the Availability Digest

www.availabilitydigest.com
@availabilitydig

HP Serviceguard Cluster Arbitration and Fencing Mechanisms

Ravi Krishnamurthy, HP Master Architect

January 2014

High availability clusters protect services and applications from unplanned downtimes and allow for shorter planned downtimes by facilitating faster and easier maintenance. They monitor hardware and software for faults and fail over applications to healthy nodes in the cluster.

Amongst the many faults that clusters handle, network faults need more care and involve more complexity in handling, since they can affect heartbeat and hence can affect cluster membership as well. Arbitration mechanisms are needed to handle network faults. In addition, fencing mechanisms are needed to avoid data corruption, a possible side-effect of network faults.

The Impact of Network Partitioning

High availability clusters can experience network partitions in spite of using designs with redundant network paths and switches. When network failures happen in a specific sequence or only in specific segments in clusters using these networks for heartbeats, they result in the cluster being partitioned in various ways. When such partitions happen, resulting in nodes of each partition unable to talk to each other, clusters typically use arbitration mechanisms to determine which partition survives and which one will be evicted from its membership. After the arbitrator determines the partition to be evicted, I/O fencing mechanisms like SCSI-3 Persistent reservation in combination with node fencing mechanisms are used to ensure that the evicted members of the cluster reboot and that no I/O is generated from them until they rejoin the cluster.

Various arbitration mechanisms such as a quorum server process, a dedicated shared LUN, or a volume group per cluster are used for the purpose of arbitration by different clusters. HP Serviceguard cluster for HP-UX, as well as Linux, supports many such mechanisms as arbitrators. There are different advantages and disadvantages for each type of arbitration mechanism. These are discussed here in the context of HP Serviceguard.

Various arbitration mechanisms

Quorum Server

A quorum server is a process that runs on a separate node outside of the cluster and arbitrates in the event of a network partition. Its advantage is that it is very simple to set up and can be shared by up to 50 clusters for the purpose of arbitration. A disadvantage, though, is that during a network partition, the cluster nodes may not be able to communicate with the quorum server due to partitions in the network. To overcome this, HP Serviceguard supports multiple paths from each cluster node to the quorum server to provide redundancy. If one path fails, the other path is used by the node to communicate with the quorum server.

Lock LUN

The lock LUN arbitration mechanism is a small, dedicated, shared LUN on which a fast mutex structure is laid out. The coordinator servers of each partition try to acquire the mutex on the disk, and the partition that obtains the lock survives while the other one is evicted from the cluster membership.

Volume Group Lock

Another arbitrator on HP-UX is the volume group lock, or vglock, where the same fast mutex is laid out on the metadata of the volume group so that coordinator members of each partition try to acquire the lock in order to survive the partition.

Disk based arbitration mechanisms like lock LUN and vglock are convenient and useful when the number of nodes in a cluster is four or less and all the nodes are co-located. A quorum server is preferable either when the number of cluster nodes exceed four or when they are located in more than one location, i.e., clusters stretched across cities or located across two Metropolis.

Handling network partitions and fencing

When a Serviceguard cluster experiences a 50 - 50 network partition, i.e., the cluster is partitioned into exactly two sets of nodes that can communicate within themselves but not to each other, the cluster requests the configured arbitration mechanism to select one of these partitions to continue in the membership of that cluster. Once the arbitrator selects a partition, the nodes of the other partition timeout and evict themselves from the cluster. When the two sides of the network partition have an unequal number of nodes, the cluster chooses the partition that has a majority number of nodes, without employing an arbitrator. Clusters are recommended to be configured with redundant network paths for heartbeat and data so that a single failure does not cause cluster reformations or workload failovers.

Clusters typically use various fencing mechanisms. HP Serviceguard uses a deadman kernel module to ensure that nodes that are evicted from the membership of a cluster reboot within a guaranteed duration during a network partition. This is the first of two guarantees that are essential to prevent data corruption. The second one is the prevention of ghost I/O. HP Serviceguard ensures this by using SCSI-3 Persistent reservation technology on all the data storage that is configured as part of the cluster. The registration for the node that is evicted from the membership is revoked from the shared cluster storage so that delayed I/O from that node is not honored by the storage after the cluster reconfiguration phase is completed.

Arbitration and fencing mechanisms ensure clusters handle network partitions and the associated requirements to prevent data corruption and ghost I/O. The mechanisms associated with these actions are critical to the stable functioning of a cluster.

Ravi Krishnamurthy is a Master Architect from HP who has experience in the domains of High Availability, Embedded Operating Systems, and Diagnostics.