ServiceGuard Cluster Configuration for Partitioned Systems

Manufacturing Part Number: B3936-90058
November 2001
Legal Notices

The information contained in this document is subject to change without notice. *Hewlett-Packard makes no warranty of any kind with regard to this manual, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose.* Hewlett-Packard shall not be liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.

Copyright © 2001 Hewlett-Packard Company.

This document contains information which is protected by copyright. All rights are reserved. Reproduction, adaptation, or translation without prior written permission is prohibited, except as allowed under the copyright laws.

HP ClusterView, HP OpenView, HP OpenView IT/Operations, MC/LockManager, and MC/ServiceGuard are products of Hewlett-Packard Company, and all are protected by copyright. VERITAS and VERITAS Volume Manager are registered trademarks of VERITAS Software Corporation. EMC², Symmetrix, and PowerPath are registered trademarks of EMC Corporation.

Corporate Offices:

*Hewlett-Packard Co.*

*3000 Hanover St.*

*Palo Alto, CA 94304*

Use, duplication or disclosure by the U.S. Government Department of Defense is subject to restrictions as set forth in paragraph (b)(3)(ii) of the Rights in Technical Data and Software clause in FAR 52.227-7013.

Rights for non-DOD U.S. Government Departments and Agencies are as set forth in FAR 52.227-19(c)(1,2).

Use of this manual and flexible disc(s), compact disc(s), or tape cartridge(s) supplied for this pack is restricted to this product only. Additional copies of the programs may be made for security and back-up purposes only. Resale of the programs in their present form or with alterations, is expressly prohibited. A copy of the specific warranty terms applicable to your Hewlett-Packard product and replacement parts can be obtained from your local Sales and Service Office.
Abstract

The MC/ServiceGuard product provides an infrastructure for the design and implementation of highly available HP-UX clusters that can quickly restore mission critical application services after hardware or software failures. To achieve the highest level of availability, clusters must be configured to eliminate all single points of failure (SPOF). This requires a careful analysis of the hardware and software infrastructure used to build the cluster. Partitioning technologies such as Superdome nPartitions and the HP-UX Virtual Partitions (VPARS) present some unique issues that must be considered when utilizing them within a ServiceGuard configuration. This document discusses these considerations.
Partition Configurations

Partitioning technologies such as nPartitions and VPARS provide increased flexibility in effectively managing system resources. They can be used to provide hardware and/or software fault isolation between applications sharing the same hardware platform. These technologies also allow hardware resources to be more efficiently utilized, based on application capacity requirements, and they provide the means to quickly re-deploy the hardware resources should the application requirements change. Given this capability, it is natural to want to utilize these technologies when designing MC/ServiceGuard clusters. Care must be taken, however, as the use of partitioning does present some unique failure scenarios that must be considered when designing a cluster to meet specific uptime requirements.

The partitioning provided by nPartitions is done at a hardware level and each partition is isolated from both hardware and software failures of other partitions. VPARS partitioning is implemented at a software level. While this provides greater flexibility in dividing hardware resources between partitions and allows partitioning on legacy systems, it does not provide any isolation of hardware failures between the partitions. Figure 1 illustrates nPartitions and VPARS configurations.

Figure 1  Sample nPartitions and VPARS Configurations
ServiceGuard Design Assumptions

To best understand issues related to using partitioning within the cluster, it will be helpful to start with a review of the ServiceGuard design philosophy and assumptions.

Hardware Redundancy

ServiceGuard, like all other HA clustering products, uses *hardware redundancy* to maintain application availability. For example, the ServiceGuard configuration guidelines require redundant networking paths between the nodes in the cluster. This requirement protects against total loss of communication to a node if a networking interface card fails. If a card should fail, there is a redundant card that can take over for it.

As can be readily seen, this strategy of hardware redundancy relies on an important underlying assumption: the failure of one component is independent of the failure of other components. That is, if the two networking cards were somehow related, then there could exist a failure event that would disable them both simultaneously. This represents a SPOF and effectively defeats the purpose of having redundant cards. It is for this reason that the ServiceGuard configuration rules do not allow both heartbeat networks on a node to travel through multiple ports on the same multi-ported networking interface. A single networking interface card failure would disable both heartbeat networks.

Cluster Membership Protocol

This same philosophy of hardware redundancy is reflected in the clustering concept. If a node in the cluster fails, another node is available to take over applications that were active on the failed node. Determining, with certainty, which nodes in the cluster are currently operational is accomplished through a cluster membership protocol whereby the nodes exchange heartbeat messages and maintain a cluster quorum.

After a failure that results in loss of communication between the nodes, active cluster nodes execute a cluster reformation algorithm that is used to determine the new cluster quorum. This new quorum, in conjunction with the previous quorum, is used to determine which nodes remain in
ServiceGuard Cluster Configuration for Partitioned Systems

ServiceGuard Design Assumptions

the new active cluster.

The algorithm for cluster reformation generally requires a cluster quorum of a strict majority, that is, more than 50% of the nodes that were previously running. However, exactly 50% of the previously running nodes are allowed to re-form as a new cluster, provided there is a guarantee that the other 50% of the previously running nodes do not also re-form. In these cases, some form of quorum arbitration or tie-breaker is needed. For example, if there is a communication failure between the nodes in a two-node cluster, and each node is attempting to re-form the cluster, then ServiceGuard must only allow one node to form the new cluster. This is accomplished by configuring a cluster lock.

The important concept to note here is that if more than 50% of the nodes in the cluster fail at the same time, then the remaining nodes have insufficient quorum to form a new cluster and fail themselves. This is irrespective of whether or not a cluster lock has been configured. It is for this reason that cluster configuration must be carefully analyzed to prevent failure modes that are common amongst the cluster nodes. One example of such area of concern is the power circuit considerations that are documented in HP 9000 Enterprise Servers Configuration Guide, Chapter 6. Another area where it is possible to have a greater than 50% node failure is the use of partitioned systems within the cluster. Configuration considerations for preventing this situation are described in the Section, “Partition Interactions”.

Quorum Arbitration

One form of quorum arbitration is a shared disk device configured as a cluster lock.

The cluster lock disk is a disk area located in a volume group that is shared by all nodes in the cluster. The cluster lock disk is used as a tie-breaker only for situations in which a running cluster fails and, as ServiceGuard attempts to form a new cluster, the cluster is split into two sub-clusters of equal size. Each sub-cluster attempts to acquire the cluster lock. The sub-cluster that gets the cluster lock forms the new cluster and the nodes that were unable to get the lock cease activity. This prevents the possibility of split-brain activity, that is, two sub-clusters running at the same time. If the two sub-clusters are of unequal size, the sub-cluster with greater than 50% of the previous quorum forms the new cluster and the cluster lock is not used.

For obvious reasons, two node cluster configurations are required to
ServiceGuard Cluster Configuration for Partitioned Systems

ServiceGuard Design Assumptions

configure some type of quorum arbitration. By definition, failure of a node or loss of communication in a two-node cluster results in a 50% partition. Due to the assumption that nodes fail independently of each other (independent failure assumption), the use of quorum arbitration for cluster configurations with three or more nodes is optional, though highly recommended.
Partition Interactions

With this background in mind, we next need to examine to what extent the partitioning schemes either meet or violate the independent failure assumption.

The partitioning provided by nPartitions is done at a hardware level and each partition is isolated from both hardware and software failures of other partitions. This provides very good isolation between the OS instances running within the partitions. So in this sense, nPartitions meets the assumption that the failure of one node (partition) will not affect other nodes. However, within the Superdome infrastructure, there does exist a very small possibility of a failure that can affect all partitions within the cabinet. So, to the extent that this infrastructure failure exists, nPartitions violates the independent failure assumption.

The VPARS form of partitioning is implemented at a software level. While this provides greater flexibility in dividing hardware resources between partitions and allows partitioning on legacy systems, it does not provide any isolation of hardware failures between the partitions. Thus the failure of a hardware component being used by one partition can bring down other partitions within the same hardware platform. From a software perspective, VPARS provides isolation for most software failures, such as kernel panics, between partitions. Due to the lack of hardware isolation however, there is no guarantee that a failure, such as a misbehaving kernel that erroneously writes to the wrong memory address, will not affect other OS partitions. Based on these observations, one can conclude that VPARS will violate the independent failure assumption to a greater degree than will nPartitions.

In addition to the failure case interactions, VPARS exhibit a behavior that should also be considered when including a VPARS as a node in a ServiceGuard cluster. Due to the nature of the hardware/firmware sharing between VPARS, it is possible for one partition to induce latency in other partitions. For example, during bootup, when the booting partition requests the system firmware to initialize the boot disk, it is possible for other partitions running in the same machine to become blocked until the initialization operation completes. During ServiceGuard qualification testing, delays of up to 13 seconds have been observed on systems with a PCI bus and SCSI disks. The ramifications of this type of latency are discussed in the section “Latency Considerations”.
Cluster Configuration Considerations

Using the information from the preceding sections, we can now assess any impacts or potential issues that arise from utilizing partitions (either VPARS or nPartitions) as part of a ServiceGuard cluster. From a ServiceGuard perspective, an OS instance running in a partition, is not treated any differently than OS instances running on non-partitioned nodes. Thus partitioning does not alter the basic ServiceGuard configuration rules as described in HP 9000 Enterprise Servers Configuration Guide, Chapter 6. An example of these existing configuration requirements is the need to have dual communication paths to both storage and networks. The use of partitioning does however introduce interesting configuration situations that necessitate additional configuration requirements. These are discussed below.

Quorum Arbitration Requirements

As previously mentioned, existing ServiceGuard configuration rules for non-partitioned systems require the use of a cluster lock only in the two node cluster case. This requirement is in place to protect against failures that result in a 50% quorum with respect to the membership prior to the failure. Clusters with more than two nodes do not have this as a strict requirement because of the “independent failure” assumption. As can be seen, this assumption is no longer valid when dealing with partitions. Cluster configurations that contain OS instances running within a partition, must be analyzed to determine the impact on cluster membership based on complete failure of hardware components that support more than one partition.

Rule 1. Configurations containing the potential for a loss of more that 50% membership resulting from a single failure are not supported.

These include configurations with the majority of nodes as partitions within a single hardware cabinet. This implies that in the two cabinet case, the partitions must be symmetrically divided between the cabinets.

For example, given three N-class systems, it would not be supported to create a five node cluster with three VPARS in one N-class and no partitioning in each of the other systems because the failure of the partitioned N-class would represent loss of greater than 50% of quorum.
ServiceGuard Cluster Configuration for Partitioned Systems

Cluster Configuration Considerations

(3 out of 5 partitions). Alternatively, the cluster would be supported if the systems without nPartitions each contained two VPARS, resulting in a seven-node cluster.

**Exception:** where all cluster nodes are running within partitions in a single cabinet (such as the so-called *cluster in a box* configuration). The configuration is supported provided users understand and accept the possibility of a complete cluster failure. This configuration is discussed in the Section, “Cluster In-A-Box”.

**Rule 2. Configurations containing the potential for a loss of exactly 50% membership resulting from a single failure require the use of a cluster lock.**

This includes:

- Cluster configurations where the nodes are running in partitions that are wholly contained within two hardware cabinets.
  
  **Example:** to be supported, a four-node cluster consisting of two nPartitions in each of two Superdome cabinets, would require the use a cluster lock.

- Cluster configurations where the nodes are running as VPARS partitions that are wholly contained within two nPartitions.

**NOTE**

At the time this document was created, VPARS are not supported within nPartitions, however this configuration is mentioned to illustrate the type of analysis that should be considered. It is also assumed that VPARS support within nPartitions will eventually happen.

**Cluster Configuration and Partitions**

Given the configuration requirements described in Rule 1 and Rule 2, a few interesting observations can be made of clusters utilizing partitioning:

- If it is determined that a cluster lock is needed for a particular configuration, the cluster must be configured such that the cluster lock is isolated from failures affecting the cluster nodes. This means that the lock device must be powered independently of the cluster nodes (such as hardware cabinets containing the partitions that make
up the cluster).

- Clusters wholly contained within two hardware cabinets and that utilize the cluster lock for quorum arbitration, are limited to either two or four nodes. This is due to a combination of the existing ServiceGuard rule that limits support of the cluster lock to four nodes and Rule 1.

- Cluster configurations can contain a mixture of VPARS, nPartitions, and independent nodes as long as quorum requirements are met.

- For a cluster configuration to contain no single points of failure, it must extend beyond a single hardware cabinet, comply with both the quorum rules and the ServiceGuard configuration rules described in *HP 9000 Enterprise Servers Configuration Guide*, Chapter 6.

## Cluster In-A-Box

One unique cluster configuration possibility that is enabled by partitioning is the so-called cluster in-a-box. In this case all the OS instances (nodes) of the cluster are running in partitions within the same hardware cabinet. While this configuration is subject to single points of failure, it may provide adequate availability characteristics for some applications and is thus considered a supported ServiceGuard configuration. Users must carefully assess the potential impact of a complete cluster failure on their availability requirements before choosing to deploy this type of cluster configuration.

A cluster in-a-box configuration consisting exclusively of VPARS is susceptible to a wider variety of possible failures, that could result in a complete cluster failure, than is a cluster made up exclusively of nPartitions.

## I/O Considerations

ServiceGuard does not treat OS instances running in a partition any differently than those running on an independent node. Thus, partitions do not provide any exemptions from the normal ServiceGuard connectivity rules (such as redundant paths for heartbeat networks and to storage) nor do they impose any new requirements. There is however a couple of interesting aspects related to partitioned systems that should be noted:

- While not a strictly “partitioning” issue per-se, the Superdome
ServiceGuard Cluster Configuration for Partitioned Systems

Cluster Configuration Considerations

A platform that supports nPartitions contains its interface cards in an I/O chassis and there can be more than one I/O chassis per partition. Since the I/O chassis represents a potential unit of failure, nPartitions redundant I/O paths must be configured in separate I/O chassis. Generally speaking Superdome provides enough I/O capacity that ServiceGuard's redundant path requirement should not constrain the use of partitioning within the cluster.

- VPARS on the other hand must share essentially one node's worth of I/O capacity. In this case, the redundant path requirement can be a limiting factor in determining the number of partitions that can be configured on a single hardware platform.

Example: assume we would like to create a cluster in-a-box configuration using an N-Class machine and a fibre channel based storage device. The redundant path requirement means that each partition would need two fibre channel interface cards for storage. Each partition would also need a minimum of two network interface cards for the heartbeat lans. Because there currently is no combination fibre channel/network card, this means that each partition would require a minimum of four interface cards. Given that the N-Class systems only have twelve I/O slots, this would mean that a maximum of three partitions could be configured.

The use of “combination” cards that combine both network and storage can help in some situations. However, redundant paths for a particular device must be split across separate interface cards (for example, using multiple ports on the same network interface card for the heartbeat lans is not supported).

Latency Considerations

As mentioned previously, there is a latency issue, unique to VPARS, that must be considered when configuring a ServiceGuard cluster to utilize VPARS.

There are certain operations performed by one partition (such as initializing the boot disk during bootup) that can induce delays in other partitions on the same hardware platform. The net result to ServiceGuard is the loss of cluster heartbeats if the delay exceeds the configured NODE_TIMEOUT parameter. If this should happen, the cluster starts the cluster re-formation protocol and, providing the delay is within the failover time, the delayed node simply rejoins the cluster. This results in cluster re-formation messages appearing in the syslog(1m)
file along with diagnostic messages from the ServiceGuard cluster monitor (cmcmd) describing the length of the delay.

For this reason, it is recommended that clusters containing nodes running in a VPARS partition, increase the NODE_TIMEOUT parameter to fourteen seconds in order to eliminate cluster reformations caused by latency within the VPARS nodes.

---

**NOTE**

This does not eliminate the cmcmd diagnostic messages that record delays of greater than three seconds.
ServiceGuard Cluster Configuration for Partitioned Systems

Cluster Configuration Considerations