Perform switchover, healing, and switchback

ONTAP MetroCluster

NetApp
December 22, 2022

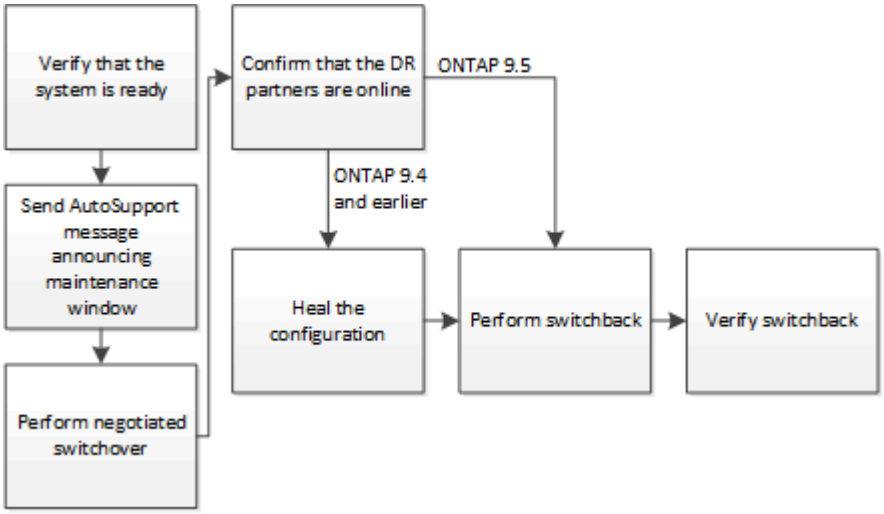
Table of Contents

Perform switchover, healing, and switchback

Perform switchover for tests or maintenance

Performing switchover for tests or maintenance

If you want to test the MetroCluster functionality or to perform planned maintenance, you can perform a negotiated switchover in which one cluster is cleanly switched over to the partner cluster. You can then heal and switch back the configuration.



Beginning with ONTAP 9.6, switchover and switchback operations can be performed on MetroCluster IP configurations with ONTAP System Manager.

Verifying that your system is ready for a switchover

You can use the `-simulate` option to preview the results of a switchover operation. A verification check gives you a way to verify that most of the preconditions for a successful run are met before you start the operation. Issue these commands from the site that will remain up and operational:

1. Set the privilege level to advanced: `set -privilege advanced`
2. From the site that will remain up and operational, simulate a switchover operation: `metrocluster switchover -simulate`
3. Review the output that is returned.

   The output shows whether any vetoes would prevent a switchover operation. Every time you perform a MetroCluster operation, you must verify a set of criteria for the success of the operation. A “veto” is a mechanism to prohibit the operation if one or more of the criteria are not fulfilled. There are two types of veto: a “soft” veto and a “hard” veto. You can override a soft veto, but not a hard veto. For example, to perform a negotiated switchover in a four-node MetroCluster configuration, one criterion is that all of the nodes are up and healthy. Suppose one node is down and was taken over by its HA partner. The switchover operation will be hard vetoed because it is a hard criterion that all of the nodes must be up and healthy. Because this is a hard veto, you cannot override the veto.

It is best not to override any veto.

Example: Verification results

The following example shows the errors that are encountered in a simulation of a switchover operation:

```
cluster4::*> metrocluster switchover -simulate

[Job 126] Preparing the cluster for the switchover operation...
[Job 126] Job failed: Failed to prepare the cluster for the switchover
operation. Use the "metrocluster operation show" command to view detailed
error
information. Resolve the errors, then try the command again.
```

Negotiated switchover and switchback will fail until you replace all of the failed disks. You can perform disaster recovery after you replace the failed disks. If you want to ignore the warning for failed disks, you can add a soft veto for the negotiated switchover and switchback.

Sending a custom AutoSupport message prior to negotiated switchover

Before performing a negotiated switchover, you should issue an AutoSupport message to notify NetApp technical support that maintenance is underway. The negotiated switchover might result in plex or MetroCluster operation failures that trigger AutoSupport messages. Informing technical support that maintenance is underway prevents them from opening a case on the assumption that a disruption has occurred.

This task must be performed on each MetroCluster site.

Steps

1. Log in to the cluster at Site_A.
2. Invoke an AutoSupport message indicating the start of the maintenance: `system node autosupport invoke -node * -type all -message MAINT=maintenance-window-in-hours`

   maintenance-window-in-hours specifies the length of the maintenance window and can be a maximum of 72 hours. If the maintenance is completed before the time has elapsed, you can issue a command to indicating that the maintenance period has ended:`system node autosupport invoke -node * -type all -message MAINT=end`

3. Repeat this step on the partner site.

Performing a negotiated switchover

A negotiated switchover cleanly shuts down processes on the partner site, and then switches over operations from the partner site. You can use a negotiated switchover to perform maintenance on a MetroCluster site or to test the switchover functionality.

- All previous configuration changes must be completed before performing a switchback operation.

  This is to avoid competition with the negotiated switchover or switchback operation.

- Any nodes that were previously down must be booted and in cluster quorum.

  The *System Administration Reference* has more information about cluster quorum in the “Understanding quorum and epsilon” section.

  System administration

- The cluster peering network must be available from both sites.
- All of the nodes in the MetroCluster configuration must be running the same version of ONTAP software.
- The option replication.create_data_protection_rels.enable must be set to ON on both of the sites in a MetroCluster configuration before creating a new SnapMirror relationship.
- For a two-node MetroCluster configuration, a new SnapMirror relationship should not be created during an upgrade when there are mismatched versions of ONTAP between the sites.
- For a four-node MetroCluster configuration, the mismatched versions of ONTAP between the sites are not supported.

The recovering site can take a few hours to be able to perform the switchback operation.

The metrocluster switchover command switches over the nodes in all DR groups in the MetroCluster configuration. For example, in an eight-node MetroCluster configuration, it switches over the nodes in both DR groups.

While preparing for and executing a negotiated switchover, you must not make configuration changes to either cluster or perform any takeover or giveback operations.

For MetroCluster FC configurations:

- Mirrored aggregates will remain in normal state if the remote storage is accessible.
- Mirrored aggregates will become degraded after the negotiated switchover if access to the remote storage is lost.
- Unmirrored aggregates that are located at the disaster site will become unavailable if access to the remote storage is lost. This might lead to a controller outage.

For MetroCluster IP configurations:

> Before performing maintenance tasks, you must remove monitoring if the MetroCluster configuration is monitored with the Tiebreaker or Mediator utility. Remove ONTAP Mediator or Tiebreaker monitoring before performing maintenance tasks

- For ONTAP 9.4 and earlier:
  - Mirrored aggregates will become degraded after the negotiated switchover.
- For ONTAP 9.5 and later:
  - Mirrored aggregates will remain in normal state if the remote storage is accessible.
  - Mirrored aggregates will become degraded after the negotiated switchover if access to the remote storage is lost.

- For ONTAP 9.8 and later:
  - Unmirrored aggregates that are located at the disaster site will become unavailable if access to the remote storage is lost. This might lead to a controller outage.
    1. Use the metrocluster check run, metrocluster check show and metrocluster check config-replication show commands to make sure no configuration updates are in progress or pending. Issue these commands from the site that will remain up and operational.
    2. From the site that will remain up and operational, implement the switchover: `metrocluster switchover`

       The operation can take several minutes to complete.

    3. Monitor the completion of the switchover: `metrocluster operation show`

```
cluster_A::*> metrocluster operation show
  Operation: Switchover
 Start time: 10/4/2012 19:04:13
      State: in-progress
   End time: -
     Errors:

cluster_A::*> metrocluster operation show
  Operation: Switchover
 Start time: 10/4/2012 19:04:13
      State: successful
   End time: 10/4/2012 19:04:22
     Errors: -
```

    4. Reestablish any SnapMirror or SnapVault configurations.

Verify that the SVMs are running and the aggregates are online

After the switchover is complete, you should verify that the DR partners have taken ownership of the disks and the partner SVMs have come online.

When you run the storage aggregate plex show command after a MetroCluster switchover, the status of plex0 of the switched over root aggregate is indeterminate and is displayed as failed. During this time, the switched over root is not updated. The actual status of this plex can only be determined after the MetroCluster healing phase.

Steps

1. Check that the aggregates were switched over by using the storage aggregate show command.

   In this example, the aggregates were switched over. The root aggregate (aggr0_b2) is in a degraded state. The data aggregate (b2_aggr2) is in a mirrored, normal state:

```
cluster_A::*> storage aggregate show

.
.
.
mcc1-b Switched Over Aggregates:
Aggregate     Size Available Used% State   #Vols  Nodes            RAID
Status
--------- -------- --------- ----- ------- ------ ----------------
------------
aggr0_b2    227.1GB   45.1GB    80% online       0 node_A_1
raid_dp,

mirror

degraded
b2_aggr1    227.1GB   200.3GB    20% online       0 node_A_1
raid_dp,

mirrored

normal
```

2. Confirm that the secondary SVMs have come online by using the vserver show command.

   In this example, the previously dormant sync-destination SVMs on the secondary site have been activated and have an Admin State of running:

```
cluster_A::*> vserver show
                                          Admin      Operational  Root
Name    Name
Vserver           Type  Subtype           State      State        Volume
Aggregate  Service Mapping
-----------       ----- ----------        ---------- -----------
--------- ---------- ------- -------
...
cluster_B-vs1b-mc data   sync-destination  running   running
vs1b_vol   aggr_b1   file    file
```

Heal the configuration

Heal the configuration in a MetroCluster FC configuration

Healing the configuration in a MetroCluster FC configuration

Following a switchover, you must perform the healing operations in specific order to restore MetroCluster functionality.

- Switchover must have been performed and the surviving site must be serving data.
- Nodes on the disaster site must be halted or remain powered off.

  They must not be fully booted during the healing process.

- Storage at the disaster site must be accessible (shelves are powered up, functional, and accessible).
- In fabric-attached MetroCluster configurations, inter-switch links (ISLs) must be up and operating.
- In four-node MetroCluster configurations, nodes in the surviving site must not be in HA failover state (all nodes must be up and running for each HA pair).

The healing operation must first be performed on the data aggregates, and then on the root aggregates.

Healing the data aggregates after negotiated switchover

You must heal the data aggregates after completing any maintenance or testing. This process resynchronizes the data aggregates and prepares the disaster site for normal operation. You must heal the data aggregates prior to healing the root aggregates.

All configuration updates in the remote cluster successfully replicate to the local cluster. You power up the storage on the disaster site as part of this procedure, but you do not and must not power up the controller modules on the disaster site.

Steps

1. Ensure that switchover has been completed by running the metrocluster operation show command.

```
controller_A_1::> metrocluster operation show
  Operation: switchover
      State: successful
 Start Time: 7/25/2014 20:01:48
   End Time: 7/25/2014 20:02:14
     Errors: -
```

2. Resynchronize the data aggregates by running the metrocluster heal -phase aggregates command from the surviving cluster.

```
controller_A_1::> metrocluster heal -phase aggregates
[Job 130] Job succeeded: Heal Aggregates is successful.
```

   If the healing is vetoed, you have the option of reissuing the metrocluster heal command with the --override -vetoes parameter. If you use this optional parameter, the system overrides any soft vetoes that prevent the healing operation.

3. Verify that the operation has been completed by running the metrocluster operation show command.

```
controller_A_1::> metrocluster operation show
    Operation: heal-aggregates
      State: successful
Start Time: 7/25/2014 18:45:55
   End Time: 7/25/2014 18:45:56
     Errors: -
```

4. Check the state of the aggregates by running the storage aggregate show command.

```
controller_A_1::> storage aggregate show
Aggregate     Size Available Used% State   #Vols  Nodes            RAID
Status
--------- -------- --------- ----- ------- ------ ----------------
------------
...
aggr_b2    227.1GB   227.1GB    0% online       0 mcc1-a2
raid_dp, mirrored, normal...
```

5. If storage has been replaced at the disaster site, you might need to remirror the aggregates.

Healing the root aggregates after negotiated switchover

After the data aggregates have been healed, you must heal the root aggregates in preparation for the switchback operation.

The data aggregates phase of the MetroCluster healing process must have been completed successfully.

Steps

1. Switch back the mirrored aggregates by running the metrocluster heal -phase root-aggregates command.

```
cluster_A::> metrocluster heal -phase root-aggregates
[Job 137] Job succeeded: Heal Root Aggregates is successful
```

   If the healing is vetoed, you have the option of reissuing the metrocluster heal command with the --override -vetoes parameter. If you use this optional parameter, the system overrides any soft vetoes that prevent the healing operation.

2. Confirm the heal operation is complete by running the metrocluster operation show command on the healthy cluster:

```
cluster_A::> metrocluster operation show
  Operation: heal-root-aggregates
      State: successful
 Start Time: 7/29/2014 20:54:41
   End Time: 7/29/2014 20:54:42
     Errors: -
```

3. Check for and remove any failed disks belonging to the disaster site by issuing the following command on the healthy site: `disk show -broken`
4. Power up or boot each controller module on the disaster site.

   If the system displays the LOADER prompt, run the `boot_ontap` command.

5. After nodes are booted, verify that the root aggregates are mirrored.

   If both plexes are present, resynchronization will occur automatically if the plexes are not synchronized. If one plex has failed, that plex must be destroyed and the mirror must be recreated using the storage aggregate mirror -aggregateaggregate-name command to reestablish the mirror relationship.

Healing the configuration in a MetroCluster IP configuration (ONTAP 9.4 and earlier)

You must heal the aggregates in preparation for the switchback operation.

> On MetroCluster IP systems running ONTAP 9.5, healing is performed automatically, and you can skip these tasks.

The following conditions must exist before performing the healing procedure:

- Switchover must have been performed and the surviving site must be serving data.
- Storage shelves at the disaster site must be powered up, functional, and accessible.
- ISLs must be up and operating.
- Nodes in the surviving site must not be in HA failover state (both nodes must be up and running).

This task applies to MetroCluster IP configurations running ONTAP versions prior to 9.5 only.

This procedure differs from the healing procedure for MetroCluster FC configurations.

Steps

1. Power up each controller module on the site that was switched over and let them fully boot.

   If the system displays the LOADER prompt, run the `boot_ontap` command.

2. Perform the root aggregate healing phase: `metrocluster heal root-aggregates`

```
cluster_A::> metrocluster heal root-aggregates
[Job 137] Job succeeded: Heal Root-Aggregates is successful
```

If the healing is vetoed, you have the option of reissuing the metrocluster heal root-aggregates command with the --override-vetoes parameter. If you use this optional parameter, the system overrides any soft vetoes that prevent the healing operation.

3. Resynchronize the aggregates: `metrocluster heal aggregates`

```
cluster_A::> metrocluster heal aggregates
[Job 137] Job succeeded: Heal Aggregates is successful
```

If the healing is vetoed, you have the option of reissuing the metrocluster heal command with the --override -vetoes parameter. If you use this optional parameter, the system overrides any soft vetoes that prevent the healing operation.

4. Confirm the heal operation is complete by running the metrocluster operation show command on the healthy cluster:

```
cluster_A::> metrocluster operation show
  Operation: heal-aggregates
      State: successful
 Start Time: 7/29/2017 20:54:41
   End Time: 7/29/2017 20:54:42
     Errors: -
```

Performing a switchback

After you heal the MetroCluster configuration, you can perform the MetroCluster switchback operation. The MetroCluster switchback operation returns the configuration to its normal operating state, with the sync-source storage virtual machines (SVMs) on the disaster site active and serving data from the local disk pools.

- The disaster cluster must have successfully switched over to the surviving cluster.
- Healing must have been performed on the data and root aggregates.
- The surviving cluster nodes must not be in the HA failover state (all nodes must be up and running for each HA pair).
- The disaster site controller modules must be completely booted and not in the HA takeover mode.
- The root aggregate must be mirrored.
- The Inter-Switch Links (ISLs) must be online.
- Any required licenses must be installed on the system.

1. Confirm that all nodes are in the enabled state: `metrocluster node show`

   The following example displays the nodes that are in the enabled state:

```
cluster_B::>  metrocluster node show

DR                         Configuration  DR
Group Cluster Node         State          Mirroring Mode
----- ------- ----------- -------------- ---------
--------------------
1     cluster_A
              node_A_1    configured     enabled   heal roots
completed
              node_A_2    configured     enabled   heal roots
completed
      cluster_B
              node_B_1    configured     enabled   waiting for
switchback recovery
              node_B_2    configured     enabled   waiting for
switchback recovery
4 entries were displayed.
```

2. Confirm that resynchronization is complete on all SVMs: `metrocluster vserver show`
3. Verify that any automatic LIF migrations being performed by the healing operations have been successfully completed: `metrocluster check lif show`
4. Perform a simulated switchback to verify the system is ready: `metrocluster switchback -simulate`
5. Check the configuration:

   `metrocluster check run`

   The command runs as a background job and might not be completed immediately.

```
cluster_A::> metrocluster check run
The operation has been started and is running in the background. Wait
for
it to complete and run "metrocluster check show" to view the results.
To
check the status of the running metrocluster check operation, use the
command,
"metrocluster operation history show -job-id 2245"
```

```
cluster_A::> metrocluster check show
Last Checked On: 9/13/2018 20:41:37

Component           Result
------------------- ---------
nodes               ok
lifs                ok
config-replication  ok
aggregates          ok
clusters            ok
connections         ok
6 entries were displayed.
```

6. Perform the switchback by running the metrocluster switchback command from any node in the surviving cluster: `metrocluster switchback`
7. Check the progress of the switchback operation: `metrocluster show`

   The switchback operation is still in progress when the output displays waiting-for-switchback:

```
cluster_B::> metrocluster show
Cluster                   Entry Name          State
------------------------- ------------------- -----------
 Local: cluster_B         Configuration state configured
                          Mode                switchover
                          AUSO Failure Domain -
Remote: cluster_A         Configuration state configured
                          Mode                waiting-for-switchback
                          AUSO Failure Domain -
```

   The switchback operation is complete when the output displays normal:

```
cluster_B::> metrocluster show
Cluster                   Entry Name          State
------------------------- ------------------- -----------
 Local: cluster_B         Configuration state configured
                          Mode                normal
                          AUSO Failure Domain -
Remote: cluster_A         Configuration state configured
                          Mode                normal
                          AUSO Failure Domain -
```

   If a switchback takes a long time to finish, you can check on the status of in-progress baselines by using the `metrocluster config-replication resync-status show` command. This command is at the advanced privilege level.

8. Reestablish any SnapMirror or SnapVault configurations.

   In ONTAP 8.3, you need to manually reestablish a lost SnapMirror configuration after a MetroCluster switchback operation. In ONTAP 9.0 and later, the relationship is reestablished automatically.

Verifying a successful switchback

After performing the switchback, you want to confirm that all aggregates and storage virtual machines (SVMs) are switched back and online.

1. Verify that the switched-over data aggregates are switched back:

   ```
   storage aggregate show
   ```

   In the following example, aggr_b2 on node B2 has switched back:

   ```
   node_B_1::> storage aggregate show
   Aggregate     Size Available Used% State   #Vols  Nodes            RAID
   Status
   --------- -------- --------- ----- ------- ------ ----------------
   ------------
   ...
   aggr_b2    227.1GB   227.1GB    0% online       0 node_B_2   raid_dp,

   mirrored,

   normal
   ```

2. Verify that all sync-destination SVMs on the surviving cluster are dormant (showing an admin state of “stopped”) and the sync-source SVMs on the disaster cluster are up and running:

   ```
   vserver show -subtype sync-source
   ```

```
node_B_1::> vserver show -subtype sync-source
                                Admin      Root
Name    Name
Vserver     Type    Subtype    State      Volume     Aggregate
Service Mapping
----------- ------- ---------- ---------- ---------- ----------
------- -------
...
vs1a        data    sync-source
                                running    vs1a_vol   node_B_2
file    file

aggr_b2

node_A_1::> vserver show -subtype sync-destination
                                Admin      Root
Name    Name
Vserver             Type    Subtype    State      Volume     Aggregate
Service Mapping
-----------         ------- ---------- ---------- ---------- ----------
------- -------
...
cluster_A-vs1a-mc  data    sync-destination
                                       stopped    vs1a_vol   sosb_
file    file

aggr_b2
```

Sync-destination aggregates in the MetroCluster configuration have the suffix “-mc” automatically appended to their name to help identify them.

3. Confirm that the switchback operations succeeded by using the `metrocluster operation show` command.

If the command output shows…	Then…
That the switchback operation state is successful.	The switchback process is complete and you can proceed with operation of the system.
That the switchback operation or switchback-continuation-agent operation is partially successful.	Perform the suggested fix provided in the output of the `metrocluster operation show` command.

You must repeat the previous sections to perform the switchback in the opposite direction. If site_A did a switchover of site_B, have site_B do a switchover of site_A.

Commands for switchover, healing, and switchback

There are specific ONTAP commands for performing the MetroCluster disaster recovery processes.

If you want to…	Use this command…
Verify that switchover can be performed without errors or vetoes.	`metrocluster switchover -simulate` at the advanced privilege level
Verify that switchback can be performed without errors or vetoes.	`metrocluster switchback -simulate` at the advanced privilege level
Switch over to the partner nodes (negotiated switchover).	`metrocluster switchover`
Switch over to the partner nodes (forced switchover).	`metrocluster switchover -forced-on-disaster true`
Perform data aggregate healing.	`metrocluster heal -phase aggregates`
Perform root aggregate healing.	`metrocluster heal -phase root-aggregates`
Switch back to the home nodes.	`metrocluster switchback`

Monitoring the MetroCluster configuration

You can use ONTAP MetroCluster commands and Active IQ Unified Manager (formerly OnCommand Unified Manager) to monitor the health of a variety of software components and the state of MetroCluster operations.

Checking the MetroCluster configuration

You can check that the components and relationships in the MetroCluster configuration are working correctly. You should do a check after initial configuration and after making any changes to the MetroCluster configuration. You should also do a check before a negotiated (planned) switchover or a switchback operation.

About this task

If the `metrocluster check run` command is issued twice within a short time on either or both clusters, a conflict can occur and the command might not collect all data. Subsequent `metrocluster check show` commands do not show the expected output.

Steps

1. Check the configuration:

```
metrocluster check run
```

The command runs as a background job and might not be completed immediately.

```
cluster_A::> metrocluster check run
The operation has been started and is running in the background. Wait
for
it to complete and run "metrocluster check show" to view the results. To
check the status of the running metrocluster check operation, use the
command,
"metrocluster operation history show -job-id 2245"
```

2. Display more detailed results from the most recent `metrocluster check run` command:

```
metrocluster check aggregate show

metrocluster check cluster show

metrocluster check config-replication show

metrocluster check lif show

metrocluster check node show
```

The `metrocluster check show` commands show the results of the most recent `metrocluster check run` command. You should always run the `metrocluster check run` command prior to using the `metrocluster check show` commands so that the information displayed is current.

The following example shows the `metrocluster check aggregate show` command output for a healthy four-node MetroCluster configuration:

```
cluster_A::> metrocluster check aggregate show

Last Checked On: 8/5/2014 00:42:58

Node                  Aggregate                 Check
Result
---------------       --------------------      ---------------------
---------
controller_A_1        controller_A_1_aggr0
                                                mirroring-status
ok
                                                disk-pool-allocation
ok
                                                ownership-state
ok
                      controller_A_1_aggr1
                                                mirroring-status
ok
```

```
                                                    disk-pool-allocation
ok
                                                    ownership-state
ok
                      controller_A_1_aggr2
                                                    mirroring-status
ok
                                                    disk-pool-allocation
ok
                                                    ownership-state
ok


controller_A_2        controller_A_2_aggr0
                                                    mirroring-status
ok
                                                    disk-pool-allocation
ok
                                                    ownership-state
ok
                      controller_A_2_aggr1
                                                    mirroring-status
ok
                                                    disk-pool-allocation
ok
                                                    ownership-state
ok
                      controller_A_2_aggr2
                                                    mirroring-status
ok
                                                    disk-pool-allocation
ok
                                                    ownership-state
ok

18 entries were displayed.
```

The following example shows the `metrocluster check cluster show` command output for a healthy four-node MetroCluster configuration. It indicates that the clusters are ready to perform a negotiated switchover if necessary.

```
Last Checked On: 9/13/2017 20:47:04

Cluster               Check                          Result
--------------------- ------------------------------ ---------
mccint-fas9000-0102
                      negotiated-switchover-ready    not-applicable
                      switchback-ready               not-applicable
                      job-schedules                  ok
                      licenses                       ok
                      periodic-check-enabled         ok
mccint-fas9000-0304
                      negotiated-switchover-ready    not-applicable
                      switchback-ready               not-applicable
                      job-schedules                  ok
                      licenses                       ok
                      periodic-check-enabled         ok
10 entries were displayed.
```

Commands for checking and monitoring the MetroCluster configuration

There are specific ONTAP commands for monitoring the MetroCluster configuration and checking MetroCluster operations.

Commands for checking MetroCluster operations

If you want to…	Use this command…
Perform a check of the MetroCluster operations. **Note:** This command should not be used as the only command for pre-DR operation system validation.	`metrocluster check run`
View the results of the last check on MetroCluster operations.	`metrocluster show`
View results of check on configuration replication between the sites.	`metrocluster check config-replication show metrocluster check config-replication show-aggregate-eligibility`
View results of check on node configuration.	`metrocluster check node show`
View results of check on aggregate configuration.	`metrocluster check aggregate show`
View the LIF placement failures in the MetroCluster configuration.	`metrocluster check lif show`

Commands for monitoring the MetroCluster interconnect

If you want to…	Use this command…
Display the HA and DR mirroring status and information for the MetroCluster nodes in the cluster.	`metrocluster interconnect mirror show`

Commands for monitoring MetroCluster SVMs

If you want to…	Use this command…
View all SVMs in both sites in the MetroCluster configuration.	`metrocluster vserver show`

Using the MetroCluster Tiebreaker or ONTAP Mediator to monitor the configuration

See Differences between ONTAP Mediator and MetroCluster Tiebreaker to understand the differences between these two methods of monitoring your MetroCluster configuration and initiating an automatic switchover.

Use these links to install and configure Tiebreaker or Mediator:

- Install and configure the MetroCluster Tiebreaker software
- xref:./manage/../install-ip/concept_mediator_requirements.html

How the NetApp MetroCluster Tiebreaker software detects failures

The Tiebreaker software resides on a Linux host. You need the Tiebreaker software only if you want to monitor two clusters and the connectivity status between them from a third site. Doing so enables each partner in a cluster to distinguish between an ISL failure, when inter-site links are down, from a site failure.

After you install the Tiebreaker software on a Linux host, you can configure the clusters in a MetroCluster configuration to monitor for disaster conditions.

How the Tiebreaker software detects intersite connectivity failures

The MetroCluster Tiebreaker software alerts you if all connectivity between the sites is lost.

Types of network paths

Depending on the configuration, there are three types of network paths between the two clusters in a MetroCluster configuration:

- **FC network (present in fabric-attached MetroCluster configurations)**

  This type of network is composed of two redundant FC switch fabrics. Each switch fabric has two FC switches, with one switch of each switch fabric co-located with a cluster. Each cluster has two FC switches, one from each switch fabric. All of the nodes have FC (NV interconnect and FCP initiator) connectivity to each of the co-located IP switches. Data is replicated from cluster to cluster over the ISL.

- **Intercluster peering network**

This type of network is composed of a redundant IP network path between the two clusters. The cluster peering network provides the connectivity that is required to mirror the storage virtual machine (SVM) configuration. The configuration of all of the SVMs on one cluster is mirrored by the partner cluster.

- **IP network (present in MetroCluster IP configurations)**

  This type of network is composed of two redundant IP switch networks. Each network has two IP switches, with one switch of each switch fabric co-located with a cluster. Each cluster has two IP switches, one from each switch fabric. All of the nodes have connectivity to each of the co-located FC switches. Data is replicated from cluster to cluster over the ISL.

Monitoring intersite connectivity

The Tiebreaker software regularly retrieves the status of intersite connectivity from the nodes. If NV interconnect connectivity is lost and the intercluster peering does not respond to pings, then the clusters assume that the sites are isolated and the Tiebreaker software triggers an alert as "AllLinksSevered". If a cluster identifies the "AllLinksSevered" status and the other cluster is not reachable through the network, then the Tiebreaker software triggers an alert as "disaster".

How the Tiebreaker software detects site failures

The NetApp MetroCluster Tiebreaker software checks the reachability of the nodes in a MetroCluster configuration and the cluster to determine whether a site failure has occurred. The Tiebreaker software also triggers an alert under certain conditions.

Components monitored by the Tiebreaker software

The Tiebreaker software monitors each controller in the MetroCluster configuration by establishing redundant connections through multiple paths to a node management LIF and to the cluster management LIF, both hosted on the IP network.

The Tiebreaker software monitors the following components in the MetroCluster configuration:

- Nodes through local node interfaces
- Cluster through the cluster-designated interfaces
- Surviving cluster to evaluate whether it has connectivity to the disaster site (NV interconnect, storage, and intercluster peering)

When there is a loss of connection between the Tiebreaker software and all of the nodes in the cluster and to the cluster itself, the cluster will be declared as “not reachable” by the Tiebreaker software. It takes around three to five seconds to detect a connection failure. If a cluster is unreachable from the Tiebreaker software, the surviving cluster (the cluster that is still reachable) must indicate that all of the links to the partner cluster are severed before the Tiebreaker software triggers an alert.

All of the links are severed if the surviving cluster can no longer communicate with the cluster at the disaster site through FC (NV interconnect and storage) and intercluster peering.

Failure scenarios during which Tiebreaker software triggers an alert

The Tiebreaker software triggers an alert when the cluster (all of the nodes) at the disaster site is down or unreachable and the cluster at the surviving site indicates the "AllLinksSevered" status.

The Tiebreaker software does not trigger an alert (or the alert is vetoed) in the following scenarios:

- In an eight-node MetroCluster configuration, if one HA pair at the disaster site is down
- In a cluster with all of the nodes at the disaster site down, one HA pair at the surviving site down, and the cluster at the surviving site indicates the "AllLinksSevered" status

  The Tiebreaker software triggers an alert, but ONTAP vetoes that alert. In this situation, a manual switchover is also vetoed

- Any scenario in which the Tiebreaker software can either reach at least one node or the cluster interface at the disaster site, or the surviving site still can reach either node at the disaster site through either FC (NV interconnect and storage) or intercluster peering

How the ONTAP Mediator supports automatic unplanned switchover

The ONTAP Mediator stores state information about the MetroCluster nodes in mailboxes located on the Mediator host. The MetroCluster nodes can use this information to monitor the state of their DR partners and implement a Mediator-assisted automatic unplanned switchover (MAUSO) in the case of a disaster.

When a node detects a site failure requiring a switchover, it takes steps to confirm that the switchover is appropriate and, if so, performs the switchover.

MAUSO is only initiated if both SyncMirror mirroring and DR mirroring of each node's nonvolatile cache is operating and the caches and mirrors are synchronized at the time of the failure.

Monitoring and protecting the file system consistency using NVFAIL

The `-nvfail` parameter of the `volume modify` command enables ONTAP to detect nonvolatile RAM (NVRAM) inconsistencies when the system is booting or after a switchover operation. It also warns you and protects the system against data access and modification until the volume can be manually recovered.

If ONTAP detects any problems, database or file system instances stop responding or shut down. ONTAP then sends error messages to the console to alert you to check the state of the database or file system. You can enable NVFAIL to warn database administrators of NVRAM inconsistencies among clustered nodes that can compromise database validity.

After the NVRAM data loss during failover or boot recovery, NFS clients cannot access data from any of the nodes until the NVFAIL state is cleared. CIFS clients are unaffected.

How NVFAIL impacts access to NFS volumes or LUNs

The NVFAIL state is set when ONTAP detects NVRAM errors when booting, when a MetroCluster switchover operation occurs, or during an HA takeover operation if the NVFAIL option is set on the volume. If no errors are detected at startup, the file service is started normally. However, if NVRAM errors are detected or NVFAIL processing is enforced on a disaster switchover, ONTAP stops database instances from responding.

When you enable the NVFAIL option, one of the processes described in the following table takes place during bootup:

If…	Then…

ONTAP detects no NVRAM errors	File service starts normally.
ONTAP detects NVRAM errors	• ONTAP returns a stale file handle (ESTALE) error to NFS clients trying to access the database, causing the application to stop responding, crash, or shut down. ONTAP then sends an error message to the system console and log file. • When the application restarts, files are available to CIFS clients even if you have not verified that they are valid. For NFS clients, files remain inaccessible until you reset the `in-nvfailed-state` option on the affected volume.
If one of the following parameters is used: • `dr-force-nvfail` volume option is set • `force-nvfail-all` switchover command option is set.	You can unset the `dr-force-nvfail` option after the switchover, if the administrator is not expecting to force NVFAIL processing for possible future disaster switchover operations. For NFS clients, files remain inaccessible until you reset the `in-nvfailed-state` option on the affected volume. Using the `force-nvfail-all` option causes the `dr-force-nvfail` option to be set on all of the DR volumes processed during the disaster switchover.
ONTAP detects NVRAM errors on a volume that contains LUNs	LUNs in that volume are brought offline. The `in-nvfailed-state` option on the volume must be cleared, and the NVFAIL attribute on the LUNs must be cleared by bringing each LUN in the affected volume online. You can perform the steps to check the integrity of the LUNs and recover the LUN from a Snapshot copy or backup as necessary. After all of the LUNs in the volume are recovered, the `in-nvfailed-state` option on the affected volume is cleared.

Commands for monitoring data loss events

If you enable the NVFAIL option, you receive notification when a system crash caused by NVRAM inconsistencies or a MetroCluster switchover occurs.

By default, the NVFAIL parameter is not enabled.

If you want to…	Use this command…

Create a new volume with NVFAIL enabled	`volume create -nvfail on`
Enable NVFAIL on an existing volume	`volume modify` **Note:** You set the `-nvfail` option to "on" to enable NVFAIL on the created volume.
Display whether NVFAIL is currently enabled for a specified volume	`volume show` **Note:** You set the `-fields` parameter to "nvfail" to display the NVFAIL attribute for a specified volume.

Related information

See the man page for each command for more information.

Accessing volumes in NVFAIL state after a switchover

After a switchover, you must clear the NVFAIL state by resetting the `-in-nvfailed-state` parameter of the `volume modify` command to remove the restriction of clients to access data.

Before you begin

The database or file system must not be running or trying to access the affected volume.

About this task

Setting `-in-nvfailed-state` parameter requires advanced-level privilege.

Step

1. Recover the volume by using the volume modify command with the -in-nvfailed-state parameter set to false.

After you finish

For instructions about examining database file validity, see the documentation for your specific database software.

If your database uses LUNs, review the steps to make the LUNs accessible to the host after an NVRAM failure.

Related information

Monitoring and protecting the files system consistency using NVFAIL

Recovering LUNs in NVFAIL states after switchover

After a switchover, the host no longer has access to data on the LUNs that are in NVFAIL states. You must perform a number of actions before the database has access to the LUNs.

Before you begin

The database must not be running.

Steps

1. Clear the NVFAIL state on the affect volume that hosts the LUNs by resetting the `-in-nvfailed-state` parameter of the `volume modify` command.
2. Bring the affected LUNs online.
3. Examine the LUNs for any data inconsistencies and resolve them.

   This might involve host-based recovery or recovery done on the storage controller using SnapRestore.

4. Bring the database application online after recovering the LUNs.

Where to find additional information

You can learn more about MetroCluster configuration and operation.

MetroCluster and miscellaneous information

Information	Subject
MetroCluster Documentation	• All MetroCluster information
NetApp Technical Report 4375: NetApp MetroCluster for ONTAP 9.3	• A technical overview of the MetroCluster configuration and operation. • Best practices for MetroCluster configuration.
Fabric-attached MetroCluster installation and configuration	• Fabric-attached MetroCluster architecture • Cabling the configuration • Configuring the FC-to-SAS bridges • Configuring the FC switches • Configuring the MetroCluster in ONTAP
Stretch MetroCluster installation and configuration	• Stretch MetroCluster architecture • Cabling the configuration • Configuring the FC-to-SAS bridges • Configuring the MetroCluster in ONTAP
MetroCluster IP installation and configuration	• MetroCluster IP architecture • Cabling the configuration • Configuring the MetroCluster in ONTAP
MetroCluster Tiebreaker 1.21 software installation and configuration	• Monitoring the MetroCluster configuration with the MetroCluster Tiebreaker software
Active IQ Unified Manager documentation NetApp Documentation: Product Guides and Resources	• Monitoring the MetroCluster configuration and performance

Copy-based transition	• Transitioning data from 7-Mode storage systems to clustered storage systems

Copyright information

Trademark information