

How to non-disruptively create a new root aggregate and have it host the root volume in clustered Data ONTAP 8.2 and 8.3, ONTAP 9.0

Document ID

Document ID HO454

Answer ID

Answer ID 1030179

Published Date

Published Date 08/27/2018

Views

Views 1000

Ratings

Ratings 4

Categories

Categories ONTAP 9,ONTAP 9.1,ONTAP 9.2,Clustered Data ONTAP 8.2,Clustered Data ONTAP 8.3,ONTAP 9.3

Description

This article describes the procedure that should be followed to create a new root aggregate non-disruptively and have it host the root volume in clustered Data ONTAP 8.2 and 8.3 and ONTAP 9.0.

Data ONTAP 8.3 and later introduced a new feature called Advanced Disk Partitioning. This feature allows the root aggregate to be hosted on a thin partition using disks that are shared for data aggregates. On system initialization, 12-24 disks are divided into a large 'P1' partition for data and into a small 'P2' partition for the root. Only re-initializing each node (option 4) will place the root aggregate onto shared disks and it is not possible to do this non-disruptively. However, it is possible to non-disruptively move the root volume off of shared disks and onto normal un-partitioned disks located on an external disk shelf using this KB, but this will reduce storage efficiently and is not recommended unless advised by Technical Support. For more information on how to configure the root aggregate to use shared disks using Advanced Disk Partitioning (ADP) on Data ONTAP 8.3, see article [1031828](#): How to upgrade and re-initialize FAS25xx, FAS22xx, and All-Flash FAS platforms to use the Advanced Drive Partitioning feature in Data ONTAP 8.3 .

ONTAP 9.0 and later introduced a new CLI command to migrate root onto new disks. In the background, this command will automatically follow the same steps as the KB article. You can find the ONTAP 9.0 and later 'system node migrate-root' command syntax at the end of the KB. If the command fails, Technical Support should be engaged to perform manual cleanup.

Review the following before continuing:

There are many steps to ensure that maintenance is non-disruptive, and it is critical all steps are reviewed and not deviated from.

Due to the possible impact of a system that is down for the duration of this maintenance on other nodes in the cluster, the first steps are somewhat different for 2-node, 4-node, or larger clusters, and exist to protect data access in the rest of the cluster. These steps must be followed prior to performing activity on each node.

- Although not required, it is recommended to upgrade Data ONTAP to 8.2.1P1, 8.2.2 (or later release) before doing this maintenance to avoid BUG [810014](#). If upgrading Data ONTAP is not possible, avoid moving the root aggregate more than one time; multiple moves increase the likelihood of encountering the condition of this bug. Steps 11 and 12 below provide checks on whether this bug is encountered and Technical Support has an internal workaround for this issue.
Check the administration documentation on size requirements for the root volume for the platform you are using. If the new root aggregate drives you are using are smaller in size, you might need to have a root aggregate of more than 3 disks to accommodate for the space.
- In the example below, root aggregate `aggr0_node1` on `node01` is going to be hosted on 3 manually specified drives.
- OnCommand System Manager (OCSM), `spi`, `offbox viruscanfpolicy` will not work after the root vol is moved to a new aggregate for that node. Creating and applying a new ssl certificate will resolve this. To create and apply a new certificate, see article [1032196](#): How to renew an SSL certificate in clustered Data ONTAP
- Determine CPU and disk utilization
 - `node run -node node_name -command sysstat -c 10 -x 3`
 - You should monitor CPU and disk utilization for 30 seconds. The values in the CPU and Disk Util columns should not exceed 50% for all 10 measurements reported. No additional load should be added to the cluster until the upgrade is complete.
- If working on multiple nodes, the procedure must be completed start to finish on each individual node. In an HA pair, complete all steps on the first node before performing action on the second node. Pay close attention to move Epsilon off of the node that is being working on (Step 1). This can and will cause an outage if not properly done.
- **If you are running a cluster with only a single node**, see article [1032066](#): How to move mroot to a new root aggregate in a single node cluster. Single node cluster root volume migration cannot be done non-disruptively

Procedure

ONTAP 9

Run the following command:

```
::*> system node migrate-root
```

Availability: This command is available to cluster administrators at the advanced privilege level.

Description: The `system node migrate-root` command migrates the root aggregate of a node to a different set of disks. You need to specify the node name and the list of disks on which the new root aggregate will be created. The command starts a job that backs up the node configuration, creates a new aggregate, set it as new root aggregate, restores the node configuration and restores the names of original aggregate and volume. The job might take as long as a few hours depending on time it takes for zeroing the disks, rebooting the node and restoring the node configuration.

Parameters:

-node {<nodename>|local} - Node Specifies the node that owns the root aggregate that you wish to migrate. The value local specifies the current node.

-disklist <disk path name>, ... - List of Disks for New Root Aggregate Specifies the list of disks on which the new root aggregate will be created. All disks must be spares and owned by the same node. Minimum number of disks required is dependent on the RAID type.

-raid-type {raid_tec|raid_dp|raid4} - RAID Type for the New Root Aggregate Specifies the RAID type of the root aggregate. The default value is raid-dp.

Example:

```
::> system node migrate-root -node node1 -disklist 1.11.8, 1.11.9, 1.11.10, 1.11.11, 1.11.12 -raid-type raid-dp
```

Data ONTAP 8.2 and 8.3

Perform the following steps to create the new root aggregate and have it host the new root volume:

Warning: The following steps marked in green are very important to be performed first. If the steps for 2-node or 4-node clusters are not performed, you will risk an outage of all nodes in the cluster, and will need to contact Support to get data access recovered, this can possibly take hours. There have been users who have skipped these pre-steps and have experienced real outages as a result.

1. **If you are running a 2-node cluster**, make sure to disable cluster HA prior to maintenance (this will not disable failover but will change quorum voting for 2-node clusters) and make sure the node that will not be changed (the partner node of the system being worked on) is set to be epsilon.

```
::> cluster ha modify -configured false
::> set adv
::*> cluster modify -node node01 -epsilon false
::*> cluster modify -node node02 -epsilon true
::*> cluster show
```

The commands above will allow halting node01 to maintenance mode without a takeover (required for some of the steps later on), and prevent node02 from going out of quorum as a result. Prior to this halt, relevant storage will be moved to node02, so it can continue to be served in the steps below. If above steps are not followed in a 2-node cluster, the surviving node hosting all the storage will not serve any of it or its partner's data and there will be an outage of all data on both nodes of the 2-node cluster.

If you are running a 4-node cluster, run the following commands prior to maintenance. Check if the node you are working on is epsilon; if it is, move epsilon to a different node in the cluster to reduce the risk of a cluster quorum loss in case of an unexpected failover in other parts of the cluster. Also mark the node as ineligible so it will not count towards cluster quorum:

```
::> set adv
::*> cluster show
::*> cluster modify -node node01 -epsilon false
::*> cluster modify -node node03 -epsilon true
::*> system node modify -node node01 -eligibility false
```

Important: Verify epsilon has moved off the node that will be worked on before proceeding to Step 2.

```
::*> cluster show
```

If you are running a cluster with more than 4 nodes, no additional protection is necessary besides marking the node as ineligible:

```
::> set adv
::*> system node modify -node node01 -eligibility false
```

2. Relocate the data aggregates on the node you want to change, and move them to the partner. Include all SFO aggregates that have data volumes on them.
Aggregate relocate will cause the relevant aggregates to be reassigned to the partner. Any volumes in aggregates not relocated will be inaccessible during the maintenance, so be sure to verify afterwards if all aggregates with user data are moved. Once maintenance is done, be sure to move the aggregates back.

```
::> aggr relocation start -node node01 -destination node02 -aggregate-list
aggr1_node1,aggr2_node1
```

3. Once you have successfully moved all data aggregates, only the root aggregate and at least 3 spare drives should remain on the maintenance node.
Run the following command to verify:

```
::> aggr show -node node01 -fields has-mroot
aggregate    has-mroot
-----
aggr0_node1  true
```

4. Create the new root aggregate with 3 disks, specifying the exact drives:

```
::*> aggr create -aggregate newroot -disklist
node01:0a.11.16,node01:0a.11.19,node01:0a.11.22 -force-small-aggregate true
---->Notice the requirement for the -force-small-aggregate true flag due to the fact that there are
only three drives. This is an advanced level option.
```

5. Migrate all LIFs on the relevant node to other nodes in the cluster (making sure failover groups are configured correctly):

```
::*> net int migrate-all -node node01
```

6. Verify that the aggregate is fully created. If the drives used for the new aggregate were not zeroed beforehand, the aggregate creation might take some time. Do not start the next step until after the newly created root aggregate is fully ready.
7. You can verify the status with `sysconfig -r` in the nodeshell of the relevant node. Again, this might take a few hours if no zeroed disks were available for disk creation:

```
::*> run -node node01 sysconfig -r
```

8. Verify all of the above steps have been done correctly before performing the steps below. Once you are confident of quorum settings (epsilon and cluster ha) are correct, the data aggregates and LIFs are moved to the right nodes using `net int migrate-all`, `aggregate relocate` respectively. After the new root aggregate is fully created and ready, then reboot the node without takeover with the following command, and go to Maintenance Mode.

```
::*> reboot -node node01 -inhibit-takeover true
```

```
md2.uzip: 7360 x 16384 blocks
*****
*                                     *
* Press Ctrl-C for Boot Menu. *
*                                     *
*****
      ^C^C^C^C          <----- Ctrl-C
Boot Menu will be available.
```

Select one of the following:

- (1) Normal Boot.
- (2) Boot without /etc/rc.
- (3) Change password.
- (4) Clean configuration and initialize all disks.
- (5) Maintenance mode boot.
- (6) Update flash from backup config.
- (7) Install new software first.
- (8) Reboot node.

Selection (1-8)? 5 <----- option 5

9. Set the new aggregate to CFO, which allows it to become the new root aggregate. Then set the root flag, which will make it the new root aggregate. After booting, this aggregate will automatically have a new root volume pre-created. This new root volume is called AUTOROOT. (or AUTOROOT-1 if a volume with that name already exists)

```
*> aggr options newroot ha_policy cfo
Setting ha_policy to cfo will substantially increase the client outage during
giveback for volumes on aggregate "newroot".
Are you sure you want to proceed (y/n)? y
*> aggr options newroot root
Aggregate 'newroot' will become root at the next boot.
```

Bring the system back up so you can clear the recovery environment variable:

```
*> halt
LOADER-A> boot_ontap
```

10. The system will now boot with a newly created skeleton root volume. Based on the data stored in the cf card and NVRAM, the node knows its identity in the cluster. Because the node previously had cluster database data in its root volume, and this volume is now empty, the system will set a recovery flag and give a warning at boot:

```
Sep 18 17:00:33 [node01:mgmtgwd.rootvol.recovery.new:EMERGENCY]: A new root volume
was detected. This node is not fully operational. Contact technical support to
obtain the root volume recovery procedures.
Sep 18 17:00:33 [node01:callhome.root.vol.recovery.reqd:EMERGENCY]: Call home for
ROOT VOLUME NOT WORKING PROPERLY: RECOVERY REQUIRED.
Wed Sep 18 17:00:35 EST 2013
login: admin
Password:
*****
* This is a serial console session. Output from this *
* session is mirrored on the SP/RLM console session. *
*****
*****
** SYSTEM MESSAGES **
*****
A new root volume was detected. This node is not fully
```

operational. Contact support personnel for the root volume recovery procedures.

11. To reset the recovery flag and let the node synchronize its cluster database with the rest of the cluster, halt the system:

```
::*> halt -node node01 -inhibit-takeover true
(system node halt)
```

12. Reset the recovery flag at the loader prompt and boot the node back up.

```
LOADER-A*> unsetenv bootarg.init.boot_recovery
LOADER-A*> boot_ontap
```

The node should boot normally, without the recovery warning this time. It will take a few seconds at the end of the logging before the login prompt is visible while the node is synchronizing the cluster database.

13. **If the cluster is 4 nodes or greater and the node was previously set ineligible**, log into another node in the cluster and declare the node eligible. If this step is not followed, cluster services will not come online.

```
::> set adv
::*> cluster modify -node node01 -eligibility true
```

Return to the node where the root volume is being moved and continue the process.

14. Check for the health of the node with `cluster ring show`; all rings should show numbers.

```
::*> set adv
Warning: These advanced commands are potentially dangerous; use them only when
directed to do so by NetApp personnel.
Do you want to continue? {y|n}: y
```

```
::*> cluster ring show
```

Node	UnitName	Epoch	DB Epoch	DB Trnxs	Master	Online
node01	mgmt	22	22	2971287	node01	master
node01	vldb	20	20	1	node01	master
node01	vifmgr	20	20	932	node01	master
node01	bcomd	20	20	1	node01	master
node01	crs	4	4	2	node01	master
node02	mgmt	22	22	2971287	node01	secondary
node02	vldb	20	20	1	node01	secondary
node02	vifmgr	20	20	932	node01	secondary
node02	bcomd	20	20	1	node01	secondary

Warning: If there are dashes or RPC errors for any of the rings shown instead of numbers in the 3 columns, wait for at least 10 more minutes and check `cluster ring show` again. If after waiting for 20 minutes after logging dashes or errors still show for some of the rows, contact NetApp support before taking further action.

15. **If you are running a 2-node cluster**, run the following to re-enable HA after all rings show numbers:

```
::> cluster ha modify -configured true
```

Check the health of the HA relationship:

```
::> storage failover show
```

Node	Partner	Takeover Possible	State Description
node01	node02	true	Connected to node02
node02	node01	true	Connected to node01

2 entries were displayed.

```
::> set diag
```

Warning: These diagnostic commands are for use by NetApp personnel only.

Do you want to continue? {y|n}: y

```
::*> cluster ha show
```

High Availability Configured: true

High Availability Backend Configured (MBX): true

Warning: If you see false after enabling HA, contact NetApp support before taking further action.

Important Additional clean-up steps:

1. With the system back and in fully redundant state, you can now delete the old volume and aggregate. The old root volume is likely called vol0 (the default) but might be called something else. Check aggr status in the nodeshell to see what volume resides in the old root aggregate. In this example, it is called vol0.

```
::*> run -node node01
```

```
node01> vol offline vol0
```

Volume 'vol0' is now offline.

```
node01> vol destroy vol0
```

Are you sure you want to destroy volume 'vol0'? y

Volume 'vol0' destroyed.

2. Return to the cluster-shell to delete the old root aggregate:

```
::*> aggr delete -aggregate aggr0_node1
```

Warning: Are you sure you want to destroy aggregate "aggr0_node1"?

{y|n}: y

[Job 110] Job succeeded: DONE

3. When moving the root volume, some of the volume and aggregate changes are done outside of the knowledge of the Volume Location Database (VLDB). As a result, it is important to make sure that the VLDB is modified to know about the aggregate and volume changes made during this maintenance. To make the VLDB aware of the changes, run the following diag level commands:

```
::*> set diag
```

```
::*> volume remove-other-volume -volume vol0 -vserver node01
```

```
::*> volume add-other-volumes -node node01
```

Verify the correctness of the VLDB with the following diag level command:

```
::*> debug vreport show
```

This table is currently empty.

Info: WAFL and VLDB volume/aggregate records are consistent.

If the above message is displayed, it is confirmed that there are no issues reported.

4. Move back the relocated aggregates as there are no issues:

```
::> aggr relocation start -node node02 -destination node01 -aggregate-list  
aggr1_node1,aggr2_node1
```

5. Add the NVFAIL option to the new root vol.

```
::*> node run -node node01 vol options AUTOROOT nvfail on  
::*> volume show -volume AUTOROOT -fields nvfail  
vserver      volume    nvfail  
-----  
node01      AUTOROOT on
```

6. Rename the new root volume and new aggregate to the name before. The new root volume is most likely named AUTOROOT and it can be changed if desired.

In this example, AUTOROOT volume is renamed to vol0 and the aggregate created as newroot is renamed as aggr0.

```
::*> vol rename -volume AUTOROOT -newname vol0 -vserver node01  
(volume rename)  
[Job 111] Job succeeded: Successful  
::*> aggr rename newroot -newname aggr0_node1  
[Job 112] Job succeeded: DONE
```

7. Revert all LIF(s) back to their home port(s):

```
::> network interface revert -vserver * -lif *
```

8. There are size restrictions for the root volume. Ensure the root volume size is still correct and adheres to the size requirements as stated in the Data ONTAP System Administration Guide for your release. You might need to increase the size of the volume. If the new aggregate consists of smaller drives than the drives used before, it might need an additional disk to hold the space required.

Note: OnCommand System Manager (OCSM), spi, offbox viruscanfpolicy will not work after the root vol is moved to a new aggregate for that node. Creating and applying a new ssl certificate will resolve this. Also note that each SVM certificate under the new root aggr/volume will need to be re-created. To create and apply a new certificate, see article [1032196](#): How to renew an SSL certificate in clustered Data ONTAP

Related Bugs:

- [976779](#) - migrate-root operation fails with an error message: 'Internal error: Failed to verify the new root aggregate status
- [1025006](#) - migrate-root operation fails when copying contents from the old root volume
- [1026948](#) - migrate-root operation fails with an internal error message

Note: If you are unable to view the entire content of this article please log in to [kb.netapp.com](#)

NetApp provides no representations or warranties regarding the accuracy or reliability or serviceability of any information or recommendations provided in this publication or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.