

**NetApp Solution Deployment Guidelines**

**Thomson Reuters – cDOT Base Configuration**

**Updated for cDOT 8.3**

**Synopsis:** This document details the NetApp clustered Data ONTAP (cDOT) base configuration used for various storage solutions within Thomson Reuters.

**Authors:** Ken Zola (zola@netapp.com)

**Contributors:** Ian Daniel, Joel Edstrom

**Document Version:** V13

**Date:** April 2017

**CONFIDENTIAL INFORMATION**

This document contains information proprietary to Thomson Reuters and may not be reproduced, disclosed or used in whole or part without express permission of Thomson Reuters.

© Thomson Reuters 2014

Contents

[1 Introduction 4](#_Toc479335711)

[1.1 Management Summary 4](#_Toc479335712)

[1.2 Assumptions 4](#_Toc479335713)

[1.3 Change History 4](#_Toc479335714)

[1.4 Initial Distribution List 5](#_Toc479335715)

[1.5 Glossary 5](#_Toc479335716)

[2 Networking Standards 6](#_Toc479335717)

[2.1 Networking 6](#_Toc479335718)

[2.1.1 Required network cabling 6](#_Toc479335719)

[2.1.2 Required IP addresses 6](#_Toc479335720)

[3 Cluster Interconnect Configuration 7](#_Toc479335721)

[3.1 Default network ports by system model 7](#_Toc479335722)

[3.2 Single node clusters 7](#_Toc479335723)

[3.3 Two node switchless clusters 7](#_Toc479335724)

[3.4 CN1610 switched clusters 8](#_Toc479335725)

[3.5 Cluster interconnect failure scenarios, quorum, and epsilon 9](#_Toc479335726)

[4 Storage System Configuration 10](#_Toc479335727)

[4.1 Cluster setup 10](#_Toc479335728)

[4.1.1 Creating a new cluster 10](#_Toc479335729)

[4.1.2 Joining a cluster 10](#_Toc479335730)

[4.1.3 Two node switchless cluster considerations 10](#_Toc479335731)

[4.2 Aggregates 10](#_Toc479335732)

[4.2.1 Dedicated root aggregates 10](#_Toc479335733)

[4.2.2 Data aggregates 10](#_Toc479335734)

[4.3 Networking 11](#_Toc479335739)

[4.3.1 Physical network connections 11](#_Toc479335740)

[4.3.2 Logical interfaces (LIFs) 11](#_Toc479335741)

[4.3.3 IPspaces (ONTAP 8.3+) 12](#_Toc479335742)

[4.3.4 Broadcast Domains (ONTAP 8.3+) 12](#_Toc479335743)

[4.3.5 Subnets (ONTAP 8.3+) 13](#_Toc479335744)

[4.3.6 Routing 13](#_Toc479335745)

[4.3.7 Failover Groups 14](#_Toc479335746)

[4.3.8 External Management Services Access (ONTAP 8.3+) 14](#_Toc479335747)

[4.3.9 Upgrade from 8.2 to 8.3 LIF Requirement 15](#_Toc479335748)

[4.3.10 Upgrade from 8.2 to 8.3 Failover Groups Requirement 15](#_Toc479335749)

[4.4 Vservers 16](#_Toc479335750)

[4.4.1 Node Vservers 16](#_Toc479335751)

[4.4.2 Admin Vserver 16](#_Toc479335752)

[4.4.3 Data Vservers 16](#_Toc479335753)

[4.5 Miscellaneous configuration 17](#_Toc479335754)

[4.5.1 Vserver services 17](#_Toc479335755)

[4.5.2 Unlocking the diag user account 17](#_Toc479335756)

[4.5.3 The wlstats account 17](#_Toc479335757)

[4.5.4 CN1610 device monitoring 17](#_Toc479335758)

[4.5.5 System core file and log access via HTTP 17](#_Toc479335759)

[4.5.6 Vserver and Volume language setting 17](#_Toc479335760)

[4.6 IPspaces and Cluster Peering for SnapVault 18](#_Toc479335761)

[4.6.1 Intercluster LIF and peering relationship 18](#_Toc479335762)

[4.7 Compression and Dedupe Settings 18](#_Toc479335763)

[4.7.1 Inline Zero Detection, Inline Adaptive Compression, & Inline Dedupe 18](#_Toc479335764)

[4.7.2 Recommended Compression and Dedupe Settings for an AFF system 19](#_Toc479335765)

[4.7.3 Recommended Compression and Dedupe Settings for SnapVault secondary volumes of AFF primary volumes 19](#_Toc479335766)

[4.7.4 Recommended Compression and Dedupe Settings for non-AFF Primary Systems 19](#_Toc479335767)

[5 Single node cluster system backup configuration 20](#_Toc479335768)

[5.1 HTTP server configuration on a management server 20](#_Toc479335769)

[5.1.1 Configuration of a SLES Apache server 20](#_Toc479335770)

[5.1.2 Configuration of an OEL Apache server 20](#_Toc479335771)

[6 Clustered ONTAP configuration CLI examples 21](#_Toc479335772)

[6.1 CN1610 cluster switch configuration 21](#_Toc479335773)

[6.1.1 Set admin password and enable mode password 21](#_Toc479335774)

[6.1.2 Set management IP 21](#_Toc479335775)

[6.1.3 Configure SSH and disable telnet 21](#_Toc479335776)

[6.1.4 Set clock 21](#_Toc479335777)

[6.1.5 Set hostname 21](#_Toc479335778)

[6.1.6 Save configuration 21](#_Toc479335779)

[6.1.7 Check FastPath and RCF version 22](#_Toc479335780)

[6.1.8 Update RCF version 22](#_Toc479335781)

[6.1.9 Diagnostic commands 22](#_Toc479335782)

[6.2 Initial cluster setup 22](#_Toc479335783)

[6.2.1 Creating a new cluster on the first node 22](#_Toc479335784)

[6.2.2 Joining a cluster with additional nodes 22](#_Toc479335785)

[6.2.3 Configuration settings for two node switchless cluster 22](#_Toc479335786)

[6.2.4 Licensing 23](#_Toc479335787)

[6.3 Aggregate creation 23](#_Toc479335788)

[6.3.1 Root aggregate renaming 23](#_Toc479335789)

[6.3.2 Create a standard aggregate with free space reallocation 23](#_Toc479335790)

[6.3.3 Create a FlashPool aggregate with free space reallocation 23](#_Toc479335791)

[6.4 Network configuration 24](#_Toc479335792)

[6.4.1 Disabling Ethernet flowcontrol on cluster interconnect ports 24](#_Toc479335793)

[6.4.2 Disabling Ethernet flowcontrol on data ports 24](#_Toc479335794)

[6.4.3 Creating a LACP IFGRP with data ports 24](#_Toc479335795)

[6.4.4 Adding a VLAN tag to an IFGRP of data ports 24](#_Toc479335796)

[6.4.5 Enabling jumbo frames on an IFGRP of data ports or a VLAN tag 24](#_Toc479335797)

[6.4.6 Creating an IPspace (ONTAP 8.3+) 24](#_Toc479335798)

[6.4.7 Creating a Broadcast Domain (ONTAP 8.3+) 24](#_Toc479335799)

[6.4.8 Creating a failover group for an IFGRP 25](#_Toc479335800)

[6.4.9 Applying the failover group for management ports 25](#_Toc479335801)

[6.4.10 Create Intercluster LIF with default route 25](#_Toc479335802)

[6.4.11 Cluster Peering 25](#_Toc479335803)

[6.5 Miscellaneous admin Vserver configuration 25](#_Toc479335804)

[6.5.1 DNS 25](#_Toc479335805)

[6.5.2 Autosupport 25](#_Toc479335806)

[6.5.3 Timezone configuration 26](#_Toc479335807)

[6.5.4 NTP configuration 26](#_Toc479335808)

[6.5.5 SNMP configuration 26](#_Toc479335809)

[6.5.6 Unlocking the diag user 26](#_Toc479335810)

[6.5.7 Restricted account for wlstats 26](#_Toc479335811)

[6.5.8 SSH publickey authentication for cluster admin 26](#_Toc479335812)

[6.5.9 CN1610 cluster switch device monitoring commands 27](#_Toc479335813)

[6.5.10 System configuration backup settings 27](#_Toc479335814)

# Introduction

## Management Summary

This document details the NetApp clustered Data ONTAP (cDOT) base configuration used for various storage solutions within Thomson Reuters. The majority of the shared storage and dedicated cDOT solution deployments will be configured identically for a number of components, and those common configurations will be described in this document. A typical shared storage cDOT architecture is depicted below.



## Assumptions

It is assumed the person(s) reading this document are conversant with NetApp hardware and software. They will also be conversant with the Linux and Windows operating systems, NFS, iSCSI and CIFS protocols.

## Change History

|  |  |  |  |
| --- | --- | --- | --- |
| **Ver** | **Date** | **Author** | **Key Changes** |
| 1 | September 2013 | Michael Arndt | Initial Version |
| 2 | October 2013 | Michael Arndt | Added standards and naming conventions section, and other minor updates. |
| 3 | October 2013 | Michael Arndt | Minor corrections. |
| 4 | October 2013 | Michael Arndt | Added examples for CN1610 switch monitoring and RCF file updates. |
| 5 | November 2013 | Michael Arndt | Added IP address requirement for CN1610 switches, cleanup up admin Vserver terminology, configuration backup section for single node clusters, and cluster network connectivity details. |
| 6 | December 2013 | Michael Arndt | Minor grammatical and terminology cleanup. |
| 7 | May 2014 | Michael Arndt | Cluster interconnect ports should have flowcontrol off. Use free-space-realloc with no\_redirect for aggregates. |
| 8 | July 2014 | Ian Daniel | Updated management interfaces to cater for 8000 series. |
| 9 | August 2014 | Ian Daniel | Updated SNMP community settings for switch monitoring. |
| 10 | August 2014 | Ian Daniel | Updated SNMP community settings added note about changes in RCF |
| 11 | November 2015 | Ken Zola | Updated 1610 switch port information, cluster node count information, removed LIF and Aggregate naming references, and added Vserver and volume language requirements. |
| 12 | January 2017 | Ken Zola | Updated for ONTAP 8.3 |
| 13 | March 2017 | Ken Zola | Updated for compression and dedupe setting recommendations |

## Initial Distribution List

|  |  |
| --- | --- |
| **Name** | **Role** |
| Sridhar Chevendra | Customer |
| Stewart Bird | Customer |
| Ken Zola | Reviewer |
| Ian Daniel | Reviewer |
| Joel Edstrom | Reviewer |

## Glossary

|  |  |
| --- | --- |
| **Term** | **Definition** |
| cDOT | clustered Data ONTAP |
| Node | One physical storage controller in a cDOT system. |
| Cluster | A collection of one or more nodes that form a cDOT system. |
| **Vserver** | A logical storage virtual server, also known as a Storage Virtual Machine (SVM), which contains LIFs, Volumes, and configuration information. |
| Node Vserver | Each node in the cluster has a dedicated Vserver that can be used to manage the node. Each node Vserver has a dedicated node management LIF that does not failover to other nodes if that node is not operational. |
| Admin  Vserver | Each cluster has one cluster management Vserver called the admin Vserver. This Vserver is typically used to manage the entire cluster, and the dedicated LIF for this Vserver is configured to failover to other nodes in the cluster if the node currently hosting the admin Vserver LIF is not operational. |
| Data Vserver | Clients access data via standard protocols only on data Vservers. |
| **Port** | A physical network interface on a cDOT node. |
| Node Management Port | The ports used by administrators to connect to and manage a node. Note that the node management LIF can also reside on a data port. |
| Cluster Port | The ports used for intracluster traffic only. By default, each node has two cluster ports on 10-GbE ports enabled for jumbo frames. |
| Data Port | The ports used for data traffic. These ports are accessed by NFS, SMB (CIFS), FC, FCoE, and iSCSI clients for data requests. Each node has a minimum of one data port. |
| Intercluster Port | The ports used for cross-cluster communication. An Intercluster LIF can also reside on a data port. |
| **LIF** | Logical Interface – a cDOT logical network interface with an IP address, assigned to a single Vserver. |
| Node-management LIF | The LIF that provides a dedicated IP address for managing a particular node and gets created at the time of creating or joining the cluster. These LIFs are used for system maintenance, for example, when a node becomes inaccessible from the cluster. Node-management LIFs can be configured on either node-management or data ports. |
| Cluster-management LIF | The LIF that is used for intracluster traffic. Cluster LIFs can be configured only on cluster ports. Cluster LIFs must always be created on 10Gb network ports. |
| Cluster LIF | The LIF that is used for intracluster traffic. Cluster LIFs can be configured only on cluster ports. |
| Data LIF | The LIF that is associated with a Vserver and is used for communicating with clients. Data LIFs can be configured only on data ports. |
| Intercluster LIF | The LIF that is used for cross-cluster communication, backup, and replication. Intercluster LIFs can be configured on data ports or intercluster ports. |

# Networking Standards

## Networking

### Required network cabling

* Two 1Gb connections per node on the TR management network.
* Two 10Gb connections per node on the TR data network.
* Two 10Gb connections per node for to the private cluster switches, assuming that a switched cluster is being installed.
* Two 1Gb connections **per cluster** on the TR management network for the private cluster switches, assuming that a switched cluster is being installed.

### Required IP addresses

* One IP address per node on the 1Gb management network for the SP.
* One IP address per node on the 1Gb management network for node management.
* One IP address **per cluster** on the 1Gb management network for the admin Vserver.
* One IP address per node on the 10Gb data network for the Intercluster LIF, assuming SnapVault or SnapMirror replication is required.
* Two IP addresses **per cluster** on the 1Gb management network for the private cluster switches, assuming that a switched cluster is being installed.
* Additional IP addresses on the 10Gb data network for data Vservers.

# Cluster Interconnect Configuration

## Default network ports by system model

The cluster setup dialog will default to certain ports for the cluster and node management networks. While the cluster interconnect ports can be changed during cluster setup, it is recommended to use the default ports for the sake of consistency in the environment. The following table lists the default 10Gb network ports used for the cluster interconnect, by NetApp system model. The node management ports are also listed, but we will use the e0a or e0i port (which is a data port) for node management on all systems at Thomson Reuters.

| **Platform** | **Cluster ports** | **Node management port** | **Data ports** |
| --- | --- | --- | --- |
| 31xx,32xx | e1a, e2a | e0M (e0a data port used at TR) | All other Ethernet ports are data ports |
| 60xx | e5a, e6a | e0f (e0a data port used at TR) | All other Ethernet ports are data ports |
| 62xx | e0c, e0e | e0M (e0a data port used at TR) | All other Ethernet ports are data ports |
| 8040 | e0a, e0c | e0M (e0i data port used at TR) | All other Ethernet ports are data ports |
| 8080 | e0a, e0c  e0b & e0d (reserved for expansion) | e0M (e0i data port used at TR) | All other Ethernet ports are data ports |

## Single node clusters

Single node clusters do not need any cluster interconnect connections. While any of the 10Gb network ports could be used as data ports, it is recommended to leave the default cluster interconnect ports free so that they can be used in the future if the system is ever converted into a multi-node cluster. The SnapVault secondary backup systems in the Thomson Reuters environment will start as single node clusters, and may grow to two node switchless or multi-node switched clusters as capacity requirements increase.

## Two node switchless clusters

Two node switchless clusters are used when a single HA pair forms a cluster, and there are no immediate plans to grow the cluster with additional nodes. In this case, the 10Gb cluster interconnect ports are simply directly connected to the same port on the other controller to form the cluster interconnect. If there is a need to transition to a switched cluster, that can be accomplished with no disruption. The Log Backup solution at Thomson Reuters will typically be configured as two node switchless clusters. If smaller scale shared storage environments are used outside of the main datacenters, these systems may also be configured as two node switchless clusters.

## CN1610 switched clusters

Most of the standard shared storage cDOT environments in the main Thomson Reuters datacenters will be configured as switched clusters, using NetApp CN1610 10Gb switches. A typical cDOT shared storage cluster will initially consist of a high tier HA pair and a low tier HA pair. Based on the current cluster limits and the features used in the Thomson Reuters environment, a given cluster may grow up to 8 nodes for standard use. For tech refresh of a node the cluster may grow to 10 nodes during the transition. A variance request through NetApp is required for the 10 configuration.

The CN1610 switches have their configuration set by the NetApp Reference Configuration File (RCF). Each node in the cluster will have one connection to each of the two CN1610 switches used by the cluster. The following table describes the port assignments to be used.

|  |  |  |  |
| --- | --- | --- | --- |
| **CN1610 cluster switch A** | | **CN1610 cluster switch B** | |
| **Switch ports** | **Node/port usage** | **Switch ports** | **Node/port usage** |
| Serviceport/wrench | Management switch (1Gb) | Serviceport/wrench | Management switch (1Gb) |
| 1-12 | Nodes 1-12 cluster port 1 | 1-12 | Nodes 1-12 cluster port 2 |
| 13-16 | ISL to switch B ports 13-16 | 13-16 | ISL to switch A ports 13-16 |

Copper twinax cables up to 5 meters in length or optical LC-LC cables up to 300 meters in length may be used to connect from the cluster nodes to the CN1610 cluster switches. The CN1610 cluster switches should also have a 1Gb connection from the Serviceport (wrench port) to the management network so that any required configuration changes or RCF file updates can be performed over the network. The CN1610 switches are monitored by the nodes in the cluster, and Autosupport is used to notify NetApp support in the event that a CN1610 switch fails.

## Cluster interconnect failure scenarios, quorum, and epsilon

In a cDOT switched cluster environment, connectivity to the cluster interconnect switches is required in order for nodes in the cluster to function properly. The cluster interconnect switches are vital components of the storage system infrastructure, and must be properly monitored and maintained in order to ensure the health of the overall cluster.

Quorum and epsilon are important measures of cluster health and function that together indicate how clusters address potential communications and connectivity challenges. Quorum is a precondition for a fully-functioning cluster. When a cluster is in quorum, a simple majority of nodes are healthy and can communicate with each other. When quorum is lost, the cluster loses the ability to accomplish normal cluster operations. Only one collection of nodes can have quorum at any one time because all of the nodes collectively share a single view of the data. Therefore, if two non-communicating nodes are permitted to modify the data in divergent ways, it is no longer possible to reconcile the data into a single data view.

Because there is the possibility of a tie in a cluster that has an even number of nodes, one node has an extra fractional voting weight called epsilon. When the connectivity between two equal portions of a large cluster fails, the group of nodes containing epsilon maintains quorum, assuming that all of the nodes are healthy.

The following scenarios demonstrate just how critical it is that nodes in a cluster can communicate with each other.

* **Scenario 1** - A node in the cluster loses both cluster interconnect connections:

In this scenario, the node that has lost connectivity to the cluster interconnect is out of quorum with the cluster, and therefore will stop serving data. All other nodes in the cluster will continue serving data, as the remaining nodes are in quorum.

* **Scenario 2** – Up to half of the nodes in a cluster lose both of their cluster interconnect connections, but epsilon resides on a node that still has cluster interconnect connectivity:

In this scenario, all the nodes that have lost connectivity to the cluster interconnect are out of quorum with the cluster, and therefore will stop serving data. All other nodes in the cluster will continue serving data as long as one of the following is true:

* + More than half of the nodes still have cluster interconnect connectivity.
  + Exactly half the nodes still have cluster interconnect connectivity and epsilon also resides on a node that still has cluster interconnect connectivity.
* **Scenario 3** – More than half of the nodes in a cluster lose both of their cluster interconnect connections, or exactly half of the nodes in a cluster lose both of their cluster interconnect connections, and epsilon resides on a node that lost both cluster interconnect connections:

In this scenario, quorum cannot be maintained, and all nodes in the cluster stop serving data.

# Storage System Configuration

## ****Cluster setup****

### Creating a new cluster

You use the Cluster Setup wizard to create the cluster on the first node, via the console mode of the SP connection to the node. The wizard helps you to configure the cluster interconnect that connects the nodes (if the cluster consists of two or more nodes), create the admin and node Vservers, add feature license keys, and create the node management interface for the first node. To create the first node simply answer with “create” when prompted if you would like to create a new cluster during the cluster setup wizard.

### Joining a cluster

You also use the Cluster Setup wizard to add nodes to the cluster, again via the console mode of the SP connection to the new node. To add a node, simply answer with “join” when prompted if you would like to join an existing cluster during the cluster setup wizard. Note that the cluster interconnect must be properly connected in order for a node to join an existing cluster.

### Two node switchless cluster considerations

If you have a two-node switchless configuration in which there is no cluster interconnect switch, you must ensure that the switchless-cluster-network option is enabled. This ensures proper cluster communication between the nodes. This configuration is performed on the first node of the cluster, prior to joining the second node in the two node switchless cluster. As of cDOT 8.2.1, the cluster setup wizard will ask you if the cluster interconnect will be configured to use network switches.

## ****Aggregates****

### Dedicated root aggregates

Every node in the cluster must have a dedicated 3 disk RAID-DP aggregate for the root volume of the node. Under no circumstances should user data volumes be stored on the root aggregate. In order to maintain consistency and avoid accidental usage of the root aggregate, each root aggregate should follow a strict naming standard.

New with ONTAP 8.3 is Advance Drive Partitioning. With ADP, drives are sliced into partitions such that part of the drive is used for the root aggregate and the remainder of the drive is used for the data aggregate. ADP can only be used on AFF and 2000 systems and is configured by default from the factory. Only the first 2 shelves will come as partitioned drives. Note that upgrade from 8.2 to 8.3 will not convert a system to ADP. To convert to ADP in the field, a system has to be upgraded to 8.3 and then zeroed and re-initialized.

### Data aggregates

All non-root aggregates will be used to hold volumes for data Vservers. These data aggregates can be made up of both partitioned and full drive RAID groups. This will be the case for any Thomson Reuters AFF or 2000 configuration on cDOT 8.3. Data aggregates should be configured with *-free-space-realloc* set to *no\_redirect.* This setting will cause the aggregate to automatically manage free space for optimal write performance over the life of the aggregate, with minimal overhead. Note that this setting only applies to system that use HDDs and should not be applied to AFF systems. For most cDOT deployments, TR will have a single data aggregate per node.

## Networking

### Physical network connections

The following table describes the physical network connections that are required for each clustered ONTAP system at Thomson Reuters.

|  |  |  |  |
| --- | --- | --- | --- |
| **Port** | **Quantity** | **Port Speed** | **Notes** |
| SP | 1 per node | 1Gb | Used to connect to a controller’s Service Processor. Connected to TR’s management network. |
| Node Mgmt | 1 per node | 1Gb | 1 Gb data port (e0a/e0i) connected to TR’s management network for node/cluster management. |
| Cluster | 2+ per node | 10Gb | 10Gb cluster ports used to connect each node into the cluster network. Minimum of 2 ports. |
| Data | 2+ per node | 10Gb | Each node will have minimum of two 10Gb ports to be used as a data port. These two ports on each node will typically be configured as a LACP interface group and use VLAN tagging as required. If Intercluster LIFs are required for SnapMirror or SnapVault traffic, these will also reside on the 10Gb data port by convention at TR. |

### Logical interfaces (LIFs)

The following table describes the standard configuration of logical interfaces (LIFs) on a clustered ONTAP system at Thomson Reuters. Each LIF will require an IP address from the TR network to be assigned to it (except for the cluster LIFs, which are private to the cluster and automatically selected by the system).

|  |  |  |  |
| --- | --- | --- | --- |
| **LIF Role** | **Quantity** | **Port Speed** | **Notes** |
| node-mgmt | 1 per node | 1Gb | Each node has a management LIF, used by the node Vserver, on the TR management network. |
| cluster | 2+ per node | 10Gb | Each node must have two cluster LIFs to be used on the cluster interconnect. The IP addresses for these LIFs are private IP addresses that are automatically configured on the system. |
| cluster-mgmt | 1 per cluster | 1Gb | The cluster management LIF requires an IP address on the Thomson Reuters management network. This LIF will float from node to node as required, and is to be used by the admin Vserver for cluster management. |
| intercluster | 1 per node/ipspace | 10Gb | Intercluster LIFs are required for SnapMirror or SnapVault traffic and will require an IP address on the TR data network. |
| data | At least one per Vserver | 10Gb | Each NAS data Vserver will have at least one LIF that requires an IP address on the TR data network for serving data to clients. The data LIF will be created with a –firewall-policy of “mgmt”. This will allow the LIF to also act as a management interface for the Vserver. For iSCSI Vservers, data LIFS are handled a little different. You have one LIF per node of an HA pair and the LIFs do not failover. |
| Vserver-mgmt | One per iSCSI Vserver | 10GB | ISCSI Vserver mgmt interface. Will use 10G connection on the data network. |

### IPspaces (ONTAP 8.3+)

An IPspace is a grouping of SVMs and broadcast domains that provides network isolation. This allows SVMs on the same cluster but in different IPspaces to have overlapping subnets and IP addresses. Since Thomson Reuters controls the networks on the cluster, overlapping subnets is not an issue, but IPspaces also allow a cluster to be dual homed for SnapVault backups which is needed in the Thomson Reuters environment. Upon initial set up or upgrade to 8.3 a cluster will have a Cluster IPspace, which holds the cluster SVM and its cluster network interfaces and a Default IPspace which holds all other SVMs on the cluster.

* IPspace rules
  + If an IPspace is not specified upon SVM and broadcast domain creation they are placed in the Default IPspace.
  + Once an SVM is created in an IPspace you cannot move it to a different IPspace.
  + An IPspace can contain multiple SVMs. An SVM can only belong to one IPspace
* For new installs of 8.3 at Thomson Reuters
  + The first user data network and its SVMs will use the Default IPspace along with the Management SVMs and their network. New IPspaces can be added to the system if additional networks (ECOM, COLO) are required.
* For systems being upgraded from 8.2 to 8.3 at Thomson Reuters
  + By default, all existing SVMs and networks will be upgraded into the Default IPspace. Assuming initial 8.2 clusters only have 1 network attached this will match new 8.3 cluster installs.
  + New IPspaces can be added to the system if additional networks (ECOM, COLO) are required.

### Broadcast Domains (ONTAP 8.3+)

Broadcast domains are a grouping of network ports. The ports in a broadcast domain can either be physical or logical ports (ifgrp, VLAN tagged) and must be on the same VLAN. A LIF should have the same network connectivity across all ports in a broadcast domain.

The use of broadcast domains allows clustered ONTAP 8.3 to configure failover groups automatically. It also makes certain that LIFs will not failover to ports that aren't connected to the same network and that all the ports being used in the broadcast domain have the same MTU value.

* Broadcast Domain rules
  + A failover-group is automatically created when a broadcast domain is created.
  + A port can only belong to one broadcast domain.
  + All ports that are part of a broadcast domain must have the same MTU size.
  + Only the logical ports that are part of the VLAN should be a part of the new broadcast domain. In other words, if VLAN port e0c-2003 is part of the new broadcast domain, the port e0c itself would not be part of the domain and would remain in the Default broadcast domain.
* For new installs of 8.3 at Thomson Reuters
* A management broadcast domain containing the node management ports should be created in the Default IPspace.
* A data broadcast domain in the appropriate data IPspace should be created for the data ports associated with each data network on the cluster (e.g. ECOM & COLO)
* For systems being upgraded from 8.2 to 8.3 at Thomson Reuters
  + For existing broadcast domains, modify the name to match the standard naming convention of Thomson Reuters. The default names for broadcast domains created by an upgrade are the subnet mask for that VLAN.
  + New broadcast domains for new IPspaces would follow the same as for a new 8.3 build.

|  |  |
| --- | --- |
| **Broadcast Domain** | **Notes** |
| Default | All ports not directly associated with a data, management or cluster broadcast domain. |
| Cluster | All ports used by the Cluster network |
| Mgmt | All ports used for Node & Cluster management |
| Data  (e.g. ecom-2010) | All logical ports from the cluster nodes that are part of this data network. More than one Data Broadcast domain will exist on the cluster if multiple IPspaces are setup up. (e.g. ECOM & COLO) |

### Subnets (ONTAP 8.3+)

A subnet in is an optional construct that contains an IP address range, gateway, and netmask. A subnet lives inside of a broadcast domain. The IP address range can also be a list of IP addresses or a list of IP ranges. The subnet does not have to include every IP address in an actual IP subnet, it can be a subset of those IP addresses. Once a subnet object is defined, LIFs can be created using that subnet object. The LIF will automatically be configured using the next available IP address from the subnet object as well as the subnet’s gateway and netmask. Subnets were added to this document for completeness. **At this time Thomson Reuters will not be using Subnets**.

### Routing

With ONTAP 8.3 routing for an SVM has changed. Prior to 8.3 Routing Groups were setup and tied to each LIF of an SVM. An SVM could have 1 or more Routing Groups. With 8.3 routing is now handled by a standard route table that is assigned to the SVM not the LIF. A route table can contain multiple routes but an SVM can only have one route table. The command “network route” is used to manage the route table at ONTAP 8.3

* + Upon creation of an SVM and its LIFs a default route needs to be added to the SVMs routing table
    - “network route create -vserver x -destination 0.0.0.0/0 -gateway x.x.x.x”

### Failover Groups

LIF failover refers to the automatic migration of a LIF in response to a link failure on the LIF's current network port. When such a port failure is detected, the LIF is migrated to a working port. A failover group contains a set of network ports (physical, VLANs, and interface groups) on one or more nodes. A LIF can subscribe to a failover group. The network ports that are present in the failover group define the failover targets for the LIF.

The following table describes the failover groups that must be configured on a clustered ONTAP system at Thomson Reuters.

|  |  |  |  |
| --- | --- | --- | --- |
| **Failover Group Name** | **Ports** | **Port Speeds** | **Notes** |
| mgmt | The e0a or e0i port on all nodes | 1Gb | By convention at TR, we will configure the cluster management LIF on the admin Vserver to be part of a failover group named *mgmt*, which will consist of all the e0a or e0i ports in the cluster. |
| data-<vlan> | The a0a-<vlan> ports on all nodes | 10Gb | By convention at TR, we will configure each data Vserver LIF to be part of a failover group named *data-<vlan>*, where *<vlan>* reflects the VLAN id used on a given interface group. As new data Vservers and new LIFs are created, they will need to be properly assigned to this failover group. |

In addition to the custom failover groups that we will configure at Thomson Reuters, there is a default failover group named “system-defined”, which is used on LIFs that are not designed to failover between nodes (such as node management LIFs, cluster LIFs, and Intercluster LIFs).

For ONTAP 8.3+, Failover Groups are automatically configured when Broadcast Domains are created. No further Failover Group configuration is required. For 8.2 systems upgraded to 8.3, Failover Groups and their use should be checked/updated to match 8.3 naming standard.

### External Management Services Access (ONTAP 8.3+)

With ONTAP 8.3 all access to DNS, Active Directory, or other external management services must be done through a mgmt LIF on the SVM. Prior to 8.3 an SVM could route a DNS request through the node management LIF. This is no longer allowed at 8.3+. Each SVM must have a management capable LIF with access to all needed external management Services. For Thomson Reuter’s, iSCSI SVMs will use their SVM management LIF and non-iSCSI SVMs will use the SVM data LIF which has a firewall-policy of mgmt. NOTE: DNS services must be setup on the SVM prior to upgrading to 8.3 or risk losing name resolution as the node management SVM will not be able to provide DNS resolution after the upgrade.

### Upgrade from 8.2 to 8.3 LIF Requirement

With the network changes between ONTAP 8.2 and 8.3 an additional LIF requirement is required by the system during the upgrade. This requirement is also called out in the ONTAP Upgrade Advisor. Note that during the ONTAP image update step a LIF sufficiency check is run to make sure your system meets the LIF requirements for the upgrade. Do not ignore this check. It will pass when all needed requirements are met.

**LIF Requirements for Upgrading**

Prior to upgrading a cluster to 8.3 the following requirements must be met….

* + The first node of cluster to be upgraded must host at least 1 LIF that can access external management services (DNS, AD, LDAP, etc) from each SVM.
  + The LIF must be setup with auto-revert to ensure that it fails back once that node reboots after being upgraded to 8.3.
  + For all standard Thomson Reuters NFS/CIFS SVMs with one data/mgmt LIF the following steps would need to be done.
    - Change the home-port of the SVM data LIF to the first node of cluster to be upgraded.
    - Set auto-revert to true for that LIF.
    - Verify that the LIF’s current location is the first node to be upgraded.
    - Once all nodes have been upgraded reset all LIFs back to their original home-port and auto-revert setting.
  + For a standard Thomson Reuters iSCSI SVM with a separate management LIF, the following steps would need to be done.
    - Change the home-port of the SVM mgmt LIF to the first node of cluster to be upgraded.
    - Set auto-revert to true for that LIF.
    - Verify that the LIF’s current location is the first node to be upgraded.
    - Once all nodes have been upgraded, reset all management LIFs back to their original home-port and auto-revert setting.

### Upgrade from 8.2 to 8.3 Failover Groups Requirement

During the upgrade to 8.3, Broadcast Domains are automatically created based off of the failover groups on the 8.2 system. If the failover groups are invalid the upgrade will fail or the Broadcast Domains will be setup incorrectly. An example of an invalid failover group is one that contains ports that are used for different subnets. The failover group can be a legacy group that is not currently used and still cause problems with the upgrade.

To check for this issue prior to the upgrade, NetApp is working on a script that will analyze the 8.2 Failover group setup and provide corrections as needed. Please check with the local NetApp team for this script prior to any upgrades.

## Vservers

### Node Vservers

Each node has a Vserver that can be used to manage the entire cluster by connecting to the management LIF owned by that node. The node Vserver exists to provide a management interface in the event that the admin Vserver is not accessible, and perform operations such as AutoSupport. Normal day to day management should be performed using the admin Vserver.

### Admin Vserver

Each cluster has the concept of one admin Vserver, which is used to manage the cluster. The admin Vserver LIF will be configured to automatically failover between nodes in the cluster, in the event that the node currently holding the LIF is unavailable.

### Data Vservers

Each volume that stores end user data is associated with one data Vserver. A given volume cannot be attached to more than one Vserver, but Vservers can contain multiple volumes. Multiple data Vservers will typically be configured in a cluster. Each data Vserver must have at least one LIF, and by convention at TR we will typically have only one LIF per Vserver when serving NFS and CIFS traffic on the Vserver.

## Miscellaneous configuration

### Vserver services

Various services are configured for the cluster Vserver, including DNS, AutoSupport, NTP, SNMP, and the local time zone. Examples of these commands are provided in the *Miscellaneous admin Vserver configuration* section of the CLI examples. Note that the DNS service should also be configured on all data Vservers.

### Unlocking the diag user account

The diag account is used by the perfstat8 utility, which may be required when interacting with the NetApp support center to handle performance cases. The commands to unlock this account and configure a password for it are given in the *Miscellaneous admin Vserver configuration* section of the CLI examples.

### The wlstats account

The wlstats account is configured on each cluster to facilitate gathering granular performance statistics in the TR environment. The commands to setup this account are given in the *Miscellaneous admin Vserver configuration* section of the CLI examples.

### CN1610 device monitoring

In a switched cluster environment, cDOT 8.2.1 and above have the ability to monitor CN1610 cluster switches for faults, and automatically generate an AutoSupport in the event of a switch failure. The commands for checking on the status of the CN1610 switch monitoring are given in the *Miscellaneous admin Vserver configuration* section of the CLI examples.

### System core file and log access via HTTP

As of cDOT 8.2.1, access to core files and log files on a cluster are provided via an http interface. If support asks for this information, it can be accessed using the admin account at a URL of http://<clustername>/spi. There is no configuration required for this feature.

### Vserver and Volume language setting

Upon creation of a Vserver or a volume the language will need to be explicitly set to en\_US. Languages varying from this standard are allowed but will need to be reviewed by TR Storage Engineering. If the language is not set, Vserver language will default to C.UTF-8 and volume language will default to the language of the Vserver that owns it. Note, you cannot change the language on a volume without destroying and recreating the volume.

## IPspaces and Cluster Peering for SnapVault

### Intercluster LIF and peering relationship

Backup controllers will follow the same IPspace rules as primary storage clusters. For each network (ECOM, COLO) an IPspace is setup to hold data SVMs to handle the SnapVault destination data.

* + On both the primary and secondary system configure one intercluster LIF per IPspace per node.
    - For networks that are part of the Default IPspace the Intercluster LIF will be owned by the node SVM. For new non-Default IPspaces the Intercluster LIF will be owned by the system SVM that is created by default when you create an IPspace.
    - The LIF will use the data network port configured for that IPspace.
    - For a 4-node cluster with 2 IPspaces you would have 8 Intercluster LIFs.
    - Make sure the intercluster LIFs of an IPspace can ping each other.
  + Create a cluster peer relationship between the primary system and secondary system.
    - This is done in by running the “cluster peer” command on the primary system pointing at the intercluster LIFs on the secondary and then running the command a second time from the secondary pointing at the Intercluster LIFs of the primary.

## Compression and Dedupe Settings

### Inline Zero Detection, Inline Adaptive Compression, & Inline Dedupe

ONTAP 8.3.2 comes with storage efficiencies that are new to the Thomson Reuters environment. They include Inline Zero Dedupe, Inline Adaptive Compression, & Inline Dedupe.

* **Inline Zero Detection** – Inline detection and deduplication of zeros. Enabled by default if Inline Compression is enabled. Low CPU overhead
* **Inline Adaptive Compression** – 8K compression groups that easily decompress for excellent read performance. Note that this type of compression is different than what is being used on the TR Log Archive systems. They are using Secondary Compression that does provide better compression savings but has a performance hit on reads. Inline Adaptive Compression is designed for production use with very low performance impact and is on by default for all volumes on an AFF system.
* **Inline Dedupe** – Deduplication of data on ingest into the controller. Designed to be of very low overhead. If the performance impact is greater than 2% the dedupe process is throttled and the remaining deduplication is completed as a post process.

### Recommended Compression and Dedupe Settings for an AFF system

* + Enable Inline Compression and Dedupe for all volumes on an AFF system.
    - volume efficiency on -vserver <vserver> -volume <volume>
    - set advanced
    - volume efficiency modify -vserver <vserver> -volume <volume> -inline-compression true -inline-dedupe true -compression true -compression-type adaptive
  + If the data was pre-existing in the volume, run a onetime manual scan during off hours to compress and dedupe the old data.
    - volume efficiency start -vserver <vserver> -volume <volume> -dedupe true -compression true -qos-policy background -scan-old-data true -scan-all true -snapshot-blocks true
  + Schedule a post processing job to run hourly to dedupe any data that wasn’t handled by the inline dedupe process due to the performance threshold. This job runs in the background and is throttled as to not affect any normal IO processing.
    - volume efficiency policy create -vserver <vserver> -policy AlwaysOn -type scheduled -schedule hourly -qos-policy background
    - volume efficiency modify -vserver <vserver> -volume <volume> -policy AlwaysOn

### Recommended Compression and Dedupe Settings for SnapVault secondary volumes of AFF primary volumes

* + For all AFF primary volumes that are backed up to a SnapVault secondary, the secondary volumes should NOT have any storage efficiencies enabled. Since the primary volume contains all the storage efficiencies they will be maintained across the wire and on the secondary volume as long the secondary volume does not have storage efficiencies enabled.
  + volume efficiency off -vserver <vserver> -volume <volume>

### Recommended Compression and Dedupe Settings for non-AFF Primary Systems

* + For now, all non-AFF systems will follow the established standard of only running post dedupe processing on vmware/virtualized data volumes. Recommend testing be done on Flash Pool systems to determine the feasibility of running inline compression on these systems in the Thomson Reuters environment. Inline compression will take advantage of the Flash Pool storage and is recommended for use on a Flash Pool system. Testing is recommended prior to use.

# Single node cluster system backup configuration

In a multi-node cluster, the cDOT system configuration databases are automatically backed up to other nodes in the cluster. In a single node cluster, there are no other nodes on which to place configuration backups, and therefore an off-system backup location should be configured for single node SnapVault secondary systems in the TR environment. The *System configuration backup settings* command examples in the CLI example section show how to setup a cDOT system to automatically send backups to a remote HTTP server. By default, the cDOT system will maintain two backups for each time period, which are taken every 8 hours, every night, and every week on the remote HTTP server.

## HTTP server configuration on a management server

In the TR environment, a number of administrative tasks and scripting are performed on a Linux based management server. As of the writing of this document, this management server is typically a 7 mode NetApp OnCommand Unified Manager OCUM (DFM) server, which is also running an instance of the Apache HTTP server. This instance of Apache must be configured with the WebDav module in order to accept HTTP PUT operations for cDOT configuration backups. The *cdot\_config\_backups* directory should be placed on a NAS mount on the OCUM server.

### Configuration of a SLES Apache server

1. Add the *dav* and *dav\_fs* modules to the *APACHE\_MODULES* list in */etc/sysconfig/apache2*.
2. Put the following configuration lines in the /*etc/apache2/default-server.conf* file:

<IfModule mod\_dav\_fs.c>

DAVLockDB /var/lib/dav/lockdb

</IfModule>

<Location /netapp/cdot\_config\_backups>

Order Allow,Deny

Allow from all

Dav On

</Location>

1. Create the */srv/www/htdocs/netapp/cdot\_config\_backups* directory and change the ownership of this directory to the same account that Apache runs as.
2. Create *the /var/lib/dv/lockdb* directory, and change the ownership of this directory to the same account that Apache runs as.

### Configuration of an OEL Apache server

1. Put the following configuration lines in the /*etc/httpd/conf/httpd.conf* file:

<IfModule mod\_dav\_fs.c>

DAVLockDB /var/lib/dav/lockdb

</IfModule>

<Location /netapp/cdot\_config\_backups>

Order Allow,Deny

Allow from all

Dav On

</Location>

1. Create the */var/www/html/netapp/cdot\_config\_backups* directory and change the ownership of this directory to the same account that Apache runs as.

# Clustered ONTAP configuration CLI examples

## ****CN1610 cluster switch configuration****

### Set admin password and enable mode password

password

enable password <new password>

### Set management IP

serviceport protocol none

network protocol none

serviceport ip <ip addr> <netmask> <gateway>

### Configure SSH and disable telnet

ip ssh protocol 2

config

crypto key generate rsa

crypto key generate dsa

exit

ip ssh server enable

no ip telnet server enable

### Set clock

config

clock set <mm/dd/yyyy>

clock set <hh:mm:ss>

exit

show clock

### Set hostname

hostname <new switch name>

### Save configuration

write memory

### Check FastPath and RCF version

show version

show bootvar

show running-config

### Update RCF version

en

show running-config running-config-<MM-DD-YY>.scr

copy tftp://<tftp\_server\_ip>/CN1610\_CS\_RCF\_v<version>.scr nvram:script CN1610\_CS\_RCF\_v<version>.scr

script list

script apply CN1610\_CS\_RCF\_v<version>.scr

write memory

show running-config

### Diagnostic commands

enable

show port all

show interface <slot/port>

## ****Initial cluster setup****

### Creating a new cluster on the first node

Do you want to create a new cluster or join an existing cluster? {create, join}: create

cluster show

### Joining a cluster with additional nodes

Do you want to create a new cluster or join an existing cluster? {create, join}: join

cluster show

### Configuration settings for two node switchless cluster

set -privilege advanced

network options switchless-cluster modify -enabled true

cluster ha modify -configured true

network options switchless-cluster show

cluster ha show

### Licensing

system license add

system license show

## ****Aggregate creation****

### Root aggregate renaming

storage aggregate rename <root\_aggr> <new\_root\_aggregate\_name>

storage aggregate show

### Create a standard aggregate with free space reallocation

storage aggregate create -aggregate <aggregate\_name> -nodes <node> -diskcount <diskcount> -raidtype raid\_dp -maxraidsize <raidsize>

set diag

storage aggregate modify -aggregate <aggregate\_name> -free-space-realloc no\_redirect

storage disk show

storage aggregate show

storage aggregate show –instance

### Create a FlashPool aggregate with free space reallocation

storage aggregate create -aggregate <aggregate\_name> -nodes <node> -diskcount <diskcount> -raidtype raid\_dp -maxraidsize <raidsize>

storage aggregate modify -aggregate <aggregate\_name> -free-space-realloc no\_redirect

**NOTE:** Free Space Reallocate is only needed for spinning hard drives. For the AFF8080 or any other AFF system TR may use Free Space Reallocate should be set to off.

storage aggregate modify -aggregate <aggregate\_name> -hybrid\_enabled true

storage aggregate add-disks -aggregate <aggregate\_name> -disktype SSD -raidtype raid\_dp -diskcount <count>

## ****Network configuration****

### Disabling Ethernet flowcontrol on cluster interconnect ports

network port modify -node <node> -port <port> -flowcontrol-admin none

network port show -role cluster -fields flowcontrol-admin

**Note:** For ONTAP 8.3+ a port’s role is defined by its Broadcast Domain. The above command will work but will show the flowcontrol-admin setting for any port in the Broadcast Domain “cluster”.

### Disabling Ethernet flowcontrol on data ports

network port modify -node <node> -port <port> -flowcontrol-admin none

network port show -instance

### Creating a LACP IFGRP with data ports

network port ifgrp create -node <node> -ifgrp a0a -mode multimode\_lacp -distr\_func ip

network port ifgrp add-port -node <node> -ifgrp a0a -port <port1>

network port ifgrp add-port -node <node> -ifgrp a0a -port <port2>

network port ifgrp show

### Adding a VLAN tag to an IFGRP of data ports

network port vlan create -node <node> -port a0a -vlan-id <vlan>

network port vlan show

### Enabling jumbo frames on an IFGRP of data ports or a VLAN tag

network port modify -node <node> -port a0a -mtu 9000

network port modify -node <node> -port a0a-<vlanid> -mtu 9000

network port show

### Creating an IPspace (ONTAP 8.3+)

ipspace create -ipspace corp-ipspace-01

### Creating a Broadcast Domain (ONTAP 8.3+)

broadcast-domain create -broadcast-domain corp-504 -ipspace corp-ipspace-01 -port firkin-01:a0a-504 -mtu 9000

broadcast-domain add-ports -broadcast-domain corp-504 -port firkin-02:a0a-504

### Creating a failover group for an IFGRP

network interface failover-groups create -failover-group data-<vlan> -node <node1> -port a0a-<vlan>

network interface failover-groups create -failover-group data-<vlan> -node <node2> -port a0a-<vlan>

network interface failover-groups show

network interface show –failover

### Applying the failover group for management ports

network interface modify -vserver <cluster\_vserver> -lif <cluster\_management\_LIF\_name> -failover-group mgmt

network interface show –failover

### Create Intercluster LIF with default route

network interface create -vserver <node\_vserver> -lif <ICL\_LIF\_name> -role intercluster -home-node <node> -home-port a0a-<vlan> -address <ip> -netmask <netmask>

network routing-groups route create -vserver <node\_vserver> -routing-group d<network>/<mask> -destination 0.0.0.0/0 -gateway <gateway>

(**Note**: Creating the route for ONTAP 8.3+ system)

network route create -vserver <node\_vserver or IPspace system Vserver> -destination 0.0.0.0/0 -gateway <gateway IP>

network interface show

### Cluster Peering

cluster peer create -peer-addrs ICL-IP-Node1,ICL-IP-Node2 -ipspace <ipspace name>”

## ****Miscellaneous admin Vserver configuration****

### DNS

vserver services dns create -vserver <cluster\_vserver> -domains <domainname> -name-servers <comma\_separated\_IP\_addresses>

vserver services dns show

### Autosupport

system node autosupport modify -node <node> -transport https -proxy-url <proxy:port>

system autosupport show –node <node>

### Timezone configuration

timezone -timezone <timezone>

date

### NTP configuration

system services ntp server create -node <node> -server <ntpserver1>

system services ntp server create -node <node> -server <ntpserver2>

system services ntp server show

(**Note**: Starting with ONTAP 8.3, NTP time services switch to a cluster wide setting.)

cluster time-service ntp server create -server <ntpserver1>

cluster time-service ntp server create -server <ntpserver2>

cluster time-service ntp server show

### SNMP configuration

snmp community add ro public

### Unlocking the diag user

security login unlock -username diag

security login password -username diag

### Restricted account for wlstats

security login role create -role wlstats -cmddirname statistics -access all -vserver <cluster\_vserver>

security login role create -role wlstats -cmddirname version -access all -vserver <cluster\_vserver>

security login create -username wlstats -application ontapi -authmethod password -role wlstats -vserver <cluster\_vserver>

security login show

security login role show

security login role show-ontapi

### SSH publickey authentication for cluster admin

security login create -username admin –vserver <admin\_vserver> -application ssh -authmethod publickey –role admin

security login publickey create -username admin -vserver <admin\_vserver> -publickey “<ssh\_publickey>”

security login show

security login publickey show

### CN1610 cluster switch device monitoring commands

set advanced

cluster ping-cluster <node>

system health cluster-switch delete -device <switch\_name>

system health cluster-switch create -device <switch\_name> -address <ip> -snmp-version SNMPv2c -community cshm1! -model CN1610 -type cluster-network -disable-monitoring no

system health cluster-switch show

**Note:** The switch community name changed between RCF 1.0 and RCF 1.1. If you are using RCF 1.0 use the community name netapp. RCF 1.1 should be used for cDOT 8.2 onwards.

**Note #2:** For ONTAP 8.3+ the cluster switch monitoring commands were moved up one level to “system cluster-switch....”

### System configuration backup settings

set advanced

system configuration backup settings modify -username admin -destination http://<dfm\_server\_ip>/netapp/cdot\_config\_backups

system configuration backup settings set-password