

Consolidating HP Serviceguard for Linux and Oracle RAC 10g Clusters



| | |
|---|---|
| Executive Summary | 2 |
| Introduction | 2 |
| Audience | 3 |
| Consolidating HP Serviceguard for Linux and Oracle RAC 10g Clusters | 3 |
| Background..... | 3 |
| Dual Cluster Co-existence..... | 3 |
| Recommended Configuration | 4 |
| Clusters with more than two nodes | 6 |
| Conclusion | 6 |
| Related Materials | 7 |

Executive Summary

Oracle Real Application Clusters (RAC) 10g Release1 database server includes a clustering solution that supports Oracle database servers in a highly available environment. Although this feature eliminates the requirement for a third-party cluster product for the database, additional components are needed to provide high availability for applications to run on the same servers that are part of the RAC cluster. HP Serviceguard for Linux can provide additional value in a RAC cluster by providing high availability for other applications running on those servers. This raises the question, “With Serviceguard for Linux running on the same hardware as the Oracle Cluster Ready Services (CRS – RAC’s cluster membership mechanism), how can the two clusters be kept from interfering with each other?”

This white paper specifies the recommended configurations that will achieve a stable Serviceguard for Linux and RAC consolidated cluster, with Serviceguard for Linux coexisting in the RAC clustered database environment. The recommendations to achieve a stable consolidated cluster setup include multiple heartbeat subnets, redundant heartbeat networks (by means of channel-bonding), and the optional usage of the Quorum Service as the arbitration method for Serviceguard for Linux.

Introduction

Many customers have built highly available Oracle database environments on Linux using the “single instance” version of Oracle and HP Serviceguard for Linux protecting both the database and application layers. With Oracle 10g Real Application Cluster, database scalability as well as high availability are built into the clustered product.

Oracle states that the Oracle 10g R1 replaces “the need to purchase, install, configure, and support third-party cluster software”. In addition, on Linux, Oracle mandates the use of CRS for RAC and specifies that other clustering software should not be used for deriving cluster membership For RAC. This capability does not provide high availability for other applications. As a result, when there is a requirement to provide high availability for other applications in an Oracle RAC environment, customers must use other high availability cluster software. Before the evaluation and recommendations documented by this white paper, HP required the applications and database be deployed on separate clusters. As a result of this evaluation, HP enables customers to lower their Oracle RAC environment costs by providing high availability encapsulation with HP Serviceguard for Linux for non-RAC applications running on the same nodes of an Oracle RAC cluster. Even if Oracle eventually provides failover functionality for third party applications in a RAC environment, the use of Serviceguard for Linux instead of Oracle RAC cluster is attractive because of its proven robustness and stability.

This whitepaper provides a set of guidelines for consolidating Serviceguard for Linux with Oracle 10g RAC clusters on the same server nodes, so as to eliminate any foreseeable interference between the two clusters in the event of a failure that causes cluster reconfigurations. The guidelines are based on an investigation where Oracle 10g RAC and Serviceguard for Linux coexistence was tested for possible failure scenarios.

In an Oracle 10g R1 RAC on Linux environment, only configurations leveraging Serviceguard for Linux as the failover cluster solution for applications other than the Oracle database will be supported by HP.

Audience

This document is for users of Serviceguard on Linux who are interested in providing high availability encapsulation to applications which would run on the same servers as an Oracle 10g R1 RAC cluster. It documents the guidelines to implement a stable dual-cluster, with Serviceguard for Linux and Oracle 10g R1 RAC database.

It is assumed that the reader has a general understanding of Serviceguard for Linux and Oracle 10g R1 RAC cluster features. Please see www.hp.com/go/sglx and www.hp.com/solutions/highavailability/oracle for detailed information on each solution.

Consolidating HP Serviceguard for Linux and Oracle RAC 10g Clusters

Background

On Linux, Oracle mandates that CRS be used to derive membership for the RAC cluster and disallows the use of any other clustering software for this purpose. This is different than HP-UX. For Oracle RAC on HP-UX, a special version of Serviceguard, Serviceguard Extension for RAC (SGeRAC), provides membership control for both clusters.

However Oracle does not provide any high availability features with RAC for other applications that customers may run on the same nodes of the cluster. Serviceguard for Linux, running on the same cluster hardware as Oracle RAC, can provide the high availability encapsulation for such applications. The two clusters run independently of one another, sharing no data or control. To maintain that independence, any Oracle RAC data on shared storage should not be part of any Serviceguard package.

Independently running two cluster software products, on the same set of nodes and hardware resources poses a unique set of challenges, most important of these being the ability to reduce conflicts between the two.

Dual Cluster Co-existence

The function of a high availability cluster is to keep applications available. In order to provide that capability the cluster software needs to be able to monitor the state of applications and also the state of the various nodes in a cluster. In the case of Serviceguard for Linux and Oracle RAC co-existence, there is no “overlap” in monitoring applications, since Oracle monitors only the RAC application and Serviceguard monitors any other application. Both Serviceguard for Linux and RAC are monitoring the same nodes using

heartbeat mechanisms. Heartbeats are network messages that are sent between nodes in a cluster letting each node know that the others are “alive”.

Both Serviceguard for Linux and Oracle RAC recognize node failures by the loss of heartbeats and act to resolve it by rebooting the affected node. Most failures where there is a loss of heartbeat are due to a node failure. This could be due to a hardware problem or an OS crash. In these instances, both clusters readily recognize (and implicitly agree on) which node has failed, and they will adjust their membership accordingly.

There is one failure type that, without special considerations, might cause the entire cluster to fail when two clusters are running on the same nodes. If a Serviceguard for Linux and RAC two-node cluster were configured with a single, shared heartbeat network, then the failure of that network would result in isolation of the two nodes. Both clusters have quorum mechanisms that define how the cluster determines which node should keep running and which should be reset. However, since each cluster software has different and independent algorithms, they may choose different nodes. For example, HP Serviceguard for Linux may choose to retain node A and reboot the other node (node B); while RAC may choose to retain node B and reboot node A. Serviceguard for Linux will reboot node B and nearly simultaneously RAC will reboot node A, resulting in both nodes (and both clusters) becoming unavailable.

This possibility can be eliminated for all but the most extreme failure scenarios by having redundant and/or multiple heartbeat networks. In this case, ALL of the networks carrying heartbeats would have to fail to have the condition where the clusters could attempt to reboot different nodes causing both clusters to fail – a multiple failure within the clusters.

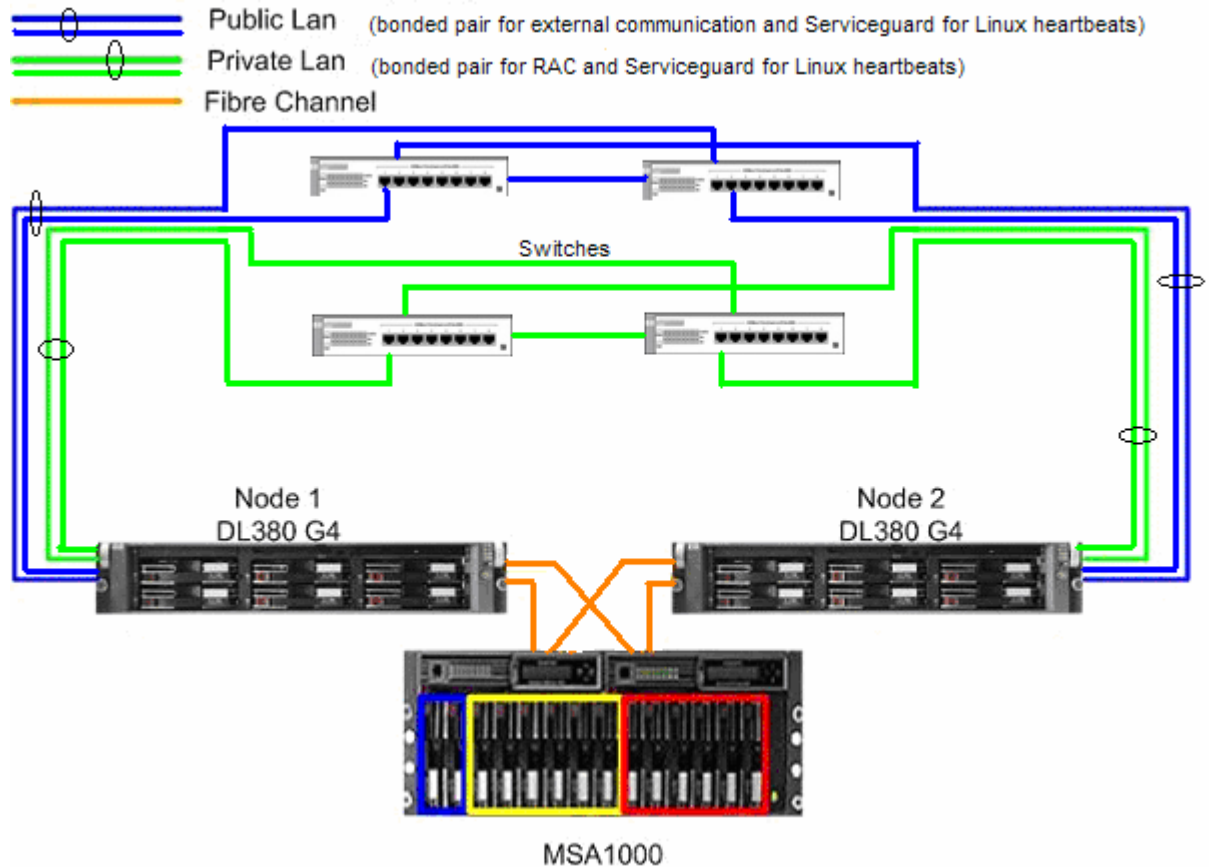
The recommended configuration (below) has 4 network paths capable of carrying heartbeats. If a user is concerned about all of these failing simultaneously, then the Serviceguard Quorum Service can be used. The Serviceguard Quorum Service runs on a system outside of the cluster such as another server or a PC. Experiments have shown that, when using the Quorum Service, Serviceguard takes down a node significantly sooner than Oracle would.¹ The Oracle membership software recognizes this and does not take down the remaining node. It should be pointed out that the Quorum Service is connected to the cluster via a network connection. If the failure case of all network paths between nodes failing also causes all paths to the Quorum Service to fail, then there is again the possibility of the entire cluster going down. It is impossible to protect against all multiple failures.

Recommended Configuration

The general recommended network configuration for Serviceguard calls for a dedicated heartbeat LAN and bonded pair of NICs (using Linux Channel Bonding) for both application traffic and Serviceguard heartbeats. Serviceguard and Oracle can share the dedicated heartbeat LANs since Serviceguard’s heartbeats are once per second and will have no measurable effect on RAC.

¹ When using the default timeout and heartbeat intervals, Oracle RAC and Serviceguard will both reset a node after about 60 seconds.

Figure 1 is an example of the network configuration for a two node consolidated cluster. It has two bonded network pairs. One dedicated to the HP Serviceguard for Linux and Oracle heartbeats. The other for communication external to the cluster and also for HP Serviceguard for Linux heartbeats. This is the minimum supported. Dual port NICs can be used, but for availability, bonded pairs should NOT use the connections from a single dual port NIC. More NICs can be added at the users' option. If possible, these should be made highly available via bonding, and also, if possible, should be configured as a Serviceguard for Linux heartbeat.



The focus of the recommendations has been to eliminate the possibility of the both clusters going down through the use of multiple heartbeat paths. If the Quorum Service is used as a Serviceguard arbitration mechanism to provide protection against loss of all heartbeats, then it is required to keep the Oracle and HP Serviceguard for Linux installation with default cluster timing settings ("pure installation"). This is necessary because there are certain timing parameters within the two clusters that cause Serviceguard to be the cluster that determines cluster membership first if all heartbeats are lost. Changing these settings can possibly cause that not to be the case, defeating the purpose of the Quorum Service. When the Quorum Service is not used the redundant heartbeat paths are protecting the cluster. In this case there is no restriction against changing the cluster timing settings.

Some extra care needs to be taken when administering the clusters. For example, if an administrator halts only HP Serviceguard for Linux on a node and does not halt CRS (RAC) on the same node, then a subsequent network partition may result in Serviceguard for Linux and the applications it is protecting becoming entirely unavailable. This would happen in case where the only remaining Serviceguard for Linux node is chosen to be rebooted by Oracle as a result of the network partition. The converse is true as well. The solution is simple, whenever an administrator takes an action that can affect the membership of one cluster (for example, "cmhaltnode" for HP Serviceguard for Linux) then the similar command for the other cluster should be performed on that same node ("srvctl stop" for RAC). Users may want to write simple scripts to handle these cases.

Clusters with more than two nodes

While all of the examples given have been for a two node cluster, this expands to larger clusters if necessary. If there is a great enough workload to justify a larger number of nodes, then consideration should be given to having a RAC cluster with only the database and a Serviceguard cluster with only the applications. If more than two nodes are needed for a combined HP Serviceguard for Linux and RAC cluster, then the failure scenarios are the same as is the configuration.

- If a single node fails, then both clusters will detect it and will have matching memberships.
- If only a single heartbeat network were used then a partition that resulted in a 50-50 split, with an equal number of nodes on each side, could cause the same problem as the loss of heartbeats in a 2 node cluster. The recommended configuration is the same, both with the number of networks carrying heartbeats, and the optional use of the Quorum Service.
- In the case of a partition with an unequal number of nodes, then the partition with more than 50% of nodes will survive. If there is a multiple partition with no part having greater than 50% of the nodes, then all nodes will go down. This is the same as with Serviceguard today.

Conclusion

HP Serviceguard for Linux adds value in a RAC cluster by providing high availability encapsulation for third party applications, while reducing the hardware requirements for the environment. HP Serviceguard for Linux can co-exist with Oracle 10g R1 RAC as a stable consolidated cluster, with a proper choice of redundant hardware and software components as mentioned below.

- ✓ Using multiple heartbeat subnets for HP Serviceguard for Linux and redundant heartbeat networks for HP Serviceguard for Linux and RAC (via channel-bonding) to prevent re-configuration on both clusters simultaneously.
- ✓ Optional use of Quorum service instead of Lock LUN for HP Serviceguard for Linux (using default parameters for both clusters) in addition to the multiple heartbeat paths to further minimize the possibility of a network partition resulting in both clusters becoming unavailable.

Related Materials

- “Oracle 10g Sales Guide” at http://oracle.hp.com/products/products.cfm?Product=oracle_10g
- “Oracle 10g Cluster Ready Services Positioning White Paper” at http://csps.fc.hp.com/ha/oracle/10g/10gRAC_on_hp_general.htm
- HP Serviceguard product documentation at <http://docs.hp.com>