the Availability Digest

HP's ServiceGuard Clustering Facility

May 2007

Clusters represent a predominant technology today to achieve high availability. A cluster is a set of independent processing nodes with access to a common database. It provides a single-server image to the users of the cluster. A cluster is managed by a cluster management facility such as HP's ServiceGuard.

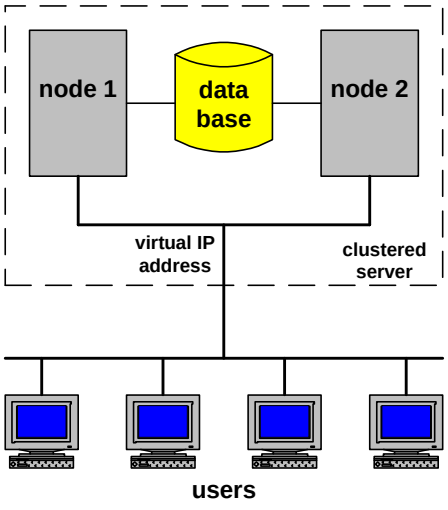
What is ServiceGuard?

HP's ServiceGuard is a cluster management facility enjoying over 150,000 licenses worldwide. It allows a company to customize and control its high availability clusters. With ServiceGuard, the business can organize its applications into packages. In the event of a hardware or software fault, the company can designate that control of specific packages be transferred to another processing node in the cluster or that communications be transferred to a standby LAN

ServiceGuard supports clusters based either on HP's HP-UX systems or on Linux.

What is a Cluster?

A cluster is a configuration of two or more servers, or nodes, in a high-availability configuration. This means that each server in the cluster is backed up by some other server in the cluster. Should one server fail, its functions are taken over by its backup server.



Typically, only one server at a time can access a particular database. This is to avoid data corruption due to two servers trying to write to the same data item at the same time. (An exception is Oracle's Real Application Clusters database, which coordinates activity among several nodes in a cluster).

There are three cluster configurations supported by ServiceGuard:

- Active/Standby – An active server is backed up by a standby server, which is otherwise idle or is being used for some noncritical function which can be instantly terminated should the active server fail.

© 2007 Sombers Associates, Inc., and W. H. Highleyman
www.availabilitydigest.com

- Active/Active[1] - Each server in the cluster is actively running different applications as well as acting as a backup for other servers. Should a server fail, its backup will continue processing its own applications while also assuming the processing of the applications that had been running on the failed processor.

- Parallel Database – Multiple servers are running the same application against the same database. This can only be done with Oracle's RAC and is a special ServiceGuard extension (HP ServiceGuard Extension for RAC, or SGeRAC). Should one server fail, the other servers take over its processing load.

ServiceGuard's Cluster Services

There are several services that ServiceGuard brings to a cluster.[2]

Packages

In a cluster, the failover unit is not a server. Rather, it is an application. Some large servers in clusters can be running several applications simultaneously. Should there be a failure, the failure may be in the application, not in the server. It may be necessary to fail over only that application to a backup server.

An application comprises the application code, the database it uses, and one or more IP addresses that are used by users to access the application. These are combined together by ServiceGuard into a *package*.[3] It is the package that is the unit of failover, not the server.

Heartbeats

Heartbeat messages are periodically exchanged between the nodes in the cluster to inform each other of their health. One node, the *cluster coordinator*, sends and receives these heartbeat messages to all other nodes in the cluster. Should it not receive a response to a heartbeat from some node within a specified time period, it will declare that node to be down. Likewise, if a node does not receive a heartbeat from the cluster coordinator, that node will declare that the cluster coordinator node is down.

Heartbeats are the very center of cluster management. It is extremely important that the heartbeat mechanism be highly reliable. Therefore, it is common to provide a redundant private connection for heartbeat messages.

Transfer of Control (TOC)

If a node cannot communicate with a majority of the other nodes in its cluster, it causes itself to fail. This *Transfer of Control (TOC)* is initiated by the ServiceGuard cluster software to ensure that only one application is modifying a particular database at any one time.

A Transfer of Control, or TOC, is the act of failing over. Should ServiceGuard decide to take down a node via a TOC, it will transfer the packages on that node to the node's backup. Likewise, if only an application fails, its package may be transferred to another processing node via a TOC.

---

[1] Note that this is not our definition of active/active.

[2] Weygant, P.S., *Clusters for High Availability*, Prentice-Hall, Inc.; 2001.

[3] A package is also known as a service group.

© 2007 Sombers Associates, Inc., and W. H. Highleyman
www.availabilitydigest.com

The backup node can be specified to be either a specific node or to be the node with the fewest packages running on it.

The TOC procedure stops the application on its current node, starts the application on its backup node, directs the application to open its database on the new node, and remaps the application's IP addresses to the new node.

Cluster Quorum

The ServiceGuard software is monitoring the health of all of the nodes in the cluster. In the event of a node failure, the cluster re-forms itself without the failed node.

Should there be a communication failure between two sets of two or more nodes, ServiceGuard re-forms the cluster around the larger set and causes the nodes in the smaller set to fail via a TOC. The larger set of surviving nodes is called the *cluster quorum*.

Should the two sets of separated nodes be of the same size, they will both attempt to become the new cluster quorum. However, it is important that only one succeed in order to prevent data corruption. This can be implemented either through a cluster lock or via a Quorum Server.

Cluster Lock

ServiceGuard can provide a *cluster lock* which must be held by the current cluster quorum. Should the cluster be separated into two sets of the same size, both sets will attempt to become the new quorum by attempting to seize the cluster lock. The successful set will become the new cluster quorum. The losing set will cause its nodes to fail via a TOC.

Quorum Server

A Quorum Server is an alternative to a cluster lock. It is software running on an independent high-availability system or on a cluster that monitors the cluster nodes via heartbeats. Should there be a node or a communication failure, the Quorum Server determines the nodes that will create the new cluster quorum. Thus, if there is a splitting of the cluster into two sets of an equal number of nodes, the Quorum Server will determine which set will become the new cluster quorum.

Hardware Monitoring and Failover

ServiceGuard monitors the processor, disks, and networks of the cluster. If it should determine that a server can no longer function properly, it will fail over all of the packages currently running on that server to backup servers.

Application Monitoring and Recovery

Under normal conditions, ServiceGuard monitors the health of all of the cluster components. This includes the applications running in the cluster.

An application is started by ServiceGuard with a special cluster command that continually monitors the health of that application. Should ServiceGuard ever receive an indication of an error exit from an application, it will initiate recovery action. ServiceGuard can be directed to attempt to restart the application, to halt it, or to fail over the package to the package's backup server via a TOC.

© 2007 Sombers Associates, Inc., and W. H. Highleyman
www.availabilitydigest.com

LAN Monitoring and Recovery

In addition to nodes and applications, ServiceGuard also monitors the cluster's LANs. It can quickly detect a LAN problem and will activate a spare LAN in the same cluster. This failover is transparent to both the users and the databases.

Workload Balancing

Following a failover, ServiceGuard can be instructed to move the packages from the failed node to other nodes in such a way as to balance the new load across the surviving nodes.

In addition, the system administrator can, at any time, move a package from one node to another to balance the cluster load.

Failover/Failback Policy

The system administrator can establish failover and failback policies. As mentioned earlier, a failover policy might be to specify that a package fails over to a specific node or to the node currently running the least number of packages.

Each package can be specified to fail back to its original node once that node is returned to service. Alternatively, it can be specified that the package remain on its backup node until the system administrator moves it.

Rolling Upgrades

An application or a node can be upgraded by removing it from the cluster, upgrading it, and returning it to the cluster. The procedure is as follows:

- Move the applications on the node to be upgraded to other nodes.
- Remove the node from the cluster.
- Perform the upgrades.
- Allow the node to rejoin the cluster.
- Move its applications back to the node.

ServiceGuard Manager

The ServiceGuard Manager is a separate HP product that provides a graphical view of all of the cluster components. It displays a cluster map showing all of the cluster components, their properties, and their status. It is invaluable to quickly determine the location of a problem and to track the actions of ServiceGuard.

Summary

ServiceGuard provides all of the services needed to efficiently manage an HP-UX or Linux cluster. With over 150,000 installations around the world, ServiceGuard is clearly an important contribution to the quest to achieve high availabilities. It is a very important element of HP's stated goal to achieve 5 9s:5 minutes of reliability (an availability of five 9s implies an average down time of five minutes per year).

© 2007 Sombers Associates, Inc., and W. H. Highleyman
www.availabilitydigest.com