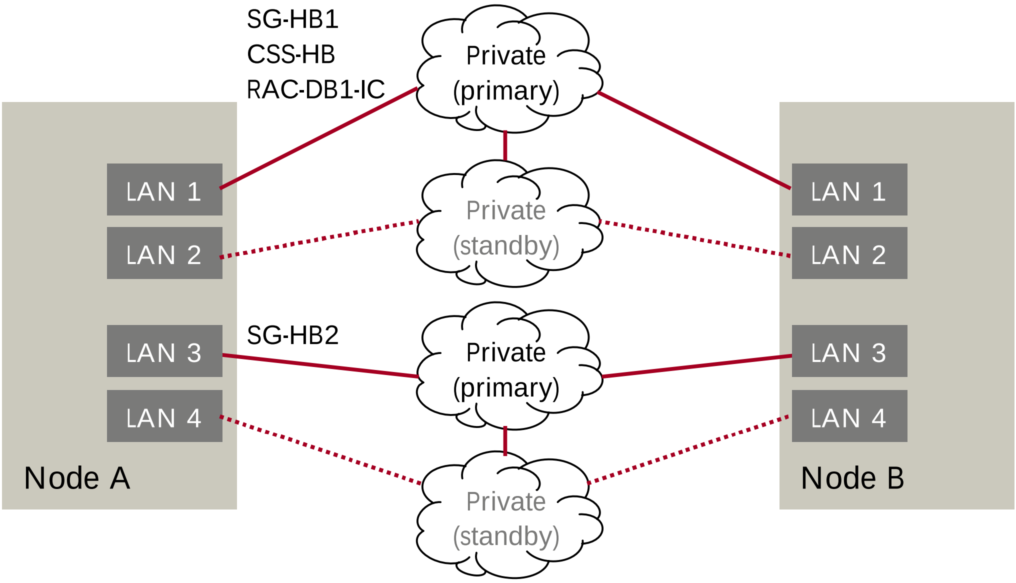
Subnet monitoring or Cluster Interconnect Subnet monitoring

Serviceguard packages can be configured with a subnet dependency on the RAC-DB2-IC. If both LAN3 and LAN4 fail, the Serviceguard package can request halting the RAC instance where the interconnect failure is detected. Serviceguard subnet monitoring has a limitation in that if all interconnects fail (e.g. primary and standby switches failed at the same time), all instances are halted. If there is a concern with simultaneous failure of both switches, Serviceguard supports multiple standbys, so you can add additional standby switches if you are concerned about simultaneous failure of both switches.

Without subnet monitoring, double networks failures on the RAC-DB2-IC network will be discovered by IMR timeout and IMR determines which instance to evict.

Starting with A.11.18, SGeRAC introduced a new feature called Cluster Interconnect Subnet (CIS) monitoring to address the limitation with subnet monitoring. The feature adds a new parameter CLUSTER_INTERCONNECT_SUBNET for use with multi-node packages (MNP). When RAC instances are configured with MNP packages and the RAC subnets are monitored using CLUSTER_INTERCONNECT_SUBNET, if the monitored subnet fails, at least one MNP package will remain.

Figure 4. Alternative configuration – low Serviceguard member timeout



When RAC-DB-IC traffic is very high and SG-HB timeout is low, there is a probability of RAC-DB-IC traffic interfering with SG-HB traffic and causing unnecessary timeouts. If SG-HB timeout can not be increased, then an alternative action is to use a second network for SG-HB. This configuration is for environments that need fast failover (low Serviceguard member timeout) for two or more nodes.

Each primary and standby pair protects against single failure. With SG-HB on more than one subnet, a single subnet failure will not trigger a SG reconfiguration. If the subnet with CSS-HB fails, unless subnet monitoring is used, CSS will resolve the subnet interconnect failure with a CSS cluster reconfiguration.

Note:
Starting with Serviceguard A.11.19, the faster failover capability is integrated with the base Serviceguard product. This configuration can be used for faster failover.

Timeouts

The SG-HB timeout (MEMBER_TIMEOUT) should be set based on service availability requirements. Optionally if CSS-HB timeout is changed, the CSS-HB timeout should be tuned to provide an opportunity for Serviceguard to complete reconfiguration and update CSS through group membership service (GMS) prior to CSS timeout. Optionally if RAC-DB-IC timeout is changed, RAC-DB-IC timeout should be 15 seconds above CSS-HB timeout. Note that the timeout relations are slightly different when using Cluster Interconnect Subnet monitoring.

Subnet monitoring or Cluster Interconnect Subnet monitoring

Serviceguard packages can be configured with a subnet dependency on the CSS-HB subnet and FAIL_FAST enabled. If both LAN1 and lan2 failed, the Serviceguard package can request halting the node where the interconnect failure is detected. Use of Serviceguard subnet monitoring has a limitation where if all interconnect fails (e.g. primary and standby switch failed at the same time), all nodes are halted. If there is a concern with simultaneous failure of both switches, Serviceguard supports multiple standby and additional standby switches may be added.

Without subnet monitoring, double networks failures on the CSS-HB network will be discovered by CSS-HB timeout and CSS determines which node to evict.

Starting with A.11.18, SGeRAC introduced Cluster Interconnect Subnet (CIS) monitoring to address the limitation with subnet monitoring, CLUSTER_INTERCONNECT_SUBNET can be used in conjunction with the node fail fast enabled option for monitoring the CSS-HB network. When CSS is configured with MNP packages and CSS-HB subnet is monitored using CLUSTER_INTERCONNECT_SUBNET, if the monitored subnet fails, the MNP package will halt the nodes where the monitored subnet has failed, and at least one MNP package and node will remain.

Note:
For 11gR1, when using CIS for CSS-HB in configurations of three or more nodes on three or more nodes configurations and nodes are halted because of a fail fast from CIS, new connections on the RAC instance be may delayed up to ten minutes.

When CIS monitoring is used on the CSS-HB network, the installed default values for various timeouts should be sufficient for most installations. If you need to change the Serviceguard member timeout, Serviceguard heartbeat interval or CSS-HB timeout, the CSS-HB timeout should be tuned to provide an opportunity for Serviceguard to complete reconfiguration and update CSS through group membership service (GMS) prior to CSS timeout.[6] Optionally if RAC-DB-IC timeout is changed, RAC-DB-IC timeout should be 15 seconds above CSS-HB timeout.

Guidelines for changing cluster parameters

These are general guidelines for changing cluster parameters for timeouts depending on whether Cluster Interconnect Subnet monitoring is used.

When Cluster Interconnect Subnet monitoring is used to monitor the CSS-HB network and if any of the following cluster parameters needs to be changed the default values of:

- Oracle Clusterware parameter CSS MISSCOUNT
- Serviceguard cluster configuration parameter MEMBER_TIMEOUT

Then the CSS MISSCOUNT parameter should be greater than:

- For SLVM: (number of nodes – 1) times (F + SLVM timeout) + 15 seconds
- For CVM/CFS: (two times number of nodes – 1) times F + 15 seconds
- When both SLVM and CVM/CFS are used, then take the max of the above two calculations

Note 1:
F is the Serviceguard failover time as given by the `max_reformation_duration` field in the ouptput of `cmviewcl –v –f line` output.

[6] The relation when using CIS is different than without using CIS is due to the need for CSS-HB timeout to take into consideration scenarios where multiple nodes may fail fast sequentially to arrive to at least one MNP instance remains.

Note 2:
SLVM timeout would be documented in the LVM Whitepaper entitled LVM Link and Node Failure Recovery Time found on the HP Technical Documentation site at docs.hp.com.

When Cluster Interconnect Subnet monitoring is not used to monitor the CSS-HB subnet and if the default values of any of the following parameters has to be changed:

- Oracle Clusterware parameter CSS MISSCOUNT
- Serviceguard cluster configuration parameter MEMBER_TIMEOUT

Then the CSS MISSCOUNT parameter should be greater than:

- For SLVM: F + SLVM timeout + 15 seconds
- For CVM/CFS: 3 times F + 15 seconds
- When both SLVM and CVM/CFS are used, then take the max of the above two calculations

Storage

SGeRAC supports a variety of storage options for Oracle RAC.

- CFS
- CVM
- SLVM
- ASM over Raw Devices
- ASM over SLVM

HP recommends you organize storage as follows, with each item using separate storage:

1. Oracle Clusterware binaries
2. RAC database binaries
3. Oracle Clusterware registry (OCR) and quorum device (Voting Disk)
4. Database data files
5. Database recovery data (archive, flash recovery)

Keeping Oracle Clusterware home and Oracle database home on separate mount points allows you flexibility to unmount the database home for maintenance without having to shutdown the Oracle Clusterware cluster. Placing the registry and quorum device on separate storage ensures that deactivation of storage for RAC database does not affect Oracle Clusterware. Finally, each RAC database uses two separate storage areas for database data files and recovery data. This is to ensure recovery data is available in case there are problems with storage for the database data files.

Oracle Clusterware may need timely access to its home directory and binaries. The default I/O timeout for LVM logical volumes is forever; therefore, for file systems on top of LVM, it may be necessary to configure a logical volume timeout, for example 30 seconds per link, so that Oracle Clusterware gets get an I/O error when the LVM physical volume is inaccessible. There are similar considerations for RAC binaries. For VxVM, there is already a default I/O timeout of 30 seconds per link. There are similar considerations for RAC binaries as well as Oracle Clusterware binaries with regards to timely access.

For Oracle Clusterware data files, starting with 10.2.0.2, Oracle introduced a CSS disktimeout to capture hang I/O hangs; however, for configurations with SGeRAC, the disktimeout is automatically set to three seconds lower than CSS MISSCOUNT and is only indirectly configurable based on CSS MISSCOUNT. Setting I/O timeouts for SLVM or CVM does not affect disktimeout. If CSS receives I/O error, CSS attempts to reopen Oracle Clusterware data files until disktimeout is reached. The default MISSCOUNT is 600 seconds for configurations with SGeRAC.

Table 1 shows a quick comparison between the various storage options and the details of each option are discussed.

Table 1. Storage comparison

	Advantages	Disadvantages	I/O Timeout
CFS	Cluster file system, online changes	Requires Serviceguard Storage Management Suite	30 seconds per link
CVM	Online changes	Archive package failover	30 seconds per link
SLVM	Established non-intrusive solution, online node and volume group changes	Online SLVM volume reconfigurations available on one node, archive package failover	Default is no timeout
ASM over Raw Devices	Zero impact on database recovery operations	Requires file system for Oracle binaries, alert log, OC data files, application binaries and data	Default timeout based on version of HP-UX 11i
ASM over SLVM	Zero impact on database recovery operations	Online SLVM volume reconfigurations available on one node	For OC data files, default is no timeout. For RAC data files, default is set to 30 seconds per link

CFS

Serviceguard supports CFS using two additional new features: simple package dependency and multi-node packages. CFS provides the following notable features:

- Single home for Oracle Clusterware
- Single ORACLE database home
- Single archive designation

The following is a sample mapping based on the preferred storage configuration:

- Disk Group (DG)/CFS for Oracle Clusterware home when single home for ease of management is important; or use local file systems for Oracle Clusterware home when minimal interference between nodes is desired.
- DG/CFS for Oracle database home when single home for ease of management is important; or use local file systems for Oracle database home when minimal interference between nodes is desired.
- DG/CFS for OCR and Voting Disk
- DG/CFS for Oracle RAC database data files
  - o Recommendation: one DG per database
- DG/CFS for flash recovery and/or archive destination
  - o Recommendation: one DG per database

Note that if you use CFS for Oracle data files, the Oracle Disk Manager (ODM) library from Symantec must be enabled to get near raw I/O performance.

When Oracle Clusterware home and/or Oracle database home is on CFS, a CFS file system full problem may negatively impact Oracle Clusterware and/or Oracle database home on all nodes, e.g. unable to log or dump. If this is a concern, another option is to place both Oracle Clusterware home and Oracle database home on local file systems to localize the effect of file system full.

Figure 7. CFS configuration

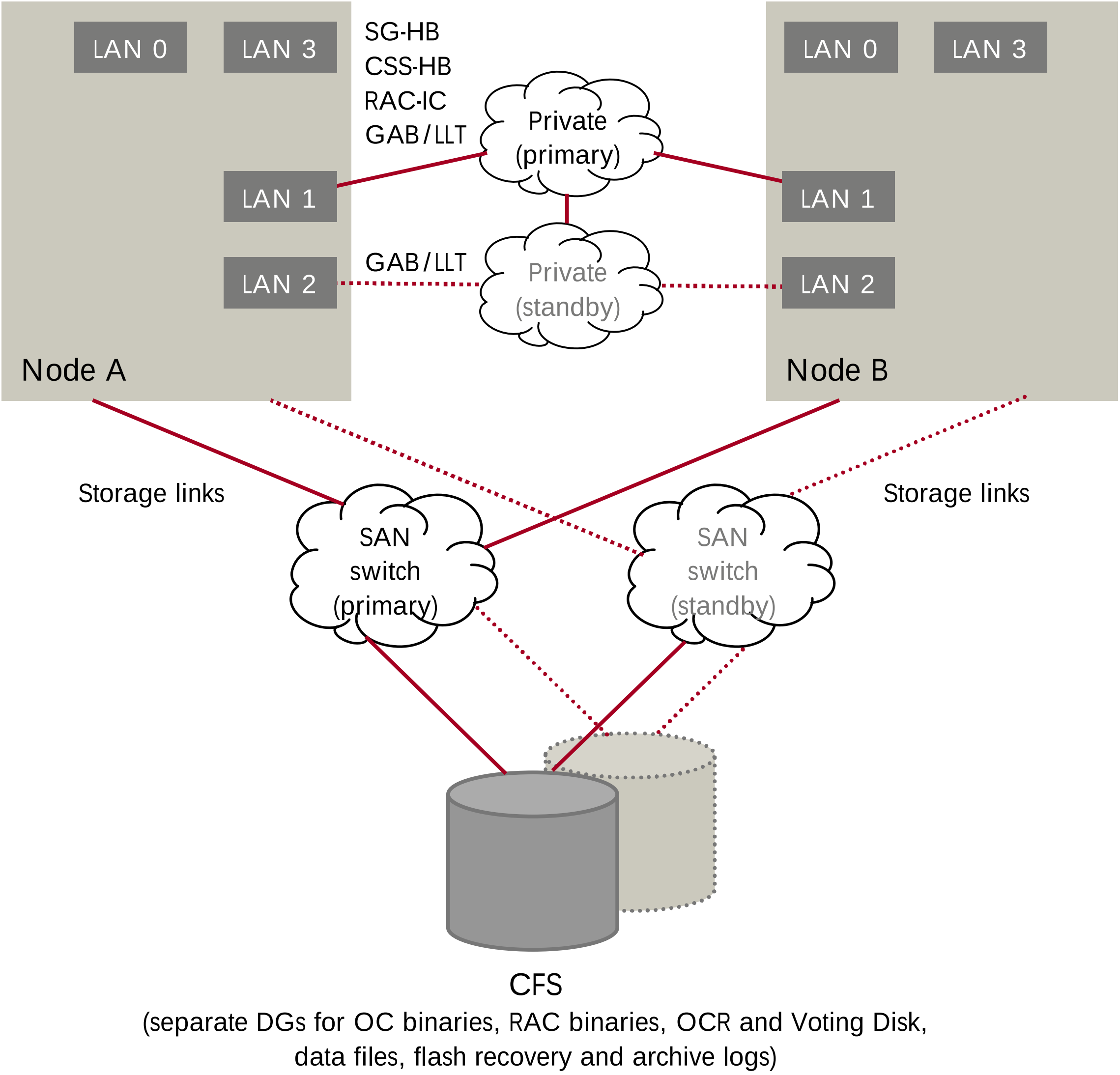


Figure 7 shows a configuration with all storage on CFS. In a Serviceguard CFS environment, there is an additional set of network traffic required by CFS/CVM (GAB/LLT) which uses both the primary and standby paths. As part of the Serviceguard CFS system multi-node package configuration, GAB/LLT is configured to run on the same set of interfaces as SG-HB, including both primary and standby paths, concurrently.

For online changes, CFS/CVM supports online reconfiguration of disk groups, volumes, file system, and nodes.

For shared storage access failure, the CVM I/O timeout default value is 30 seconds for each link.

CVM

The preferred storage configuration with CVM is as follows:

- DG/local file system for Oracle Clusterware home
- DG/local file system for Oracle home
- CVM DG for OCR and Voting Disk
- CVM DG for Oracle RAC database data files
  - Recommendation: one DG per database
- DG/local FS for flash recovery and/or archive destination
  - Recommendation: one DG per database
  - Can be placed in shared storage with failover package so that node performing Oracle recovery can have access to all archive logs

For online changes, CVM supports online reconfiguration of disk groups, volumes, and nodes.

SLVM

The preferred storage configuration for SLVM is as follows:

- VG/local FS for Oracle Clusterware home
- VG/local FS for Oracle home
- SLVM VG for OCR and Voting Disk
- SLVM VG for Oracle RAC database data files
  - Recommendation: one VG per database
- VG/local FS for flash recovery and/or archive destination
  - Recommendation: one VG per database
  - Can be placed in shared storage failover package so that node performing Oracle recovery can have access to all archive logs

For online changes, SLVM supports online volume reconfiguration through SNOR (single node online reconfiguration), online addition of new volume groups, and online node additions.

Changes through SNOR require deactivation of the SLVM volume group on all but one node. Therefore, applications using the SLVM volume group must halt on all but one node.

ASM over raw devices

SGeRAC supports ASM over raw devices. The ASM disk groups members can be shared raw devices.

- ASM supports
  - Data files, control files, online and archive redo log files, and backup files.
- ASM does not support
  - Oracle binaries, trace files, audit files, alert logs, export files, tar files, core files
  - Oracle cluster registry devices (OCR) and quorum device (Voting Disk)
  - Application binaries and data

The preferred storage configuration, with a minimum of two ASM disk groups per database, is as follows:

- VG/local FS for Oracle Clusterware home
- VG/local FS for Oracle home
- Raw devices for OCR and Voting Disk (can also use SLVM VG)
- For each RAC database
  - Two ASM disk groups
  - One for database data and the other for flash recovery

ASM over SLVM

SGeRAC supports ASM over SLVM. The ASM disk groups members must be raw logical volumes (LV) managed by SLVM.

- ASM supports
  - Data files, control files, online and archive redo log files, and backup files.
- ASM does not support
  - Oracle binaries, trace files, audit files, alert logs, export files, tar files, core files
  - Oracle cluster registry devices (OCR) and quorum device (Voting Disk)
  - Application binaries and data

The preferred storage configuration, with a minimum of two ASM disk groups per database, is as follows:

- VG/local FS for Oracle Clusterware home
- VG/local FS for Oracle home
- SLVM VG for OCR and Voting Disk
- For each RAC database
  - Two ASM disk groups, each residing in its own SLVM VG
  - One for database data and the other for flash recovery

Online changes support the same as SLVM. Therefore, changes through SNOR require deactivation and halting of applications using the shared volume group on all but one node. An I/O timeout for the shared logical volumes must be configured.

Serviceguard packages

In the combined software stack, SGeRAC provides the following to Oracle Clusterware (OC) and RAC:

- Cluster membership to CSS
- Clustered storage to meet the needs of OC and RAC database instance

Oracle Clusterware manages the following:

- Database and associated resources (database instance, services, VIP, listener, etc…)
- ASM instance, when configured.

Storage required by OC must be available before OC starts, and similarly, storage required by RAC database must be available before the RAC instance starts. There is a need to coordinate to ensure the combined stack starts up and shuts down in the proper sequence, and a need to automate startup and shutdown sequences, if desired.

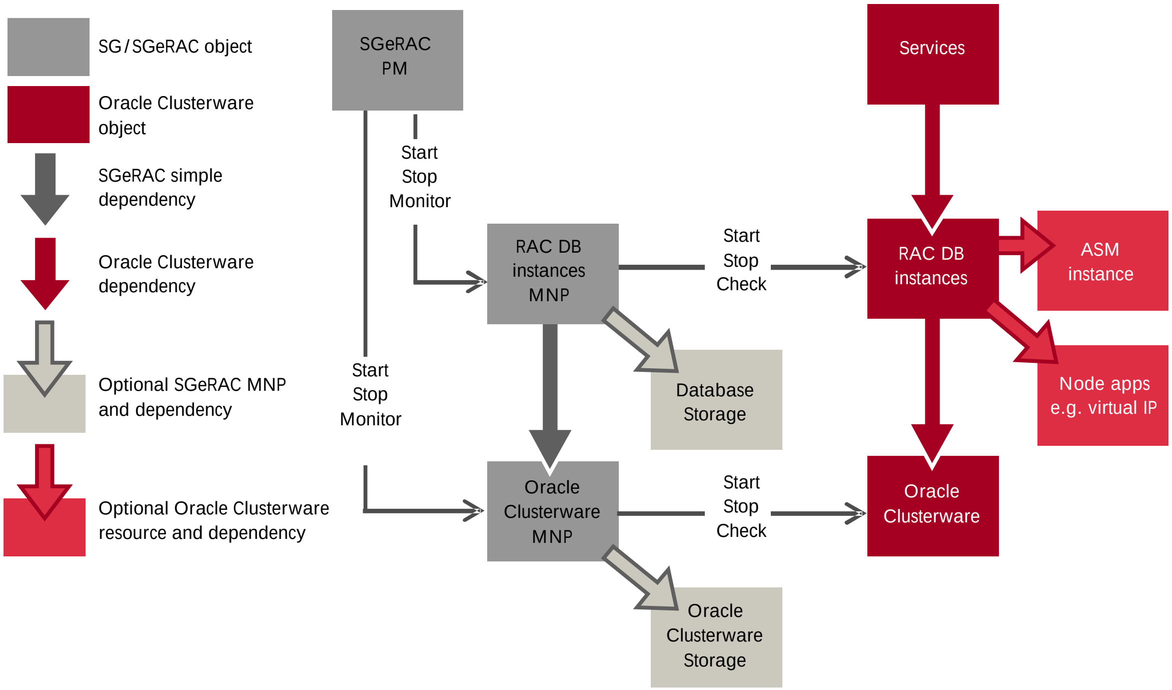
HP recommends that you use Serviceguard packages to coordinate between SG/SGeRAC and OC/RAC. Standard packages are used to encapsulate the start and stop of storage in addition to the applications. Starting with A.11.17, the support of CFS uses multi-node packages (MNP) with simple package dependencies.

Serviceguard Extension for RAC Toolkit

For SGeRAC A.11.17 or later, consider using the SGeRAC Toolkit[7]. This Toolkit is based on the contributed Package Integration Framework. This Toolkit provides a uniform, easy to manage and intuitive method to coordinate the operation of the combined RAC and SGeRAC software stack across the full range of storage management options supported by SGeRAC.

Figure 8 depicts an operation overview of the SGeRAC Toolkit.

Figure 8. Serviceguard Extension for RAC Toolkit



Oracle Clusterware is configured as one MNP. Each RAC database is configured as one MNP. All RAC MNPs are configured to depend on the Oracle Clusterware MNP.

The Oracle Clusterware MNP is configured to start and stop Oracle Clusterware. For SLVM, CVM, and ASM over SLVM, the storage is started by the MNP. For CFS, the Oracle Clusterware MNP is

[7] The SGeRAC Toolkit is available as a free download from http://software.hp.com, select High availability, select Serviceguard Extension for RAC Toolkit. Starting A.11.18, SGeRAC Toolkit is bundled in SGeRAC product.

configured with a dependency on the CFS MNP. The Oracle Clusterware MNP ensures the required storage is available before signaling the start of Oracle Clusterware.

Each RAC database MNP ensures the storage for the database is started before starting the database instance.

The init process monitors Oracle Clusterware. Oracle Clusterware monitors the database instance. The MNP monitoring is to the extent that the MNP status reflects the status of what it has started. The MNP does not restart the application. For instance, if Oracle Clusterware is halted outside the MNP, the MNP will discover that Oracle Clusterware is no longer running and the MNP is halted.

Standard packages

Please note that synchronization using standard Serviceguard packages is also supported.

Extended Distance Cluster for RAC

SGeRAC supports Extended Distance Cluster for RAC (EDC) with CFS, CVM, ASM, and SLVM. The standard basic configuration[8] consists of two datacenters where the servers and storage reside, and a third site where the quorum server resides. The quorum server is required as a tie-breaker in the event of loss of communications between the datacenters.

OCR and Voting Disk

Oracle 10.2.0.1 and later supports up to two copies of the OCR and up to three vote disks. For Extended Distance Clusters (EDC), each copy of OCR and each vote disk are required to be physically mirrored by software (CFS, CVM, SLVM) between the two data centers.

In the event of a loss of communication between the datacenters, the quorum server serves as a tie-breaker for Serviceguard. The mirrored OCR and Voting Disks ensure that OC has access to local physical copies for OC cluster reformation.

Binaries

For EDC configurations, it is recommended to maintain local storage for OC and RAC database binaries and home to reduce inter-site traffic.

Database data files

Database data files are required to be physically mirrored to both datacenters when using CFS, CVM, ASM or SLVM.

Database recovery data

For EDC, when using SLVM or CVM, the database recovery data reside locally and needs to be made available to the RAC instance performing recovery.

The benefit of using CFS for database recovery data is the simplicity for which the recovery data is available to all instances.

Conclusion

While there are many different supported configurations and combinations of SGeRAC, Oracle Clusterware and Oracle RAC, by understanding the most common best practices, an administrator may select the best choice for optimally management and availability.

[8] See document "Understanding and Designing Serviceguard Disaster Tolerant Architectures, 1st Edition, December 2006" available at http://docs.hp.com, Select High Availability, Select Metrocluster.

Document revision history

Date	Description
March 2009	Updated for SG/SGeRAC A.11.19
March 2008	Updated document for general Oracle Clusterware and RAC usage
June 2007	Updated recommendation for Oracle Clusterware home and Oracle database home when using CFS
March 2007	Updated document with Cluster interconnect configurations, Cluster interconnect subnet monitoring, SGeRAC Toolkits, and Extended Distance Cluster
September 2006	Added multiple RAC database configurations. Updated CFS/CVM configuration

References

HP documentation

All of the following materials can be found on the HP Technical Documentation web site at http://docs.hp.com.

- HP Serviceguard Storage Management Suite
  http://docs.hp.com/en/ha.html#HP%20Serviceguard%20Storage%20Management%20Suite
- Serviceguard Version A.11.19 Release Notes
  http://docs.hp.com/en/ha.html
- Serviceguard Extension for RAC Version A.11.19 Release Notes
  http://docs.hp.com/en/ha.html
- Managing Serviceguard, Sixteenth Edition, March 2009
  http://docs.hp.com/en/ha.html
- Using Serviceguard Extension for RAC, Eighth Edition, March 2009
  http://docs.hp.com/en/ha.html
- Understanding and Designing Serviceguard Disaster Tolerant Architectures
  http://docs.hp.com/en/B7660-90018/B7660-90018.pdf
- Support of Oracle 10g RAC ASM with SGeRAC Whitepaper, January 2008
  http://docs.hp.com à High Availability à Serviceguard Extension for Real Application Cluster
- Sample Configuration with SGeRAC and Oracle RAC 11gR1 Whitepaper, March 2009
  http://docs.hp.com à High Availability à Serviceguard Extension for Real Application Cluster
- Sample Configuration with SGeRAC and Oracle RAC 10gR2 Whitepaper, March 2009
  http://docs.hp.com à High Availability à Serviceguard Extension for Real Application Cluster
- Use of Serviceguard Extension for RAC Toolkit with Oracle RAC 10g Release 2 or later Whitepaper, March 2009
  http://docs.hp.com à High Availability à Serviceguard Extension for Real Application Cluster

Oracle documentation

All of the following materials can be found on the Oracle Technical Documentation web site at http://www.oracle.com/technology/documentation/database10gr2.html and http://www.oracle.com/pls/db111/homepage.

- Oracle Clusterware Installation Guide 11g Release 1 (11.1) for HP-UX, B28259-05, November 2007
  http://download.oracle.com/docs/cd/B28359_01/install.111/b28259.pdf

- Oracle Real Application Clusters Installation Guide 11g Release 1 (11.1) for Linux and UNIX, B28264-03, November 2007
  http://download.oracle.com/docs/cd/B28359_01/install.111/b28264.pdf
- Oracle Clusterware Administration and Deployment Guide 11g Release 1 (11.1), B28255-03, September 2007
  http://download.oracle.com/docs/cd/B28359_01/rac.111/b28255.pdf
- Oracle Real Application Clusters Administration and Deployment Guide 11g Release 1 (11.1), B28254-04, November 2007
  http://download.oracle.com/docs/cd/B28359_01/rac.111/b28254.pdf
- Oracle Clusterware and Oracle Real Application Clusters Installation Guide version 10g Release 2 (10.2) for HP-UX
  http://download-west.oracle.com/docs/cd/B19306_01/install.102/b14202.pdf
- Oracle Clusterware and Oracle Real Application Clusters Administration and Deployment Guide version 10g Release 2 (10.2)
  http://download-west.oracle.com/docs/cd/B19306_01/rac.102/b14197.pdf
- Oracle Net Services Administrator's Guide
  http://download-west.oracle.com/docs/cd/B19306_01/network.102/b14212.pdf
- Client Failover Best Practices for Highly Available Oracle Database: Oracle Database 10gR2
  http://www.oracle.com/technology/deploy/availability/pdf/MAA_WP_10gR2_ClientFailoverBestPractices.pdf
- Note:296874.1 Configuring the HP-UX Operating System for the Oracle 10g VIP
  https://metalink.oracle.com (Oracle MetaLink account required)

For more information

To learn more about HP Serviceguard Solutions for HP-UX 11i, please visit:
www.hp.com/go/serviceguardsolutions

For technical documentation:
www.docs.hp.com/hpux/ha

March 2009