HA RSM and mirrored aggregates

ONTAP Select

NetApp
November 21, 2019

Table of Contents

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.

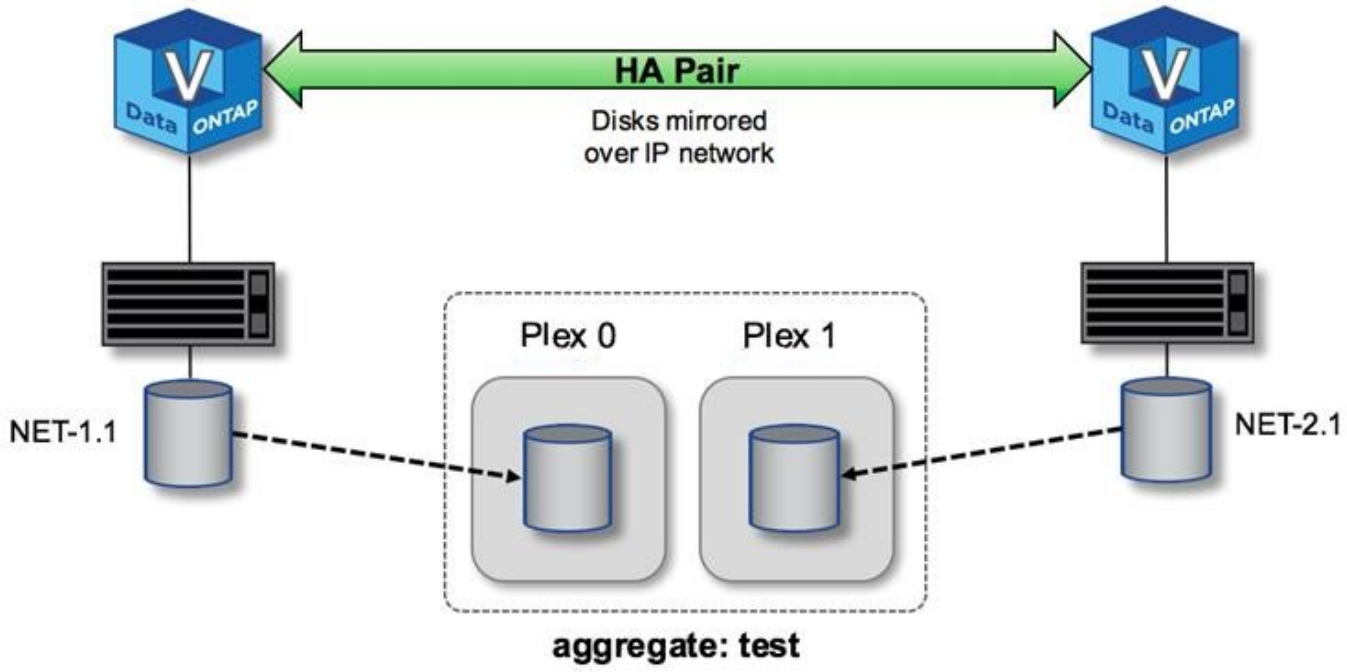
This approach is fundamentally different from the way standard ONTAP clusters work. This applies to all root and data disks within the ONTAP Select cluster. The aggregate contains both local and mirror copies of data. Therefore, an aggregate that contains N virtual disks offers N/2 disks' worth of unique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

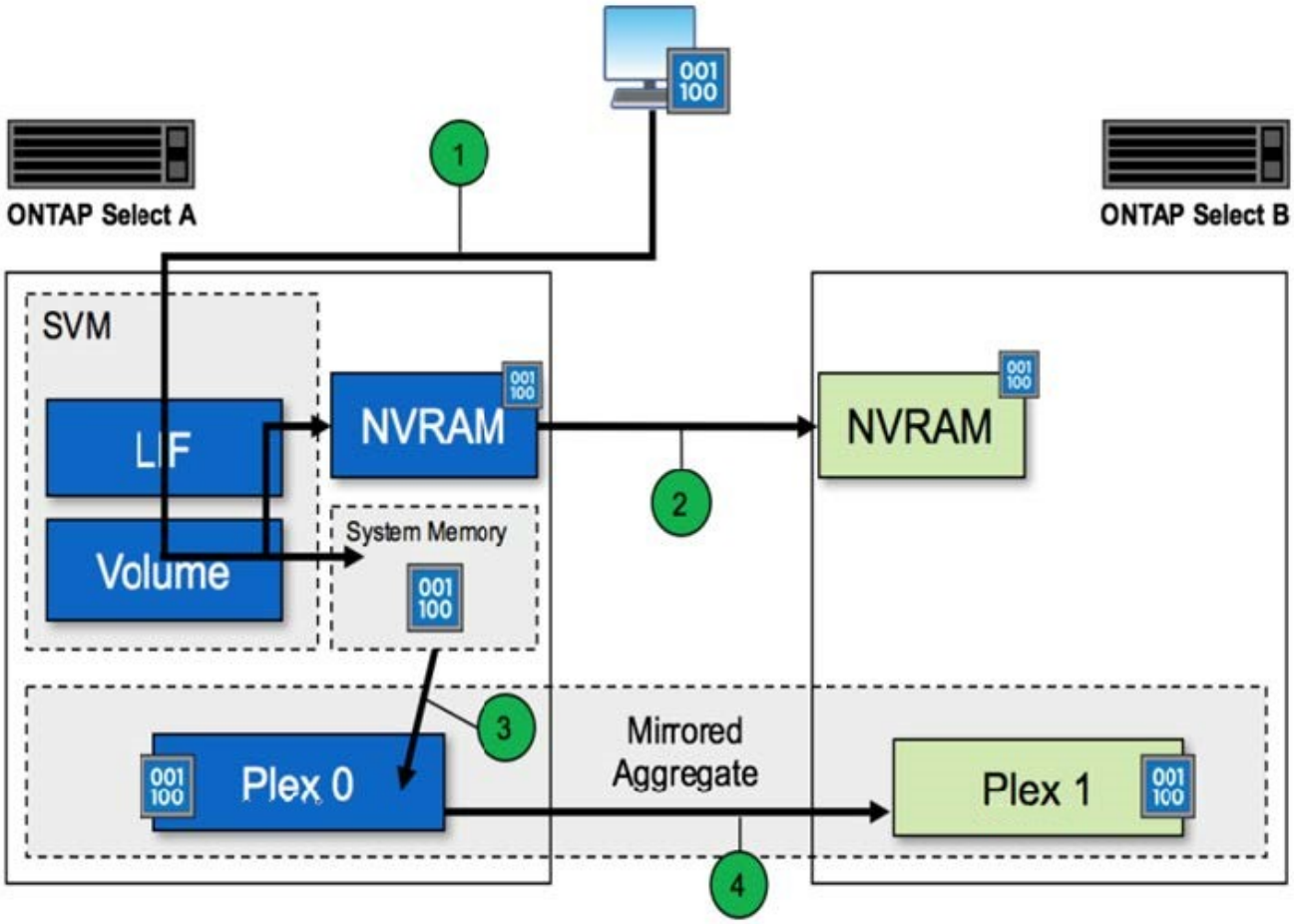
- Acknowledgment
- Destaging

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.