High availability architecture

ONTAP Select

NetApp
November 29, 2022

Table of Contents

High availability architecture

High availability configurations

Discover high availability options to select the best HA configuration for your environment.

Although customers are starting to move application workloads from enterprise-class storage appliances to software-based solutions running on commodity hardware, the expectations and needs around resiliency and fault tolerance have not changed. An HA solution providing a zero recovery point objective (RPO) protects the customer from data loss due to a failure from any component in the infrastructure stack.

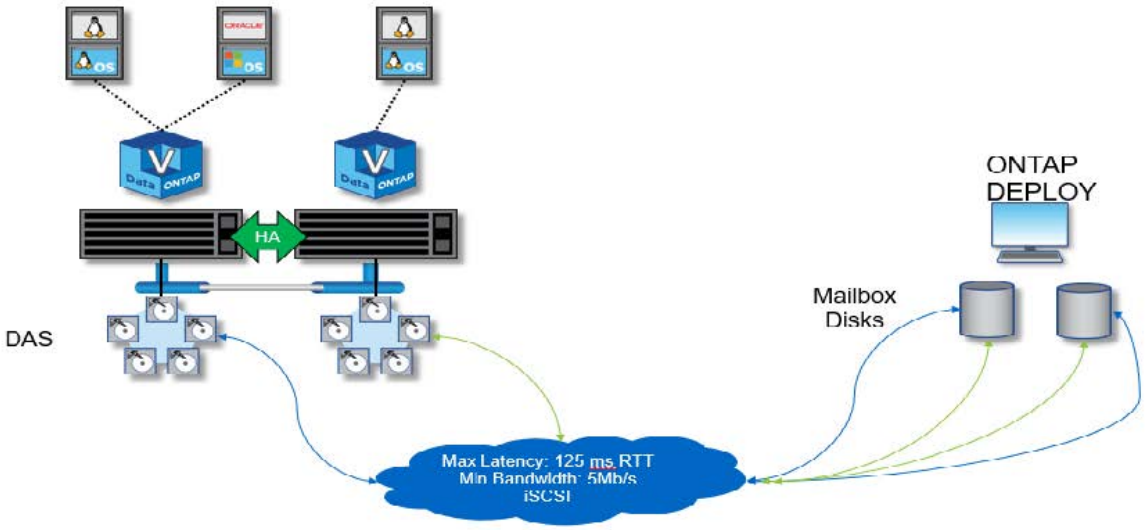
A large portion of the SDS market is built on the notion of shared-nothing storage, with software replication providing data resiliency by storing multiple copies of user data across different storage silos. ONTAP Select builds on this premise by using the synchronous replication features (RAID SyncMirror) provided by ONTAP to store an extra copy of user data within the cluster. This occurs within the context of an HA pair. Every HA pair stores two copies of user data: one on storage provided by the local node, and one on storage provided by the HA partner. Within an ONTAP Select cluster, HA and synchronous replication are tied together, and the functionality of the two cannot be decoupled or used independently. As a result, the synchronous replication functionality is only available in the multinode offering.

> In an ONTAP Select cluster, synchronous replication functionality is a function of HA implementation, not a replacement for the asynchronous SnapMirror or SnapVault replication engines. Synchronous replication cannot be used independently from HA.

There are two ONTAP Select HA deployment models: the multinode clusters (four, six, or eight nodes) and the two-node clusters. The salient feature of a two-node ONTAP Select cluster is the use of an external mediator service to resolve split-brain scenarios. The ONTAP Deploy VM serves as the default mediator for all the two-node HA pairs that it configures.

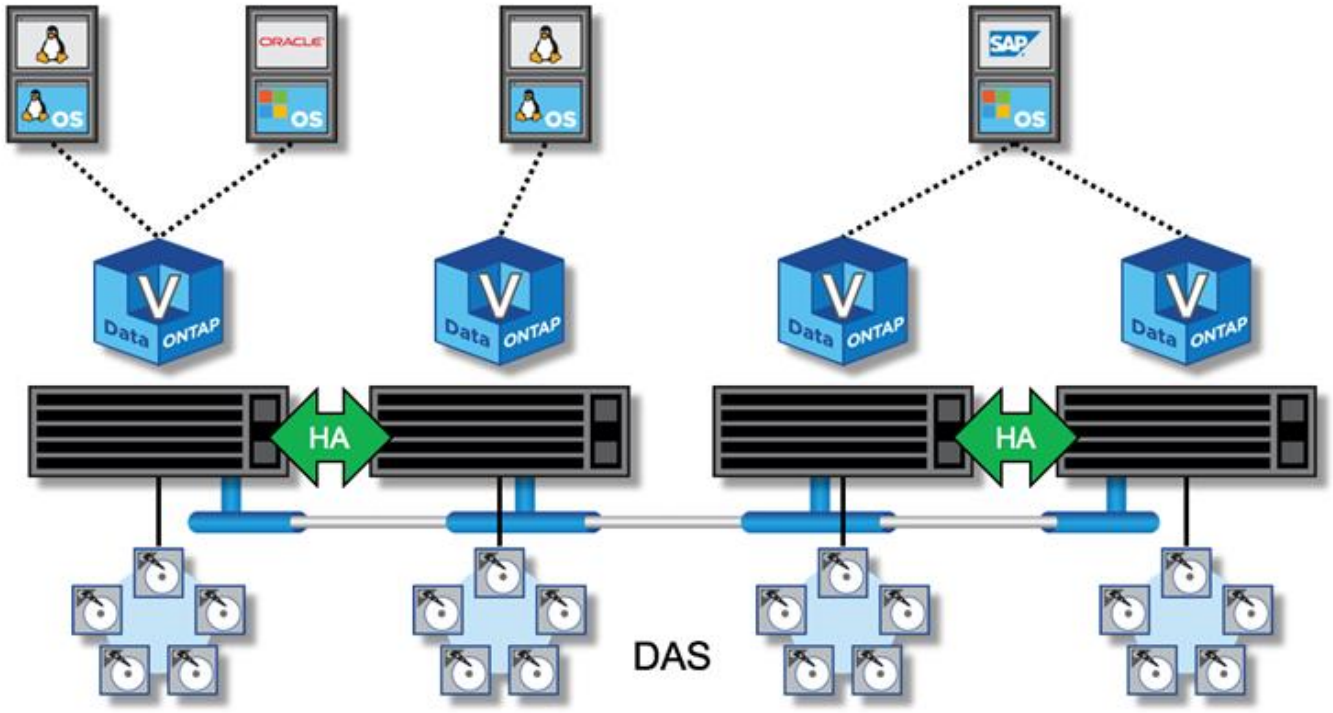
The two architectures are represented in the following figures.

Two-node ONTAP Select cluster with remote mediator and using local-attached storage



The two-node ONTAP Select cluster is composed of one HA pair and a mediator. Within the HA pair, data aggregates on each cluster node are synchronously mirrored, and, in the event of a failover, there is no loss of data.

Four-node ONTAP Select cluster using local-attached storage



- The four-node ONTAP Select cluster is composed of two HA pairs. Six-node and eight-node clusters are composed of three and four HA pairs, respectively. Within each HA pair, data aggregates on each cluster node are synchronously mirrored, and, in the event of a failover, there is no loss of data.
- Only one ONTAP Select instance can be present on a physical server when using DAS storage. ONTAP Select requires unshared access to the local RAID controller of the system and is designed to manage the locally attached disks, which would be impossible without physical connectivity to the storage.

Two-node HA versus multi-node HA

Unlike FAS arrays, ONTAP Select nodes in an HA pair communicate exclusively over the IP network. That means that the IP network is a single point of failure (SPOF), and protecting against network partitions and split-brain scenarios becomes an important aspect of the design. The multinode cluster can sustain single-node failures because the cluster quorum can be established by the three or more surviving nodes. The two-node cluster relies on the mediator service hosted by the ONTAP Deploy VM to achieve the same result.

The heartbeat network traffic between the ONTAP Select nodes and the ONTAP Deploy mediator service is minimal and resilient so that the ONTAP Deploy VM can be hosted in a different data center than the ONTAP Select two-node cluster.

The ONTAP Deploy VM becomes an integral part of a two-node cluster when serving as the mediator for that cluster. If the mediator service is not available, the two-node cluster continues serving data, but the storage failover capabilities of the ONTAP Select cluster are disabled. Therefore, the ONTAP Deploy mediator service must maintain constant communication with each ONTAP Select node in the HA pair. A minimum bandwidth of 5Mbps and a maximum round-trip time (RTT) latency of 125ms are required to allow proper functioning of the cluster quorum.

If the ONTAP Deploy VM acting as a mediator is temporarily or potentially permanently unavailable, a secondary ONTAP Deploy VM can be used to restore the two-node cluster quorum. This results in a configuration in which the new ONTAP Deploy VM is unable to manage the ONTAP Select nodes, but it successfully participates in the cluster quorum algorithm. The communication between the ONTAP Select nodes and the ONTAP Deploy VM is done by using the iSCSI protocol over IPv4. The ONTAP Select node management IP address is the initiator, and the ONTAP Deploy VM IP address is the target. Therefore, it is not possible to support IPv6 addresses for the node management IP addresses when creating a two-node cluster. The ONTAP Deploy hosted mailbox disks are automatically created and masked to the proper ONTAP Select node management IP addresses at the time of two-node cluster creation. The entire configuration is automatically performed during setup, and no further administrative action is required. The ONTAP Deploy instance creating the cluster is the default mediator for that cluster.

An administrative action is required if the original mediator location must be changed. It is possible to recover a cluster quorum even if the original ONTAP Deploy VM is lost. However, NetApp recommends that you back up the ONTAP Deploy database after every two-node cluster is instantiated.

Two-node HA versus two-node stretched HA (MetroCluster SDS)

It is possible to stretch a two-node, active/active HA cluster across larger distances and potentially place each node in a different data center. The only distinction between a two-node cluster and a two-node stretched cluster (also referred to as MetroCluster SDS) is the network connectivity distance between nodes.

The two-node cluster is defined as a cluster for which both nodes are located in the same data center within a distance of 300m. In general, both nodes have uplinks to the same network switch or set of interswitch link (ISL) network switches.

Two-node MetroCluster SDS is defined as a cluster for which nodes are physically separated (different rooms, different buildings, and different data centers) by more than 300m. In addition, each node's uplink connections are connected to separate network switches. The MetroCluster SDS does not require dedicated hardware. However, the environment should adhere to requirements for latency (a maximum of 5ms for RTT and 5ms for jitter, for a total of 10ms) and physical distance (a maximum of 10km).

MetroCluster SDS is a premium feature and requires a Premium license. The Premium license supports the creation of both small and medium VMs, as well as HDD and SSD media.

MetroCluster SDS is supported with both local attached storage (DAS) and shared storage (vNAS). Note that vNAS configurations usually have a higher innate latency because of the network between the ONTAP Select VM and shared storage. MetroCluster SDS configurations must provide a maximum of 10ms of latency between the nodes, including the shared storage latency. In other words, only measuring the latency between the Select VMs is not adequate because shared storage latency is not negligible for these configurations.

HA RSM and mirrored aggregates

Prevent data loss using RAID SyncMirror (RSM), mirrored aggregates, and the write path.

Synchronous replication

The ONTAP HA model is built on the concept of HA partners. ONTAP Select extends this architecture into the nonshared commodity server world by using the RAID SyncMirror (RSM) functionality that is present in ONTAP to replicate data blocks between cluster nodes, providing two copies of user data spread across an HA pair.

A two-node cluster with a mediator can span two data centers. For more information, see the section Two-node stretched HA (MetroCluster SDS) best practices.

Mirrored aggregates

An ONTAP Select cluster is composed of two to eight nodes. Each HA pair contains two copies of user data, synchronously mirrored across nodes over an IP network. This mirroring is transparent to the user, and it is a property of the data aggregate, automatically configured during the data aggregate creation process.

All aggregates in an ONTAP Select cluster must be mirrored for data availability in the event of a node failover and to avoid an SPOF in case of hardware failure. Aggregates in an ONTAP Select cluster are built from virtual disks provided from each node in the HA pair and use the following disks:

- A local set of disks (contributed by the current ONTAP Select node)
- A mirrored set of disks (contributed by the HA partner of the current node)

> The local and mirror disks used to build a mirrored aggregate must be the same size. These aggregates are referred to as plex 0 and plex 1 (to indicate the local and remote mirror pairs, respectively). The actual plex numbers can be different in your installation.

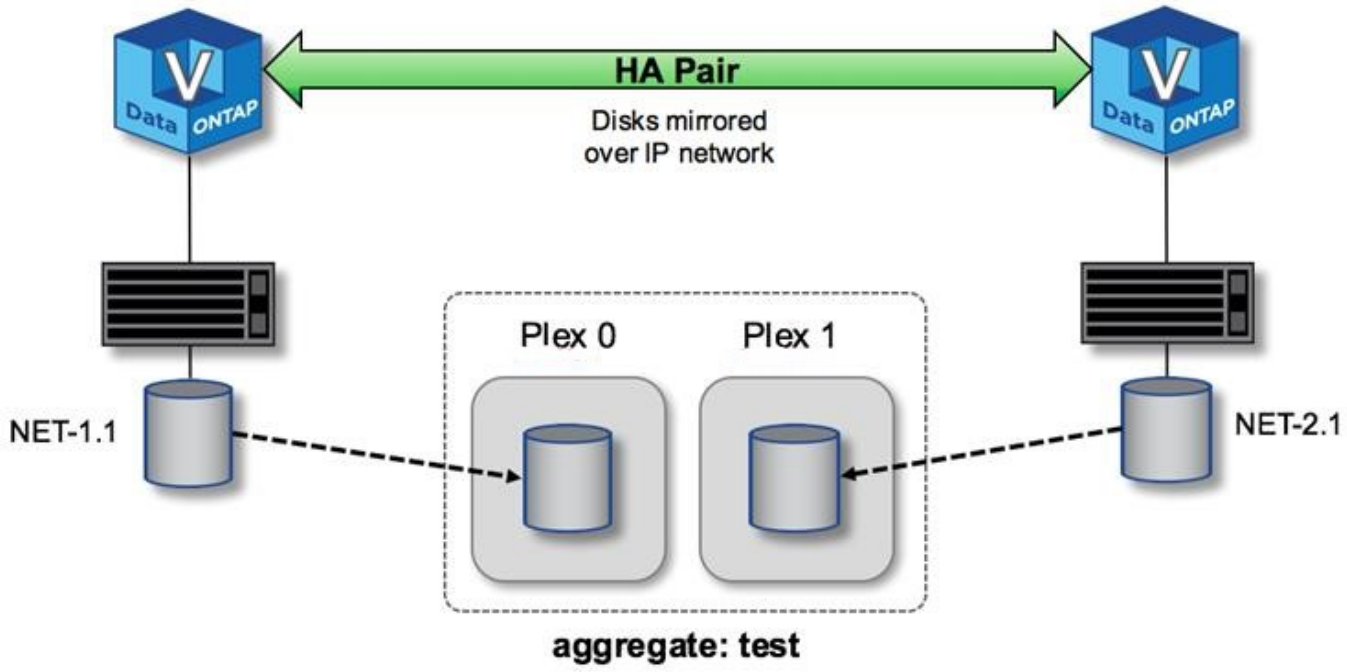
This approach is fundamentally different from the way standard ONTAP clusters work. This applies to all root and data disks within the ONTAP Select cluster. The aggregate contains both local and mirror copies of data. Therefore, an aggregate that contains N virtual disks offers N/2 disks' worth of unique storage, because the second copy of data resides on its own unique disks.

The following figure shows an HA pair within a four-node ONTAP Select cluster. Within this cluster is a single aggregate (test) that uses storage from both HA partners. This data aggregate is composed of two sets of virtual disks: a local set, contributed by the ONTAP Select owning cluster node (Plex 0), and a remote set, contributed by the failover partner (Plex 1).

Plex 0 is the bucket that holds all local disks. Plex 1 is the bucket that holds mirror disks, or disks responsible for storing a second replicated copy of user data. The node that owns the aggregate contributes disks to Plex 0, and the HA partner of that node contributes disks to Plex 1.

In the following figure, there is a mirrored aggregate with two disks. The contents of this aggregate are mirrored across our two cluster nodes, with local disk NET-1.1 placed into the Plex 0 bucket and remote disk NET-2.1 placed into the Plex 1 bucket. In this example, aggregate test is owned by the cluster node to the left and uses local disk NET-1.1 and HA partner mirror disk NET-2.1.

ONTAP Select mirrored aggregate



When an ONTAP Select cluster is deployed, all virtual disks present on the system are automatically assigned to the correct plex, requiring no additional step from the user regarding disk assignment. This prevents the accidental assignment of disks to an incorrect plex and provides optimal mirror disk configuration.

Write Path

Synchronous mirroring of data blocks between cluster nodes and the requirement for no data loss with a system failure have a significant impact on the path an incoming write takes as it propagates through an ONTAP Select cluster. This process consists of two stages:

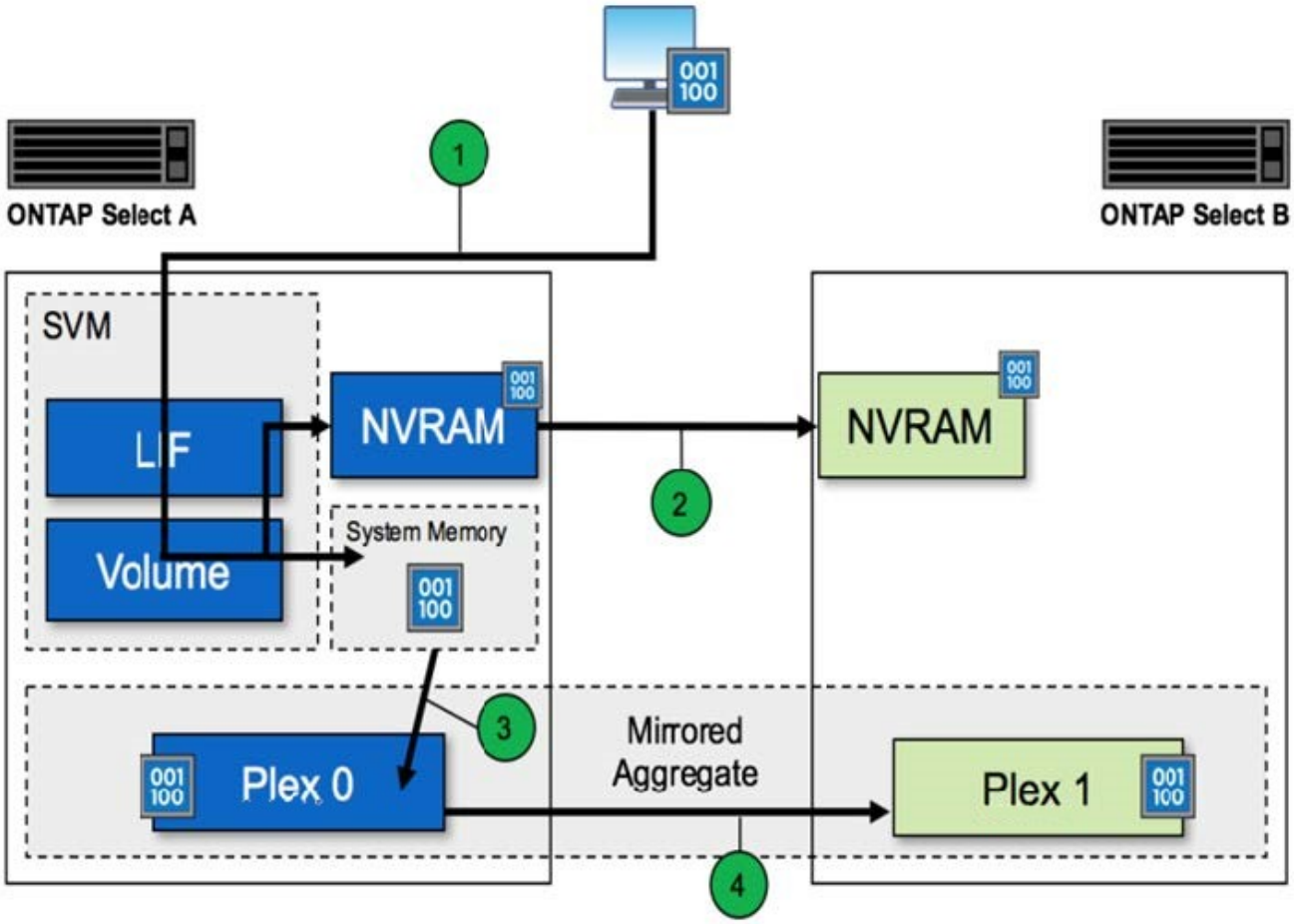
- Acknowledgment
- Destaging

Writes to a target volume occur over a data LIF and are committed to the virtualized NVRAM partition, present on a system disk of the ONTAP Select node, before being acknowledged back to the client. On an HA configuration, an additional step occurs, because these NVRAM writes are immediately mirrored to the HA partner of the target volume's owner before being acknowledged. This process makes sure of the file system consistency on the HA partner node, if there is a hardware failure on the original node.

After the write has been committed to NVRAM, ONTAP periodically moves the contents of this partition to the appropriate virtual disk, a process known as destaging. This process only happens once, on the cluster node owning the target volume, and does not happen on the HA partner.

The following figure shows the write path of an incoming write request to an ONTAP Select node.

ONTAP Select write path workflow



Incoming write acknowledgment includes the following steps:

- Writes enter the system through a logical interface owned by ONTAP Select node A.
- Writes are committed to the NVRAM of node A and mirrored to the HA partner, node B.
- After the I/O request is present on both HA nodes, the request is then acknowledged back to the client.

ONTAP Select destaging from NVRAM to the data aggregate (ONTAP CP) includes the following steps:

- Writes are destaged from virtual NVRAM to virtual data aggregate.
- Mirror engine synchronously replicates blocks to both plexes.

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

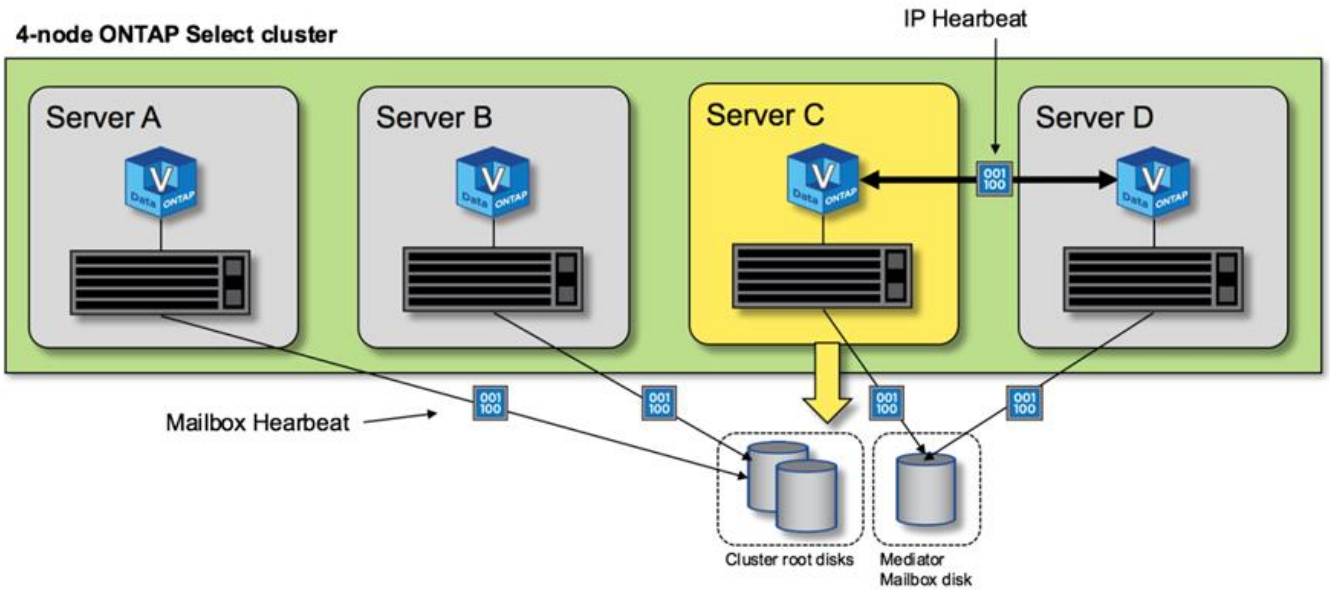
Like with NetApp FAS platforms, ONTAP Select periodically sends HA heartbeat messages over the HA interconnect. Within the ONTAP Select cluster, this is performed over a TCP/IP network connection that exists between HA partners. Additionally, disk-based heartbeat messages are passed to all HA mailbox 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>
```

Copyright information

Trademark information