Recovering from a non-controller failure

ONTAP MetroCluster

netapp-thomi, ntap-bmegan, zachary wambold

April 28, 2021

This PDF was generated from https://docs.netapp.com/us-en/ontap-metrocluster/disaster-recovery/task_recover_from_a_non_controller_failure_mcc_dr.html on May 31, 2021. Always check docs.netapp.com for the latest.

Table of Contents

Recovering from a non-controller failure

After the equipment at the disaster site has undergone any required maintenance or replacement, but no controllers were replaced, you can begin the process of returning the MetroCluster configuration to a fully redundant state. This includes healing the configuration (first the data aggregates and then the root aggregates) and performing the switchback operation.

- All MetroCluster hardware in the disaster cluster must be functional.
- The overall MetroCluster configuration must be in switchover.
- In a fabric-attached MetroCluster configuration, the ISL must be up and operating between the MetroCluster sites.

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

You must heal the data aggregates after repairing and replacing any hardware on the disaster site. This process resynchronizes the data aggregates and prepares the (now repaired) disaster site for normal operation. You must heal the data aggregates prior to healing the root aggregates.

The following example shows a forced switchover, where you bring the switched-over aggregate online. 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.

1. Verify that switchover was 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 a disaster

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.

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

   ```
   mcc1A::> 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. Ensure that the heal operation is complete by running the metrocluster operation show command on the destination cluster:

   ```
   mcc1A::> 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. Power up each controller module on the disaster site.
4. After nodes are booted, verify that the root aggregates are mirrored.

   If both plexes are present, any resynchronization will start automatically. If one plex has failed, that plex must be destroyed and the mirror recreated using the storage aggregate mirror -aggregateaggregate-name command to reestablish the mirror relationship.

Verifying that your system is ready for a switchback

If your system is already in the switchover state, you can use the -simulate option to preview the results of a switchback operation.

1. Simulate the switchback operation:

   a. From either surviving node's prompt, change to the advanced privilege level: `set -privilege advanced`

      You need to respond with `y` when prompted to continue into advanced mode and see the advanced mode prompt (*>).

   b. Perform the switchback operation with the -simulate parameter: `metrocluster switchback -simulate`

   c. Return to the admin privilege level: `set -privilege admin`

2. Review the output that is returned.

   The output shows whether the switchback operation would run into errors.

Example of verification results

The following example shows the successful verification of a switchback operation:

```
cluster4::*> metrocluster switchback -simulate
  (metrocluster switchback)
[Job 130] Setting up the nodes and cluster components for the switchback
operation...DBG:backup_api.c:327:backup_nso_sb_vetocheck : MetroCluster
Switch Back
[Job 130] Job succeeded: Switchback simulation is successful.

cluster4::*> metrocluster op show
  (metrocluster operation show)
  Operation: switchback-simulate
      State: successful
 Start Time: 5/15/2014 16:14:34
   End Time: 5/15/2014 16:15:04
     Errors: -

cluster4::*> job show -name Me*
                            Owning
Job ID Name                 Vserver    Node           State
------ -------------------- ---------- -------------- ----------
130    MetroCluster Switchback
                            cluster4
                                       cluster4-01
                                                      Success
       Description: MetroCluster Switchback Job - Simulation
```

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 the switchback by running the metrocluster switchback command from any node in the surviving cluster.
5. 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.

6. 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

node_A_1::> aggr show
Aggregate     Size Available Used% State   #Vols  Nodes            RAID
Status
--------- -------- --------- ----- ------- ------ ----------------
------------
...
aggr_b2          -         -     - unknown      - node_A_1
```

If the disaster site included unmirrored aggregates and the unmirrored aggregates are no longer present, the aggregate may show up with a State of unknown in the output of the storage aggregate show command. Contact technical support to remove the out-of-date entries for the unmirrored aggregates.

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.

Deleting stale aggregate listings after switchback

In some circumstances after switchback, you might notice the presence of stale aggregates. Stale aggregates are aggregates that have been removed from ONTAP, but whose information remains recorded on disk. Stale aggregates are displayed in the nodeshell aggr status -r command but not in the storage aggregate show command. You can delete these records so that they no longer appear.

Stale aggregates can occur if you relocated aggregates while the MetroCluster configuration was in switchover. For example:

1. Site A switches over to Site B.
2. You delete the mirroring for an aggregate and relocate the aggregate from node_B_1 to node_B_2 for load balancing.
3. You perform aggregate healing.

At this point a stale aggregate appears on node_B_1, even though the actual aggregate has been deleted from that node. This aggregate appears in the output from the nodeshell aggr status -r command. It does not appear in the output of the storage aggregate show command.

1. Compare the output of the output of the storage aggregate show command and the nodeshell aggr status -r command: ```storage aggregate show``run local aggr status -r```

   Stale aggregates appear in the run local aggr status -r output but not in the storage aggregate show output. For example, the following aggregate might appear in the run local aggr status -r output:

```
Aggregate aggr05 (failed, raid_dp, partial) (block checksums)
Plex /aggr05/plex0 (offline, failed, inactive)
  RAID group /myaggr/plex0/rg0 (partial, block checksums)

 RAID Disk Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)
Phys (MB/blks)
 --------- ------  ------------- ---- ---- ----  ----- --------------
--------------
 dparity   FAILED          N/A                        82/ -
 parity    0b.5    0b    -   -   SA:A   0 VMDISK  N/A 82/169472
88/182040
 data      FAILED          N/A                        82/ -
 data      FAILED          N/A                        82/ -
 data      FAILED          N/A                        82/ -
 data      FAILED          N/A                        82/ -
 data      FAILED          N/A                        82/ -
 data      FAILED          N/A                        82/ -
 Raid group is missing 7 disks.
```

2. Remove the stale aggregate:
   a. From either node's prompt, change to the advanced privilege level: `set -privilege advanced`

    You need to respond with `y` when prompted to continue into advanced mode and see the advanced mode prompt (*>).

    b. Remove the stale aggregate: `aggregate remove-stale-record -aggregate aggregate_name`

    c. Return to the admin privilege level: `set -privilege admin`

3. Confirm that the stale aggregate record was removed: `run local aggr status -r`

Copyright Information

Copyright © 2021 NetApp, Inc. All rights reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means-graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system- without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

RESTRICTED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.277-7103 (October 1988) and FAR 52-227-19 (June 1987).

Trademark Information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.