HA additional details

ONTAP Select

Barb Einarsen
November 19, 2019

Table of Contents

HA additional details

HA disk heartbeating, HA mailbox, HA heartbeating, HA Failover, and Giveback work to enhance data protection.

Disk heartbeating

Although the ONTAP Select HA architecture leverages many of the code paths used by the traditional FAS arrays, some exceptions exist. One of these exceptions is in the implementation of disk-based heartbeating, a nonnetwork-based method of communication used by cluster nodes to prevent network isolation from causing split-brain behavior. A split-brain scenario is the result of cluster partitioning, typically caused by network failures, whereby each side believes the other is down and attempts to take over cluster resources.

Enterprise-class HA implementations must gracefully handle this type of scenario. ONTAP does this through a customized, disk-based method of heartbeating. This is the job of the HA mailbox, a location on physical storage that is used by cluster nodes to pass heartbeat messages. This helps the cluster determine connectivity and therefore define quorum in the event of a failover.

On FAS arrays, which use a shared storage HA architecture, ONTAP resolves split-brain issues in the following ways:

- SCSI persistent reservations
- Persistent HA metadata
- HA state sent over HA interconnect

However, within the shared-nothing architecture of an ONTAP Select cluster, a node is only able to see its own local storage and not that of the HA partner. Therefore, when network partitioning isolates each side of an HA pair, the preceding methods of determining cluster quorum and failover behavior are unavailable.

Although the existing method of split-brain detection and avoidance cannot be used, a method of mediation is still required, one that fits within the constraints of a shared-nothing environment. ONTAP Select extends the existing mailbox infrastructure further, allowing it to act as a method of mediation in the event of network partitioning. Because shared storage is unavailable, mediation is accomplished through access to the mailbox disks over NAS. These disks are spread throughout the cluster, including the mediator in a two-node cluster, using the iSCSI protocol. Therefore, intelligent failover decisions can be made by a cluster node based on access to these disks. If a node can access the mailbox disks of other nodes outside of its HA partner, it is likely up and healthy.

The mailbox architecture and disk-based heartbeating method of resolving cluster quorum and split-brain issues are the reasons the multinode variant of ONTAP Select requires either four separate nodes or a mediator for a two-node cluster.

HA mailbox posting

The HA mailbox architecture uses a message post model. At repeated intervals, cluster nodes post messages to all other mailbox disks across the cluster, including the mediator, stating that the node is up and running. Within a healthy cluster at any point in time, a single mailbox disk on a cluster node has messages posted from all other cluster nodes.

Attached to each Select cluster node is a virtual disk that is used specifically for shared mailbox access. This disk is referred to as the mediator mailbox disk, because its main function is to act as a method of cluster

mediation in the event of node failures or network partitioning. This mailbox disk contains partitions for each cluster node and is mounted over an iSCSI network by other Select cluster nodes. Periodically, these nodes post health statuses to the appropriate partition of the mailbox disk. Using network-accessible mailbox disks spread throughout the cluster allows you to infer node health through a reachability matrix. For example, cluster nodes A and B can post to the mailbox of cluster node D, but not to the mailbox of node C. In addition, cluster node D cannot post to the mailbox of node C, so it is likely that node C is either down or network isolated and should be taken over.

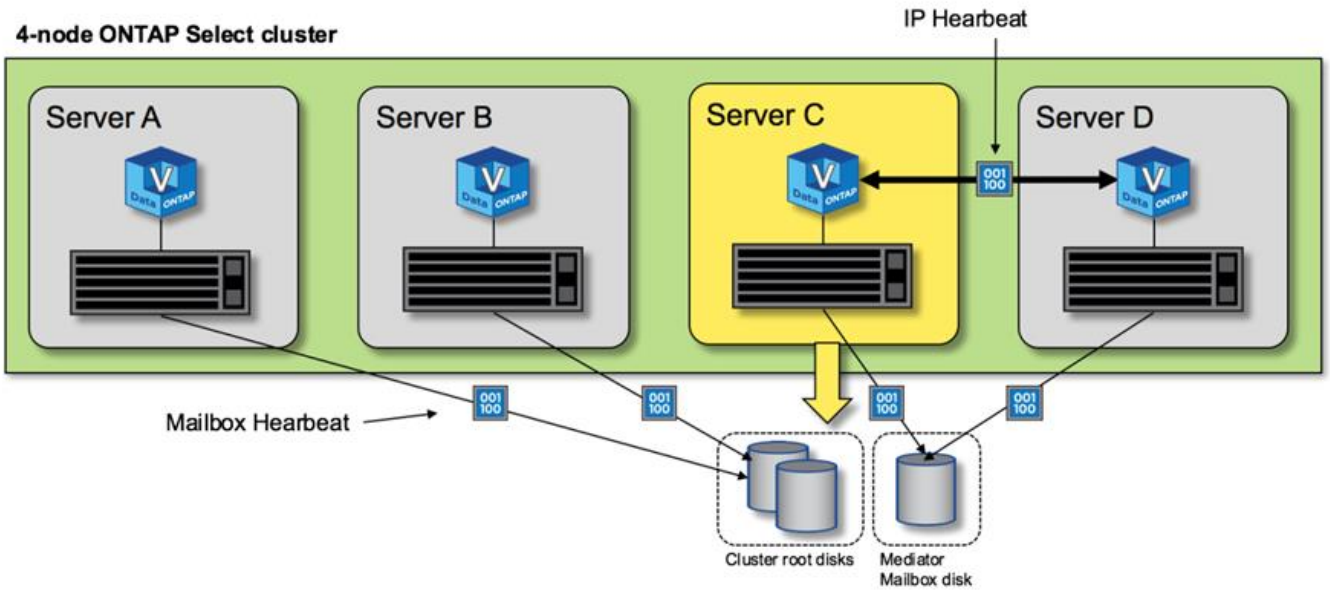
HA heartbeating

Like with NetApp FAS platforms, ONTAP Select periodically sends HA heartbeat messages over the HA interconnect. Within the ONTAP Select cluster, this is performed over a TCP/IP network connection that exists between HA partners. Additionally, disk-based heartbeat messages are passed to all HA mailbox disks, including mediator mailbox disks. These messages are passed every few seconds and read back periodically. The frequency with which these are sent and received allows the ONTAP Select cluster to detect HA failure events within approximately 15 seconds, the same window available on FAS platforms. When heartbeat messages are no longer being read, a failover event is triggered.

The following figure shows the process of sending and receiving heartbeat messages over the HA interconnect and mediator disks from the perspective of a single ONTAP Select cluster node, node C.

Network heartbeats are sent over the HA interconnect to the HA partner, node D, while disk heartbeats use mailbox disks across all cluster nodes, A, B, C, and D.

HA heartbeating in a four-node cluster: steady state



HA failover and giveback

During a failover operation, the surviving node assumes the data serving responsibilities for its peer node using the local copy of its HA partner's data. Client I/O can continue uninterrupted, but changes to this data must be replicated back before giveback can occur. Note that ONTAP Select does not support a forced giveback because this causes changes stored on the surviving node to be lost.

The sync back operation is automatically triggered when the rebooted node rejoins the cluster. The time required for the sync back depends on several factors. These factors include the number of changes that must

be replicated, the network latency between the nodes, and the speed of the disk subsystems on each node. It is possible that the time required for sync back will exceed the auto give back window of 10 minutes. In this case, a manual giveback after the sync back is required. The progress of the sync back can be monitored using the following command:

```
storage aggregate status -r -aggregate <aggregate name>
```