Dell ECS: Networking Best Practices

January 2023

H15718.10

White Paper

Abstract

This white paper describes networking and related best practices for ECS, the Dell software-defined cloud-scale object storage platform.

Dell Technologies

Copyright

Published in the USA January 2023 H15718.10.

Contents

Executive summary

Overview

Dell ECS is a cloud-scale, object-storage platform for traditional, archival, and next-generation workloads. It provides geo-distributed and multiprotocol (object, HDFS, and NFS) access to data. An ECS deployment offers a turnkey appliance with industry-standard hardware that you can use to form the hardware infrastructure. In either type of deployment, a network infrastructure is required for the interconnection between the nodes and customer environments for object storage access.

This paper describes ECS networking and provides configuration best practices. It provides details about ECS network hardware, network configurations, and network separation. It also describes how ECS connects to customer environments. Use this paper as an adjunct to the following Dell ECS documentation on the ObjectScale and ECS Info Hub:

- ECS Hardware Guide (for Gen1 and Gen2 hardware)
- ECS EX-Series Hardware Guide
- Network Guide for D- and U- Series (Gen 1 and Gen 2 hardware)
- Network Guide for EX300 and EX3000 (EX-Series hardware)

Updates to this document are completed periodically and often coincide with new features and functionality changes.

Audience

This document is for Dell field personnel and for customers who are interested in understanding ECS networking infrastructure and the role that networking plays within ECS.

Revisions

Date	Description
December 2016	Initial release
August 2017	Updated based on ECS 3.1
February 2019	Added Gen3 details
March 2020	Updated based on ECS Special Feature Configuration Support
April 2021	Updated based on ECS 3.6.1
March 2022	Updated template
March 2022	Updated based on ECS 3.7; template update
July 2022	RPQ is no longer required for customer-provided FE switches that meet the Pre-approved Configuration Requests for ECS Appliance
January 2023	Updated based on ECS 3.8.0.1

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Author: Jarvis Zhu

Note: For links to other documentation for this topic, see the ObjectScale and ECS Info Hub.

ECS overview

ECS features a software-defined architecture that promotes scalability, reliability, and availability. ECS is built as a completely distributed storage system to provide data access, protection, and geo-replication. The main use cases for ECS include storage for modern applications and secondary storage to free primary storage of infrequently used data while also keeping it reasonably accessible.

ECS software and hardware components work in concert for unparalleled object and file access. Figure 1 shows the software storage layers along with the underlying infrastructure and hardware layers. The set of layered components includes:

- **ECS portal and provisioning services**: Provides an API, CLI, and web-based portal that allows self-service, automation, reporting, and management of ECS nodes. It also handles licensing, authentication, multitenancy, and provisioning services.
- **Data services:** Provides services, tools, and APIs to support object, and HDFS and NFSv3.
- **Storage engine:** Responsible for storing and retrieving data, managing transactions, and protecting and replicating data.
- **Fabric:** Provides clustering, health, software, and configuration management as well as upgrade capabilities and alerting.
- **Infrastructure:** Uses SUSE Linux Enterprise Server 12 as the base operating system for the turnkey appliance or qualified Linux operating systems for industry-standard hardware configuration.
- **Hardware:** Includes industry-standard hardware composed of x86 nodes with internal disks or attached to disk-array enclosures with disks, and top-of-rack (ToR) switches.

For an in-depth architecture review of ECS, see *ECS: Overview and Architecture*.

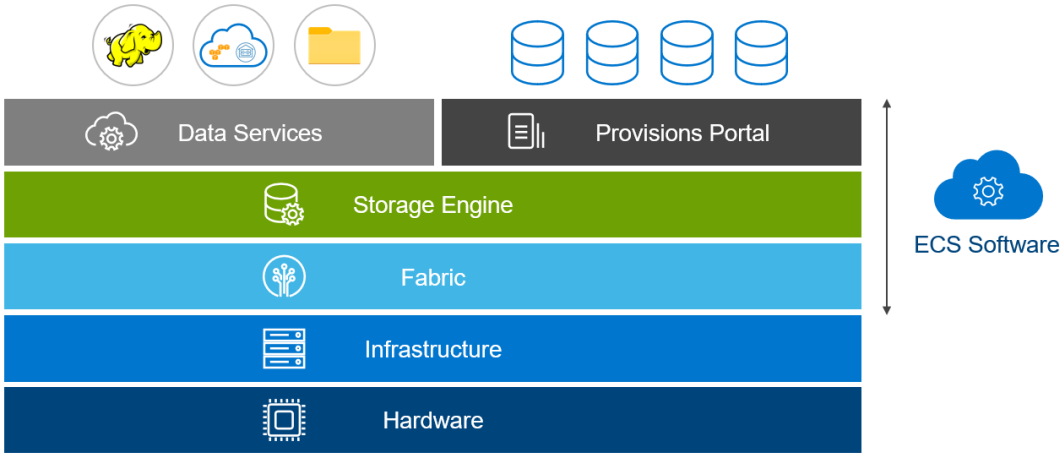

Figure 1. ECS software layers

ECS network overview

Introduction

ECS network infrastructure consists of a set of ToR switches that allow for the following two types of network connections:

- **Public Network:** Connection between the customer production network and ECS
- **Private Network:** For management of nodes and switches within and across racks

The ToR switches are dedicated to either the public (production) network or to the private, internal-to-ECS-only network. For the public network, a pair of 10/25 GbE network switches are used to service data and internal communication between the nodes. For the private network, depending on the hardware generation, use either a single 1 GbE switch for Gen1 or Gen2 (Gen1/2). Alternately, use a pair of 25 GbE switches (Gen3). The private network is used for remote management, console access, and PXE booting, which enables rack management and cluster-side management and provisioning. From this set of switches, uplink connections are presented to the customer production environment for storage access and management of ECS. The networking configurations for ECS should be redundant and highly available.

Note: Gen1/2 and Gen3 EX300, 500, 3000, and 5000 Series systems use a public network for internal communication between nodes. Gen3 EXF900 Series systems use a private network for internal communication between nodes.

Traffic types

Understanding the traffic types within ECS and the customer environment is useful for architecting the network physical and logical layout and configuration for ECS.

The public network carries the following types of traffic:

- **Data:** Customer data and I/O requests
- **Management:** Provisioning or querying of ECS through the portal or ECS Rest Management APIs, and network-services traffic such DNS, AD, and NTP
- **Internode:** Messages sent between nodes to process I/O requests depending on owner of data and internode checks
- **Replication:** Data replicated to other nodes within a replication group

In a single-site, single-rack deployment, internode traffic stays within the ECS rack switches. Alternately, in a single-site multirack deployment, internode traffic traverses from one rack set of switches up to the customer switch and to the other rack switches to process requests. In a multisite or geo-replicated deployment, the traffic also goes across the WAN.

Note: The internode traffic of EXF900 runs in a private network due to the NVMe-oF architecture of EXF900.

The private network, which is under Dell control, is entirely for node and switch maintenance. Traffic types include:

- **Segment Maintenance Management:** Traffic associated with administration, installation, or setup of nodes and switches within the rack.
- **Private.4 network:** Interconnects multiple, co-located ECS intra-rack networks into a single inter-rack network through VLAN 4. The private.4 network is also referred to as the Nile Area Network (NAN).

ECS network hardware

Introduction

Each ECS appliance rack contains three, four, or six switches. Gen1/2 appliances have three switches, two for the public network and one for the private network. Gen3 EX300, EX500, EX3000, and EX5000 systems have two public switches and two private switches. Gen3 EXF900 systems have another two dedicated aggregation switches for private switches, ensuring that all the EXF900 nodes have line rate performance to any node in any rack.

For switch details, including model numbers, along with designated switch port usage and network cabling information, see the *ECS Hardware Guide* for Gen1/2 appliances and the *ECS EX-Series Hardware Guide* for Gen3 appliances.

Public switches

Public (production or front-end) switches are used for data transfer to and from customer applications as well as internal node-to-node communication for Gen1/2 and Gen3 EX300, 500, 3000, and 5000 Series. The inter-node traffic of Gen3 EXF900 will go through the private switch. These switches connect to the ECS nodes in the same rack. For Gen1/2 appliances, two 10 GbE, 24-port or 52-port Arista switches are used. For Gen3 appliances, two 10/25 GbE (EX300) or two 25 GbE (EX500, EX3000, EX5000, and EXF900) 48-port Dell switches are used. To create a high availability (HA) network for the nodes in the rack, the public switches work in tandem using LACP/MLAG, with the Arista switches in Gen1/2 appliances, and Virtual Link Trunking (VLT), with the Dell switches in Gen3 appliances. This pairing is for redundancy and resiliency in case of a switch failure.

Across all generations of hardware, Gen1-3, each ECS node has two Ethernet ports that directly connect to one of the ToR public switches. Due to NIC bonding, the individual connections of a node appear to the outside world as one. The nodes are assigned IP addresses from the customer's network either statically or through a DHCP server. At a minimum, one uplink between each ToR public switch in the ECS appliance to the customer network is required. The public switch management ports connect to the ToR private switch or switches.

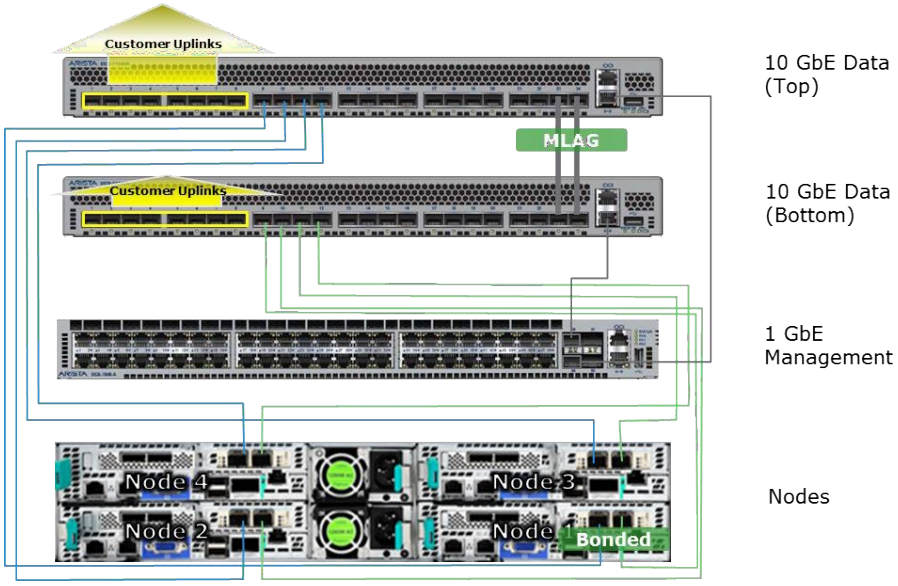

Figure 2. Public network for ECS Gen2 nodes

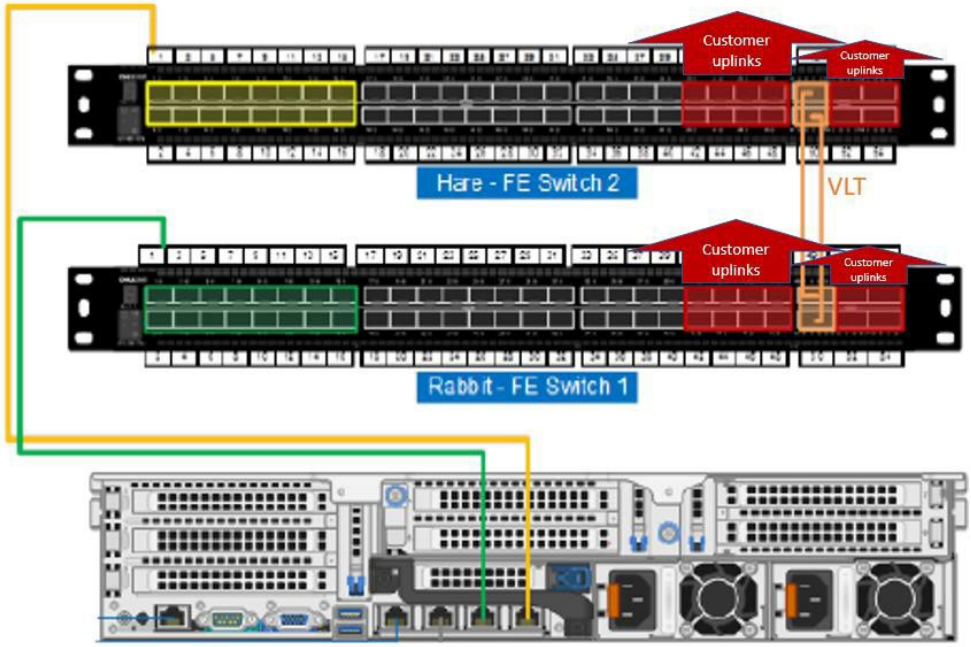

Figure 3. Public network for ECS Gen3 nodes

Best practices:

- For redundancy and to maintain a certain level of performance, have two uplinks per switch to customer switch, or four uplinks per rack minimum.
- Use 25 GbE switches for optimal performance when using customer-provided public switches.
- Have dedicated switches for ECS and do not use shared ports on the customer core network.

Private switches

Private switches are used by ECS for node management. For Gen1/2 appliances, the private switches also allow for out-of-band (OOB) management communication between customer networks and Remote Management Module (RMM) ports in individual ECS nodes. Gen1/2 appliances have a 52-port, 1 GbE Arista switch, or a Cisco switch for organizations with strict Cisco only requirements. Gen3 appliances contain two 25 GbE, 48-port Dell private switches identical in model to the public switches.

Note: Gen3 does *not* allow for OOB management communication from customer networks.

The management ports in each node connect to one or more private switches. They use private addresses such as 192.168.219.x. Each Gen1/2 node also has a connection between its RMM port and the private switch. This node also can have access to the customer network to provide OOB management of the nodes. Gen3 nodes also have a connection between their integrated Dell Remote Controller (iDRAC) and one of the private switches. However, there is no customer-facing OOB management for Gen3 ECS nodes.

Note: Dell switches are required for the private network. Private switches cannot be customer-provided.

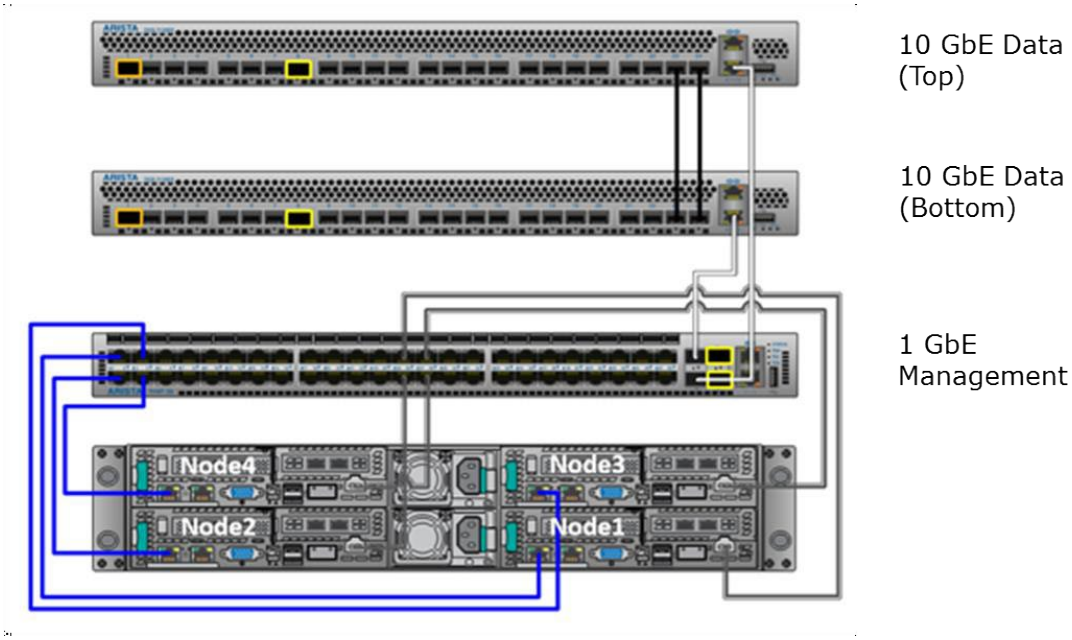

Figure 4. Private network for Gen2 ECS nodes

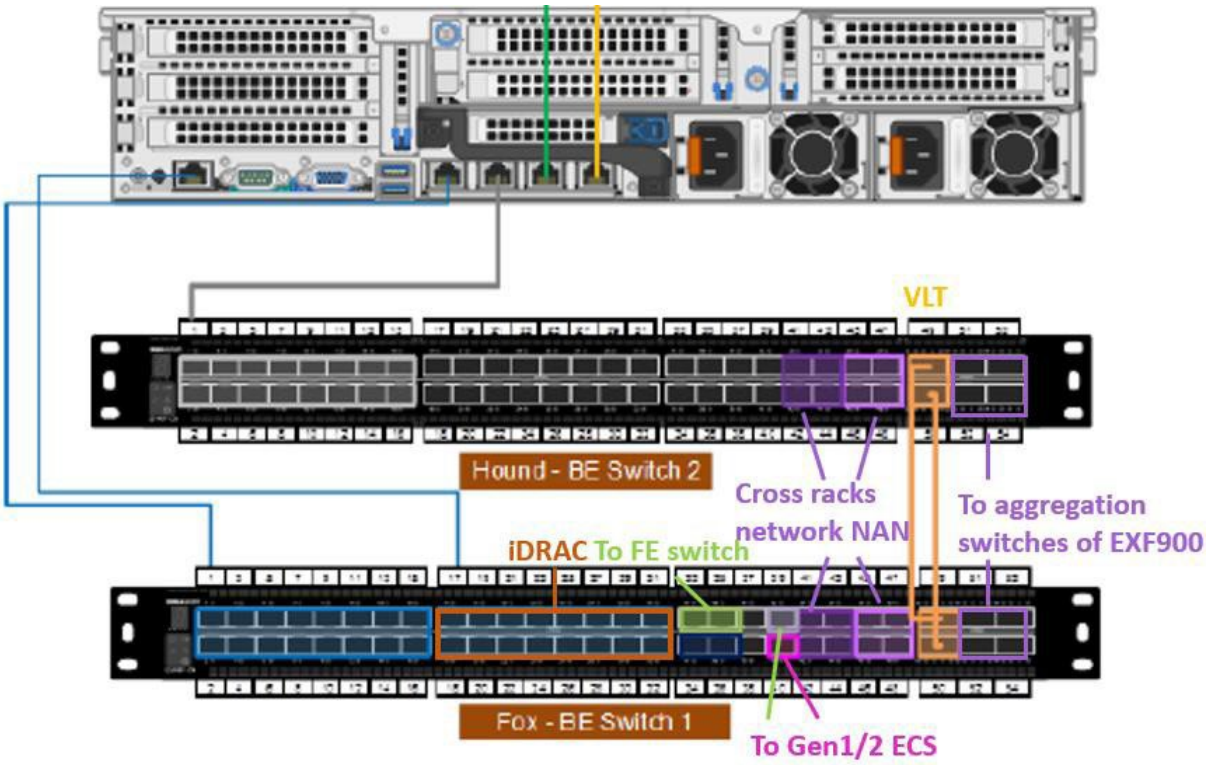

Figure 5. Private network for Gen3 ECS nodes

Best practices:

- When physically connecting nodes to the management switch, do so in an ordered and sequential fashion. For instance, node 1 should connect to port 1, node 2 to port 2, and so on. Connecting nodes to an arbitrary port between 1 through 24 can cause installation issues.
- RMM/iDRAC connections are optional. The best practice is to ask for customer requirements for these connections.

Aggregation switches for EXF900

The aggregation switches can be installed in the Dell rack or a customer provided rack. The aggregation switch allows you to connect up to seven racks of EXF900 nodes in the same cluster.

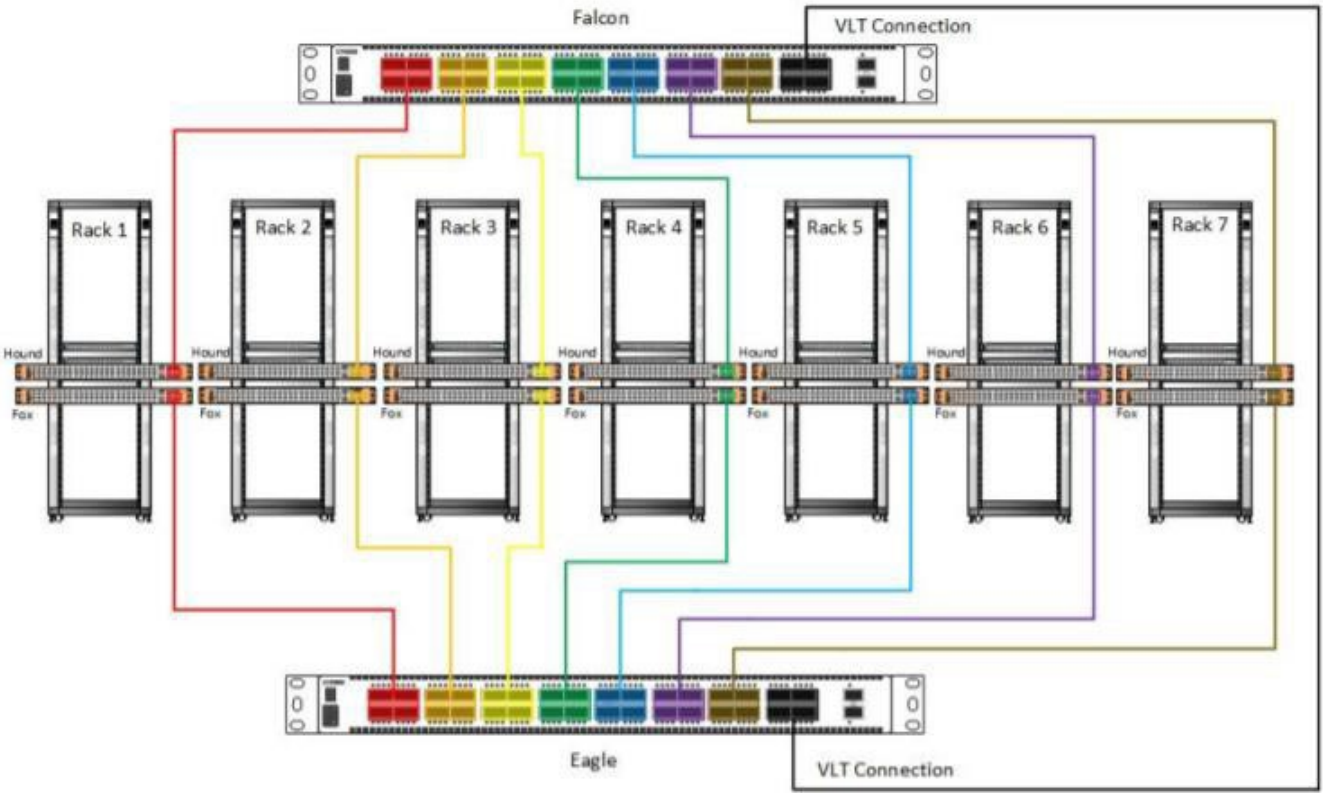

Figure 6. Aggregation network topology

Note: For more information about the network cabling, see the *ECS EX Series Hardware Guide* and the *ECS Networking Guide*.

Customer-provided switches

The flexibility of ECS allows for variations of network hardware and configurations that meet the Dell standards.

Note: ECS Appliance-Special Feature Configuration Support is an internal-only authenticated reference. Customers must ask for presales or sales assistance to read this reference.

Customers are responsible for configuration and support for customer-provided switches. These switches should be dedicated to ECS and not shared with other applications. Dell assistance is advisory for customer-provided switches.

The private network switches for an ECS appliance cannot be replaced. They are solely for administration, installation, diagnosing, and management of ECS nodes and switches. The private network switches must remain under control of Dell personnel.

ECS network configurations

Introduction

The previous section briefly describes the switches and related networks used by ECS appliances. This section explores further the public production network and the private ECS internal management network referred to as the *Nile Area Network,* or NAN. Design

considerations and best practices in both production and internal networks are discussed to offer guidance for network architects.

Production network

The production network involves the connections between the customer's network and the two ToR ECS front-end, public data switches and the connections within the ECS rack. These connections act as the critical paths for in and out client requests and data ("north to south"), and internode traffic ("east to west") for replication and processing requests, as shown in the following figure for a single rack.

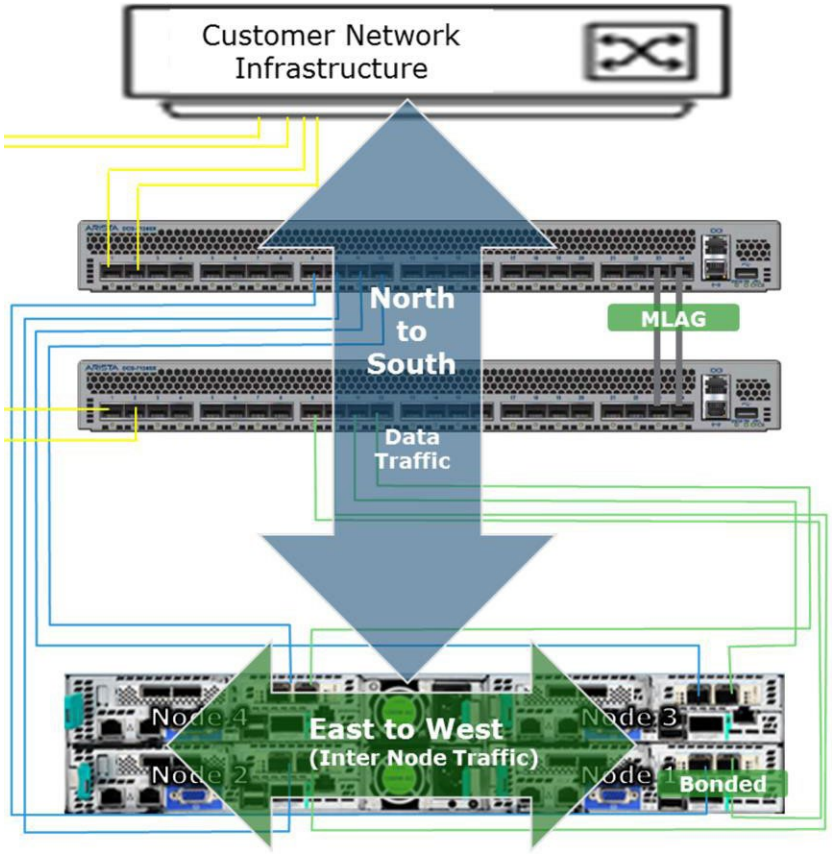

Figure 7. Network traffic flow in a single rack

For multirack Gen1/2 and Gen3 EX300, EX500, EX3000 and EX5000 systems, internode traffic flows north to south and over to the customer network and to the other ECS racks, as shown in the following figure.

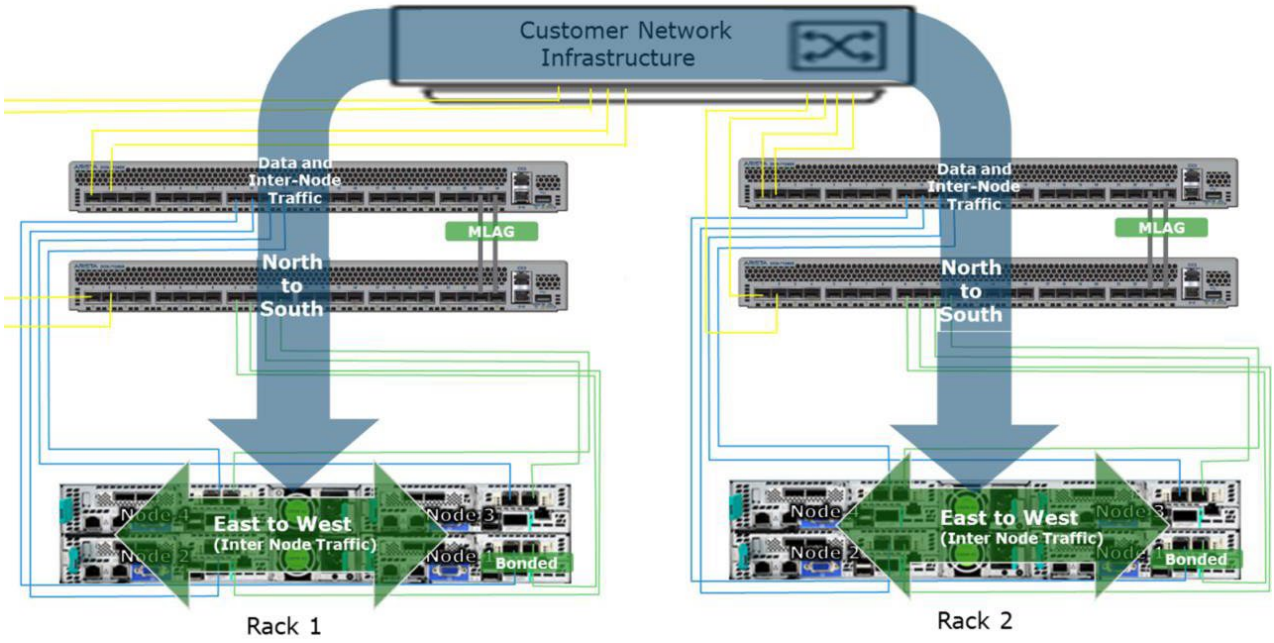

Figure 8. Network traffic flow in a multirack

Note: For Gen3 EXF900, designed with NVMe-oF architecture, the internode traffic goes through the private switch and dedicated aggregation switch to improve the performance.

As a best practice, network connections in the production network should be designed for high availability, resiliency, and optimal network performance.

ToR public switch configuration and node connectivity

The public ToR switches work in tandem using LACP/MLAG with the Arista switches in Gen1/2 appliances and Virtual Link Trunking (VLT) with the Dell switches in Gen3 appliances. This configuration creates an HA network for the nodes in the rack. Similarly, if Cisco switches are used, the vPC LAG protocol is used for HA. The aggregation of multiple ports results in higher bandwidth, resiliency, and redundancy in the data path.

Each node in the rack is connected to both switches through two NICs, which are aggregated using a Linux bonding driver. The node is configured to bond the two NICs into a single LACP bonding interface also known as a "mode 4" bond. This bonding interface connects one port to each switch.

The following two figures are examples of nodes bonded with a public switch pair.

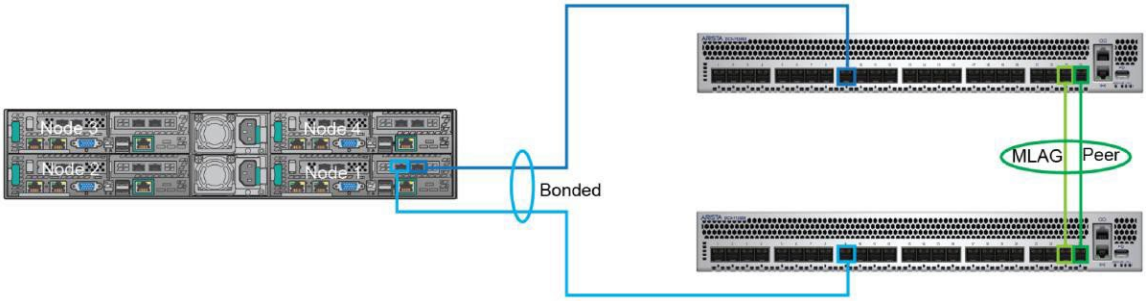

Figure 9. Gen1/2 node connectivity by MLAG

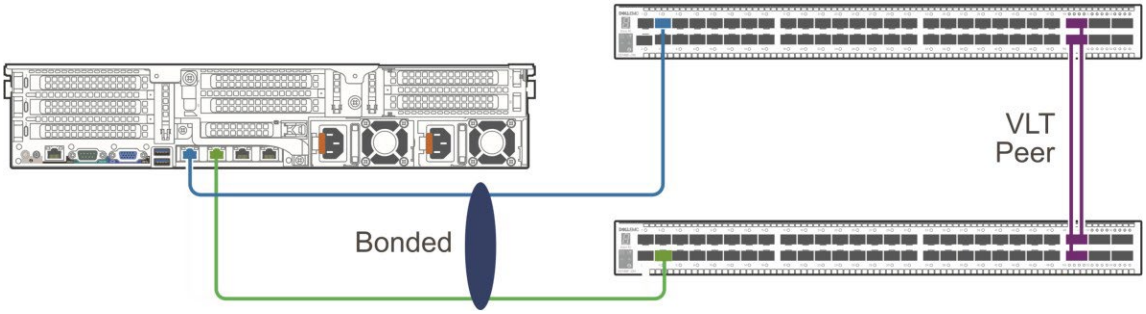

Figure 10. Gen3 node connectivity by VLT

The following terminal output displays snippets of the basic configuration files for the Gen1/2 and Gen3 node. LACP is a protocol that builds LAGs dynamically by exchanging information in Link Aggregation Control Protocol Data Units (LACDUs) relating to the link aggregation between the LAG. LACP sends messages across each network link in the group to check if the link is still active, resulting in faster error and failure detection. The port channels on each switch are MLAG-ed together and can be visible from the nodes. They are configured to allow for fast connectivity using the `spanning tree portfast` command. This command places the ports in forwarding state immediately as opposed to the transition states of listening, learning, and then forwarding, which can cause 15 seconds of delay. Port channels are also set to `lacp fallback` to allow all ports within the port channel to fall back to individual switch ports. When the node's ports are not yet configured as LAG, this setting allows for PXE booting from public switch ports of the nodes and forwarding of traffic.

```
!Gen2 node
interface Ethernet9
   description MLAG group 1
   channel-group 1 mode active
   lacp port-priority 1
!
interface Port-Channel1
   description Nile Node01
(Data) MLAG 1
   port-channel lacp fallback
   port-channel lacp fallback
timeout 1
   mlag 1
   spanning-tree portfast
spanning-tree bpduguard enable
```

```
!Gen3 node
interface ethernet1/1/9
 description "VLT Group 9"
 no shutdown
 channel-group 9 mode active
 no switchport
 mtu 9216
 flowcontrol receive off
 lacp port-priority 1000
!
interface port-channel9
description "Nile Node09 (Data)
VLT 9"
no shutdown
switchport mode trunk
switchport access vlan 1
mtu 9216
lacp fallback enable
lacp fallback preemption disable
vlt-port-channel 9
spanning-tree port type edge
```

The data switches are preconfigured on ECS supported switches. The configuration files for the data switches are on each node in directory. For example, Gen3 Dell switch configuration files are located inside `/usr/share/emc-dell-firmware/config/ecs/`.

Customer uplink configuration

Any networking device supporting Static Link Aggregation Group or IEEE 802.3ad Link Aggregation Control Protocol (LACP) can connect to the MLAG switch pair. With Static Link Aggregation, all settings are defined on all participating LAG components, whereas LACP sends messages across each link in the group to check their state. An advantage of LACP over Static Link Aggregation is faster error or failure detection and handling.

Each Gen1/2 Arista public data switch has eight ports available to connect to the customer network, providing sixteen uplink ports per rack. Gen3 Dell switches each have eight 10/25 GbE and four 100 GbE ports, providing either sixteen or eight uplink ports per rack. For complete details, including switch configuration examples, see the appropriate (Gen1/2 or Gen3) Networks Guide for ECS Hardware.

For Gen1/2 appliances, similarly to the ports used for the node, the eight uplink ports on each of the data switches are configured as a single LACP/MLAG interface. This configuration is shown in the code output that follows. The port channels are also configured to be in *lacp fallback* mode for customers who are unable to present LACP to the ECS rack. This mode is activated only if no LACP is detected by the protocol. If there is no LACP discovered between the customer link and the ECS switches, the lowest active port will be activated and all other linked ports in the LAG will be disabled until a LAG is detected. At this point, there is no redundancy in the paths.

In addition, the data switches are not configured to participate in the customer's spanning tree topology. They are presented as edge or host devices because a single LAG is

created for the eight ports in each switch. It also simplifies the setup of the ECS switches in the customer network. Here is the output for public switches basic configuration file:

```
!Gen1/2 node
interface Ethernet1
   description MLAG group 100
   channel-group 100 mode active
   lacp port-priority 1
!
interface Port-Channel100
   description Customer Uplink
(MLAG group 100)
   port-channel lacp fallback
   port-channel lacp fallback
timeout 1
   spanning-tree bpdufilter
enable
   mlag 100
```

```
!Gen3 node
interface ethernet1/1/41
 description "Customer Conn1"
 no shutdown
 channel-group 100 mode active
 no switchport
 mtu 9216
 flowcontrol receive off
!
interface port-channel100
 description "SFP Customer
Connect"
  no shutdown
  switchport mode trunk
  switchport access vlan 1
  mtu 9216
  vlt-port-channel 100
```

Connections from the customer network to the data switches can be linked in several different ways. For instance, they can be linked as a single link, as a multi-link to a single switch using LACP, or as multi-link to multiple switches using a multiple-switch LACP protocol such as Arista MLAG, Dell VLT, or Cisco vPC. Customers must provide the necessary connection information to establish communication to the nodes in the rack.

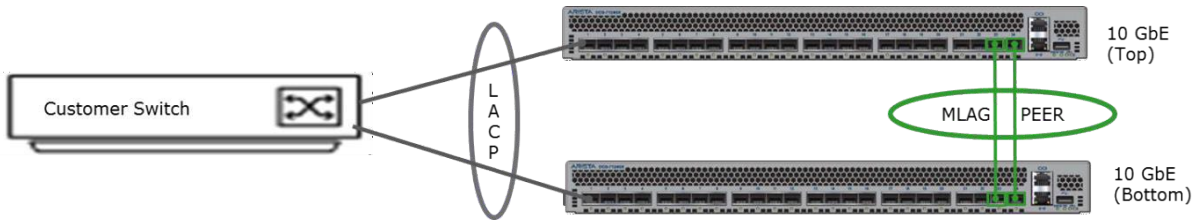

Figure 11. Single customer switch with multiple links

The following code shows an example of a two-port LAG for an ECS public switch and Cisco single switch with multiple links:

```
!Gen1/2 Arista configuration
interface Ethernet 1-2
channel-group 100 mode active
```

```
!Gen3 Dell configuration
interface ethernet1/1/41-1/1/42
channel-group 100 mode active
no switchport
mtu 9216
```

```
!Customer Cisco configuration
interface Ethernet1/1
channel-group 100 mode active
```

```
interface Ethernet1/2
channel-group 100 mode active
```

Figure 12 shows a multiple port uplink to multiple switches with a LAG configuration. A better approach would be to configure more than two links per ECS switch, as shown in Figure 13. Spread the links in a bowtie fashion (links on each customer switch should be distributed evenly between the data switches) for redundancy and optimal performance during failures or scheduled downtime.

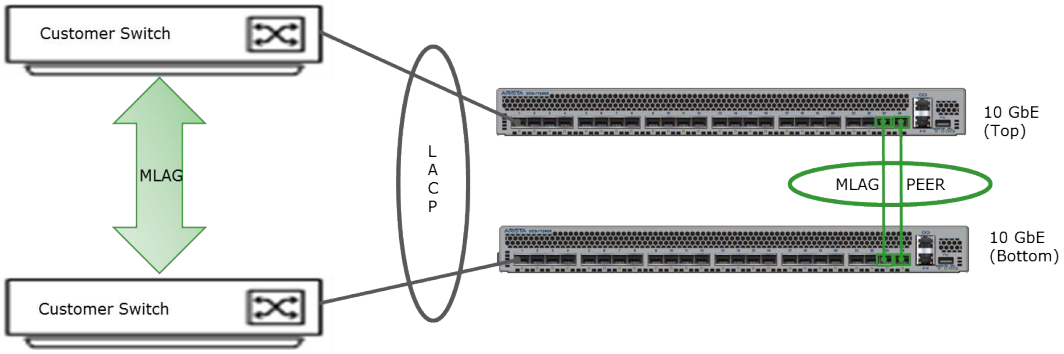

Figure 12. Multiple customer switches with single link per switch

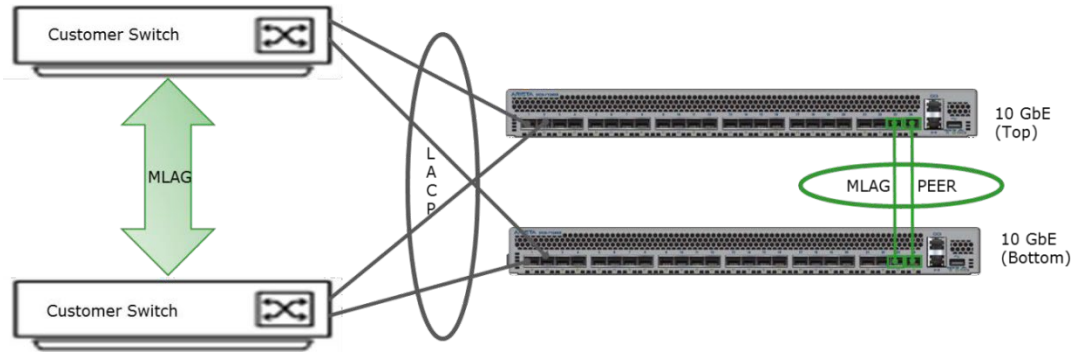

Figure 13. Multiple customer switches with multiple links per switch

In either of these configurations, both port channels must be connected using a multiswitch LAG protocol such as Arista MLAG or Cisco virtual Port Channel (vPC) to connect to the ECS MLAG switch pair port channel. Also, customers must create port channels using LACP in active or passive mode on all switches participating in the multiswitch LAG. The following output shows sample configurations for Arista and Cisco with multiswitch LAG protocol definitions. The vPC or MLAG numbers on each switch must match to create a single port channel group.

```
!Gen1/2 Arista Configuration
!Switch A
interface Ethernet 1-2
channel-group 100 mode active
mlag 100

!Switch B
interface Ethernet 1-2
channel-group 100 mode active
mlag 100
```

```
!Gen3 Dell EMC Configuration
!Switch A
interface ethernet1/1/41-1/1/42
channel-group 100 mode active
interface port-channel100
vlt-port-channel 100

!Switch B
interface ethernet1/1/41-1/1/42
channel-group 100 mode active
interface port-channel100
vlt-port-channel 100
```

```
!Customer Cisco Configuration
!Switch A
interface Ethernet1/1
channel-group 100 mode active
interface Ethernet1/2
channel-group 100 mode active
interface port-channel 100
vpc 100

!Switch B
interface Ethernet1/1
channel-group 100 mode active
interface Ethernet1/2
channel-group 100 mode active
interface port-channel 100
vpc 100
```

Best practices:

- For multiple links, set up LACP on the customer switch. If LACP is not configured on customer switches to ECS switches, one of the data switches will have the active connection or connections. The port or ports connected to the other data switch will be disabled until a LAG is configured. The connection or connections on the other switch will only become active if one switch goes down.
- Balance the number of uplinks from each switch for proper network load balancing to the ECS nodes.
- When using two customer switches, you must use multiswitch LAG protocols.
- For multirack environments of Gen1/2 and Gen3 EX300, EX500, EX3000 and EX5000 systems, consider using an aggregation switch to keep internode traffic separated from customer core network.

Note: Gen1/2 switches are configured not to participate in the customer's spanning tree topology. Gen3 switches are configured to participate in the customer's spanning tree topology with Rapid Spanning Tree Protocol (rstp). For Gen3 switch configuration details, see the *ECS Networks Guide for EX300 and EX3000 (EX-Series) Hardware* documentation.

Network configuration custom requests

Customers might have needs that require modifications to ECS basic configuration files for the data switches. For instance, customers might require physical network isolation between traffic types for security purposes. The following figure is an example of multiple ports uplinked to multiple domains. In this setup, it changes in the data switches basic configuration files would be needed to support two LAGs on the uplinks and to change VLAN membership for the LAGs.

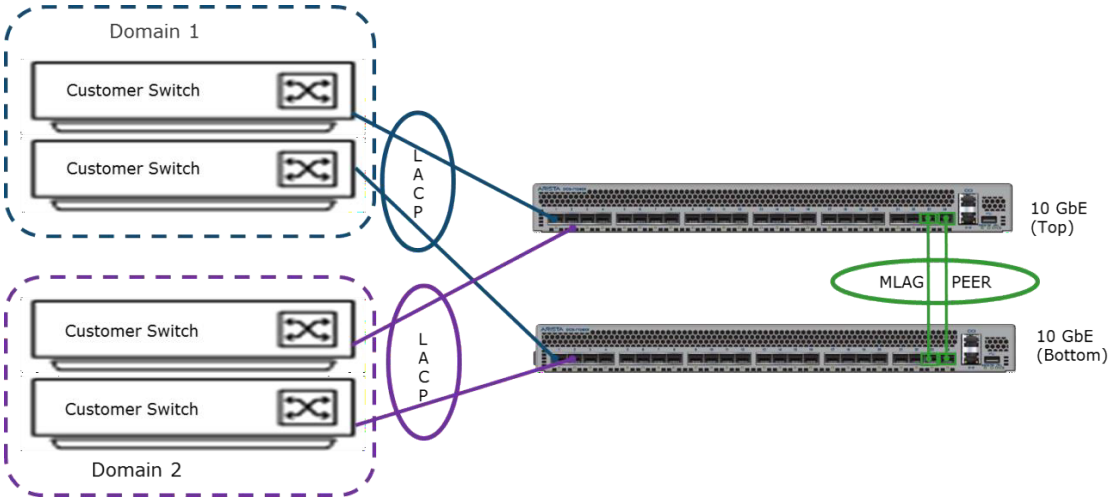

Figure 14. Multiple ports uplinked to multiple domains

Another example is if the customer needs to configure the uplinks for a specific VLAN. The VLAN membership should only be changed if the customer requirement is to set the uplink ports to VLAN trunk mode. Only the port channels for the uplink and nodes need to be changed to set up the VLAN. The following output shows an example of how to change the VLAN membership. Both data switches would need to have the same VLAN configuration.

```
# Create new vlan
vlan 10
exit
# change to vlan trunk mode for uplink
interface port-channel100
switchport mode  trunk
switchport trunk allowed vlan 10
# change vlan membership for access port to the nodes
interface port-channel1-12
switchport access vlan 10
copy running-config startup-config
```

Best practices:

- For custom configurations, see the document ECS Appliance-Special Feature Configuration Support.

Network configuration for HDD and All Flash Array (AFA) intermix node

Starting with ECS version 3.7, ECS supports HDD and AFA disks co-existing in a VDC. With this usage, deploying a load balancer in front of the ECS system but not configuring it properly might result in negative performance impacts on front-end requests.

To ensure proper load-balancer configuration, use the following the best practices:

- Ensure that each load balancer manages a type of storage pool. The client should go to the AFA bucket through the AFA load balancer and the HDD bucket through the HDD load balancer.

- If the load balancer has multiple front ends, each front end can forward requests to the specified type of storage pool. For example, HAProxy can configure multiple front ends with different port numbers.
- If the load balancer has one front end, it can forward requests to a different type of storage pool based on the data in the request. For example, HAProxy can forward requests to different back ends based on the hostname in the HTTP request.

Internal private network

The internal private network, also known as the Nile Area Network (NAN), is mainly used for maintenance and management of the ECS nodes and switches within a rack and across racks. Ports on the management switch can be connected to another management switch on another rack, creating a NAN topology. From these connections, nodes from any rack or segment can communicate to any other node within the NAN. The management switch is split in different LANs to separate the traffic to specific ports on the switch for segment-only traffic, cluster traffic, and customer traffic to RMM:

- **Segment LAN:** Includes nodes and switches within a rack
- **Private.4 LAN:** Includes all nodes across all racks
- **RMM/iDRAC LAN:** Includes uplink ports to customer LAN for RMM/iDRAC access from customer's network

NAN topologies

The NAN is where all maintenance and management communications traverse within a rack and across racks. A NAN database contains information such as IP addresses, MAC addresses, node name, and ID on all nodes within the cluster. This database is locally stored on every node and is synchronously updated by the primary node using the `setrackinfo` command. Information on all nodes and racks within the cluster can be retrieved by querying the NAN database. One command that queries the NAN database is `getrackinfo`.

The racks are connected to the management switches on designated ports. These connections allow nodes within the segments to communicate with each other. There are different ways to connect the racks or rack segments. Each rack segment is specified by a unique color during installation, thus identifying the racks within the cluster.

The following figure shows a simple topology linearly connecting the segments through ports of the management switches in a daisy-chain fashion. The disadvantage of this topology is that when one of the physical links breaks, there is no way to communicate to the segment or segments that has been disconnected from the rest of the segments. This event causes a "split-brain" issue in NAN and forms a less reliable network.

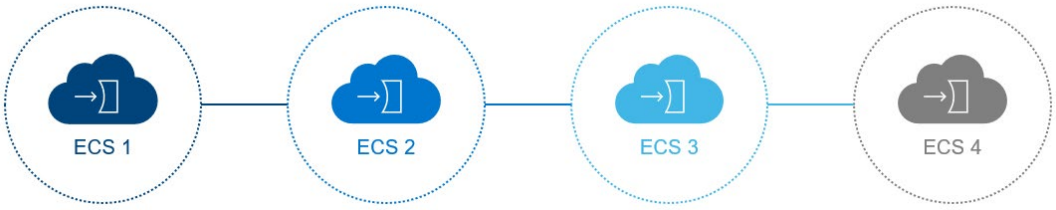

Figure 15. Linear or daisy chain topology

Another way to connect the segments is in a ring topology, as shown in the following figure. The advantage of the ring topology over the linear topology is that two physical links would need to be broken to encounter the split-brain issue.

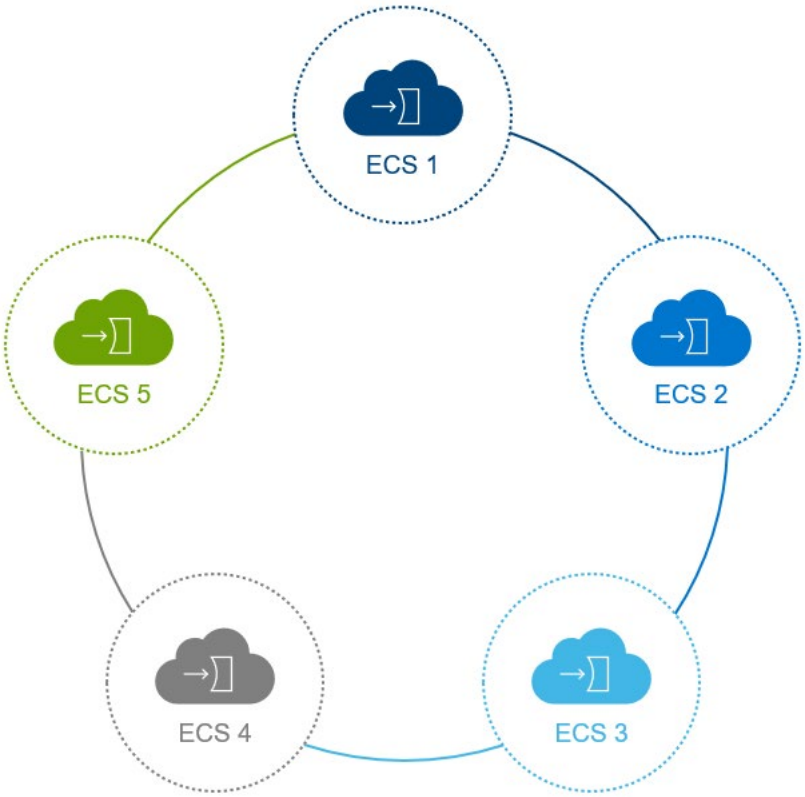

Figure 16. Ring topology

For large installations, the split-brain issue in the ring or linear topologies could be problematic for the overall management of the nodes. A star topology is recommended for an ECS cluster where there are ten or more racks or customers wanting to reduce the issues that ring or linear topologies pose. In the star topology, an aggregation switch, as shown in the following figure, must be added and would be an extra cost; however, the star topology is the most reliable among the NAN topologies.

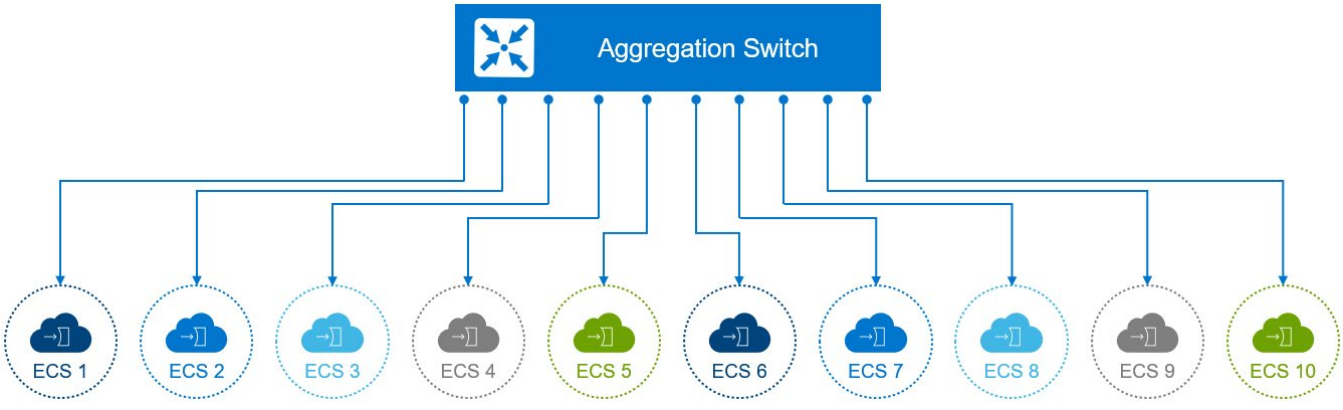

Figure 17. Star topology

Best practices:

- Do not use linear topology.
- For large installations of ten or more ECS racks, a star topology is recommended for better failover.
- EXF900 uses dedicated aggregation switches for ECS racks. Use port 53-56 on each node for connections.

Segment LAN

The segment LAN logically connects nodes and switches within a rack to a LAN identified as VLAN 2. This LAN consists of designated ports on the management switch or switches and are referred to as the "blue network**."** All traffic is limited to members of this segment for ease of management and isolation from the customer network and other segments within the cluster. The Ethernet ports on the nodes are configured with a private IP address derived from the segment subnet and node ID number. Thus, the IP address is of the form `192.168.219.{NodeID}`. The IPs are *not* routable, and packets are untagged. These addresses are reused by all segments in the cluster. To avoid confusion, do not use these IP addresses in the topology file required when installing the ECS software on the nodes. Several IP addresses are reserved for specific uses:

- **192.168.219.254:** Reserved for the primary node within the segment. Recall from the previous section that there is a primary node designated to synchronize the updates to the NAN database.
- **192.168.219.250:** Reserved for the private switch (bottom).
- **192.168.219.251:** Reserved for the private switch (top).
- **192.168.219.252:** Reserved for the public switch (bottom).
- **192.168.219.253:** Reserved for the public switch (top).
- **169.254.255.252:** Reserved for the aggregation switch for EXF900 (bottom).
- **169.254.255.253:** Reserved for the aggregation switch for EXF900 (top).

Note: Gen1/2 only have one private switch named turtle with 192.168.219.251.

Best practices:

- For troubleshooting a Gen1/2 node, administer the node through the segment LAN (connect a laptop to port 24) to not interfere with configurations of other segments within the cluster.
- For troubleshooting a Gen3 node, the administrator can access the cluster by the tray that is connected with private switch named fox in port 34, or connect a laptop to port 36 in the fox switch.

Private.4 LAN

Multiple segment LANs are logically connected to create a single cluster LAN for administration and access to the entire cluster. Designated interconnect ports on management switches provide interconnectivity between management switches. All members will tag their IP traffic with VLAN ID 4 and communicate through the IPv4 link local subnet. During software installation, all nodes in the rack are assigned a unique

color number. The color number acts as the segment ID and is used together with the node ID to consist of the new cluster IP address for every node in the cluster. The IP addresses of the nodes in the cluster LAN will be in the form of `169.254.{SegmentID}.{NodeID}`. This unique IP address would be the recommended IP address to specify in the topology file for the nodes within the cluster.

Best practices:

- ECS does not support IPv6, so do not enable IPv6 on these switches or send IPv6 packets.
- If troubleshooting a segment within the cluster, administer the segment through the segment LAN so that the configuration of the entire cluster is not affected.
- Use the IP address in the topology file to provide a unique IP for all nodes within the cluster.

RMM/iDRAC access from customer network (optional)

RMM/iDRAC access from the customer network is optional. It is recommended to meet specific customer requirements. A relevant use of the RMM/iDRAC connection would be for ECS software-only deployments where the hardware is managed and maintained by customers. Another use is when customers have a management station in which they would require RMM/iDRAC access to all hardware from a remote location for security reasons.

Note: For Gen1/2 ECS, RMM connects to the private switch named turtle; for Gen3 ECS, iDRAC connects to the private switch named fox.

For a Gen1/2 node, to allow for RMM connections from customer switch, ports 51 and 52 on the management switch are configured in a hybrid mode. This configuration allows the ports to handle both tagged and untagged traffic. In this setup, the ports can be used for multiple purposes. The uplinks to the customer switch are on VLAN 6, and packets are untagged.

Best practices:

- Providing RMM to customers is optional and only available in Gen1/2 appliances.
- Determine if customer access to RMM is required before configuration.
- Ensure that NAN traffic on VLAN 4 does not leak to the customer network when adding RMM access to the customer network.

ECS network separation

Overview

Network separation allows for the separation of different types of network traffic for security, granular metering, and performance isolation. The types of traffic that can be separated include:

- **Management:** Traffic related to provisioning and administering through the ECS portal and traffic from the operating system such as DNS, NTP, and SRS
- **Replication:** Traffic between nodes in a replication group

- **Data:** Traffic associated with data

There is a mode of operation called the network separation mode. When enabled during deployment, each node can be configured at the operating system level with up to three IP addresses or logical networks for each of the different types of traffic. This feature has been designed for flexibility by either creating three separate logical networks for management, replication, and data, or combining them to create two logical networks. For instance, management and replication traffic is in one logical network and data traffic in another logical network.

ECS implementation of network separation requires each network traffic type to be associated with specific services and ports. For instance, the portal services communicate through ports 80 or 443, so these ports and services will be tied to the management logical network. The following table highlights the services fixed to a logical network. For a complete list of services and their associated ports, refer to the most recent version of the ECS Security Configuration Guide.

Table 1. Services associated with logical network

Services	Logical network
ECS Portal, provisioning, metering and management API, SSH, DNS, NTP, AD, and SMTP	Management network (public.mgmt)
Data across NFS, Object, and HDFS	Data network (public.data) CAS-only data network (public.data2)
Replication data and XOR	Replication network (public.repl)
SRS (Dell Secure Remote Services)	Based on the network to which the SRS Gateway is attached (public.data or public.mgmt)

Note: Starting with ECS 3.7, ECS allows S3 data access on both public.data and public.data2 networks.

Network separation is achievable logically using virtual IP addresses, using VLANs or physically using different cables. The command `setrackinfo` is used to configure the IP addresses and VLANs. Any switch-level or client-side VLAN configuration is the customer's responsibility.

Network separation configurations

In addition to the default network configuration, networks can be partially separated, or all separated using the following configurations:

- **Standard (default):** All management, data, and replication traffic in one VLAN referred to as public.
- **Partial (Dual):** Two VLANs where one VLAN is the default public, which can have two traffic types, and another VLAN for any traffic not defined in the public VLAN.
- **Partial (Triple):** One VLAN for public VLAN and two VLANs where one traffic type is placed in the public VLAN and two different VLANS are defined for the other two traffic types not in public.

Network separation configures VLANs for specific networks and uses VLAN tagging at the operating system level. There is an option to use virtual or secondary IPs where no VLAN is required; however, it does not actually separate traffic but instead just provides another access point. For the public network, traffic can be tagged at the switch level. At a minimum, the default gateway is in the public network and all the other traffic can be in separate VLANs. If needed, the default public VLAN can also be part of the customer's upstream VLAN; in this case, the VLAN ID for public must match the customer's VLAN ID.

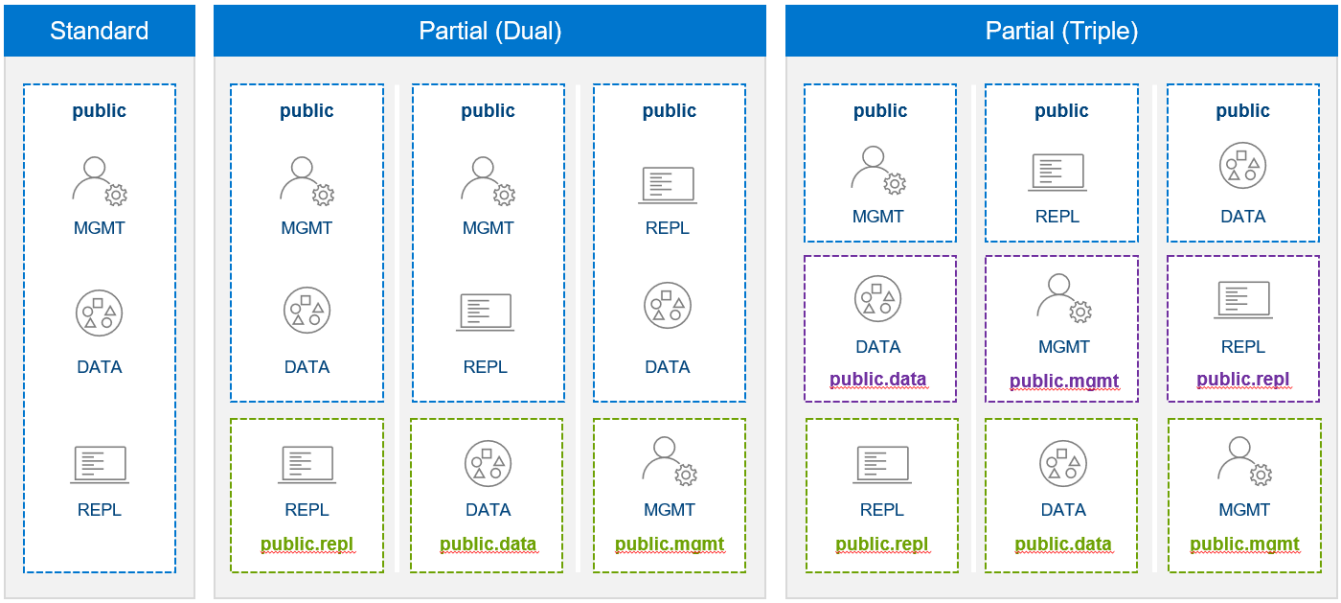

Figure 18. Examples of network separation configurations

Network separation is conducted during ECS installation before the installation of Hardware Abstraction Layer (HAL) or in an existing ECS environment. It requires static IP addresses. Planning for network separation requires decisions about how traffic should be separated in VLANs and the static IP addresses required, and the subnet and gateway information must be determined. After network separation has been completed, virtual interfaces are created for the VLANs. The interface configuration files will be of the form `ifcfg- public.{vlanID}`. The following input is an example:

```
admin@memphis-pansy:/etc/sysconfig/network> ls ifcfg-public*
ifcfg-public ifcfg-public.data ifcfg-public.mgmt ifcfg-public.repl
```

The operating system presents the interfaces with a managed name in the form of `public.{trafficType}` such as `public.mgmt`, `public.repl`, or `public.data`, as shown by the `ip addr` command output in the following code:

```
admin@memphis-pansy:/etc/sysconfig/network> ip addr | grep public
inet 10.10.20.55/24 scope global public.mgmt
40: public.repl@public: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500
qdisc noqueuest
inet 10.10.30.55/24 scope global public.repl
41: public.data@public: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500
qdisc noqueuest
inet 10.10.10.55/24 scope global public.data
```

The HAL searches for these managed names based on the **active_template.xml** in **/opt/emc/hal/etc**. It finds those interfaces and presents them to the Fabric layer. The following figure shows the `cs_hal list nics` output:

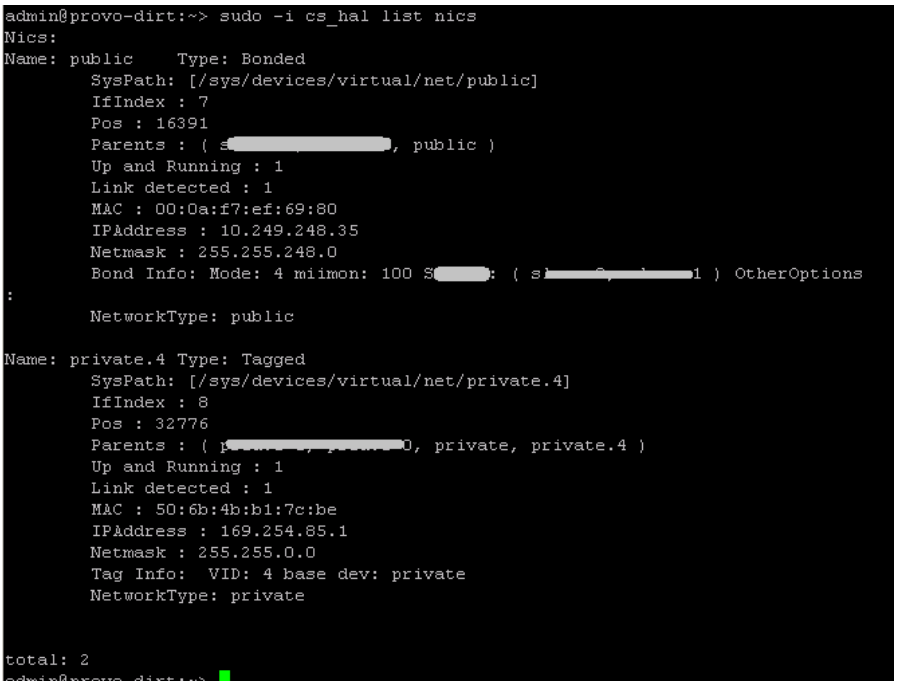

Figure 19. Output of NICs

The HAL gives the information to the Fabric layer, which creates a JavaScript Object Notation (JSON file) with IP addresses and interface names and supplies this information to the object container. The following output from the Fabric Command Line (fcli) shows the format of the JSON structure:

```
admin@memphis-pansy:/opt/emc/caspian/fabric/cli> bin/fcli agent
node.network
{
  "etag": 12,
  "network": {
    "mgmt_interface_name": "public.mgmt",
    "mgmt_ip": "10.10.20.55",
    "data_interface_name": "public.data",
    "data_ip": "10.10.10.55",
    "hostname": "memphis-pansy.ecs.lab.emc.com",
    "private_interface_name": "private.4",
    "private_ip": "169.254.78.17",
    "public_interface_name": "public",
    "public_ip": "10.245.132.55",
    "replication_interface_name": "public.repl",
    "replication_ip": "10.10.30.55"
  },
  "status": "OK"
}
```

The mapped content of this JSON structure is placed in an object container in the file */host/data/network.json*, as shown in the following terminal output.

```
{
   "data_interface_name": "public.data",
   "data_ip": "10.10.10.55",
   "hostname": "memphis-pansy.ecs.lab.emc.com",
   "mgmt_interface_name": "public.mgmt",
   "mgmt_ip": "10.10.20.55",
   "private_interface_name": "private.4",
   "private_ip": "169.254.78.17",
   "public_interface_name": "public",
   "public_ip": "10.245.132.55",
   "replication_interface_name": "public.repl"
   "replication_ip": "10.10.30.55"
}
```

Network separation in ECS uses source-based routing to specify the route that packets take through the network. In general, the path that packets come in on will be the same path going out. Based on the IP rules, the local node that originates the packet looks at the IP and looks at the local destination; if it is not local, it looks at the next destination. Using source-based routing reduces static routes that need to be added.

ECS switch configuration for network separation

Depending on customer requirements, network separation might require modification of the basic configuration files for the data switches. This section explores examples of different network separation implementations in the switch level, such as the default, single domain, single domain with public set as a VLAN, and physical separation.

Standard (default)

The default settings use configuration files that are bundled with ECS. In this scenario, there is no VLAN and there is only the public network. Also, there is no tagged traffic in the uplink connection. All ports are running in access mode. The following table and figure provide an example of a default ECS network setup with customer switches.

Table 2. Standard default switch configuration

Interface	VLAN ID	Tagged	Uplink connection
Public	None	No	MLAG:po100 No tagged traffic

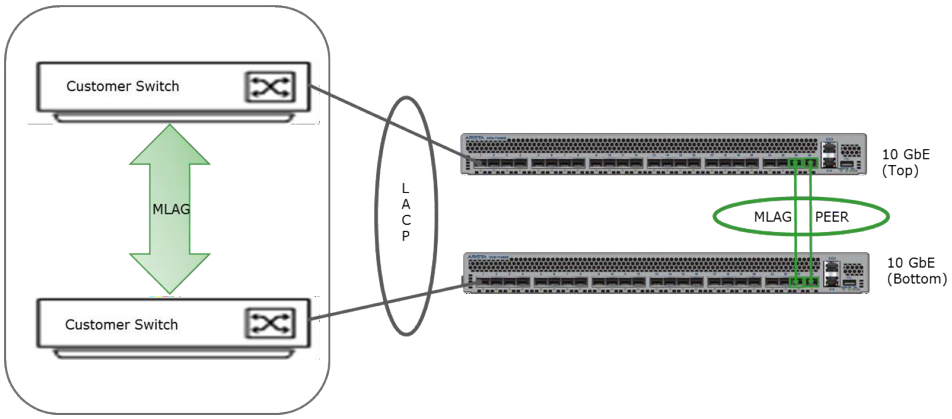

Figure 20. Standard default switch setup

Single domain

In a single domain, a LACP switch or an LACP/MLAG switch pair is configured on the customer side to connect to the ECS MLAG switch pair. Network separation is achieved by specifying VLANs for the supported traffic types. In the example in the following table and figure, data and replication traffic are separated into two VLANs and the management stays in the public network. The traffic on the VLANs will be tagged at the operating system level with their VLAN ID, which in this case is 10 for data and 20 for replication traffic. The management traffic on the public network is not tagged.

Table 3. An example of single domain switch configuration

Interface	VLAN ID	Tagged	Uplink connection
Public	None	No	MLAG:po100 All named traffic tagged
Data	10	Yes	
Repl	20	Yes	

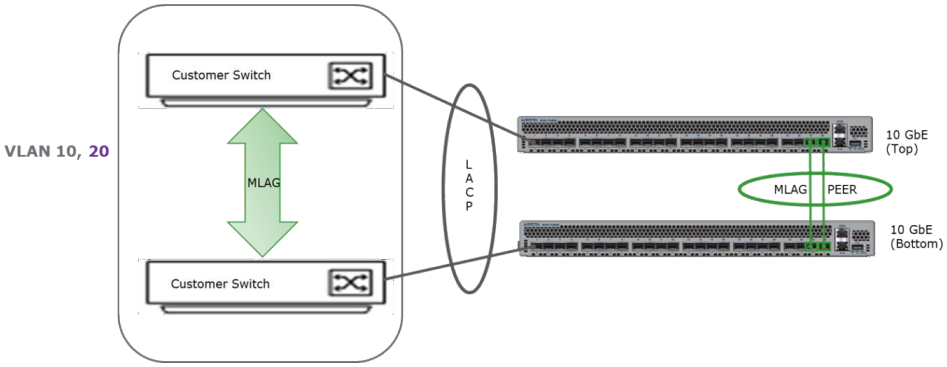

Figure 21. An example of a single domain switch with two VLANs

Both data switch configurations files must be modified to handle the VLANs in the preceding example. The following terminal output shows how this can be specified for Arista switches. Items to note from the configuration file include:

- The switchports have been modified from access to trunk.
- VLANs 10 and 20 created to separate data and replication traffic are allowed. They also need to be created first.

- VLAN 1 corresponds to the public.
- If port channels are used, it will supersede and ignore Ethernet-level configurations.

This example shows a single domain's switch settings with two VLANs for public switches:

```
vlan 10, 20
interface po1-12
switchport trunk native vlan 1
switchport mode trunk
switchport trunk allowed vlan 1,10,20

!For 7050S-52 and 7050SX-64, the last port channel is 24

interface po100
switchport mode trunk
switchport trunk allowed vlan 1, 10,20
```

Single domain and public VLAN

Customers might want to have the public network in a VLAN. In this scenario, the traffic going through the public network will be tagged at the switch level and the other VLANs will be tagged at the operating system level. The following table and figure provide switch and configuration details for a single domain with public VLAN setup.

Table 4. An example of single domain switch configuration

Interface	VLAN ID	Tagged	Uplink connection
Public	100	Yes (switch)	MLAG:po100 All traffic tagged
Data	10	Yes (operating system level)	
Repl	20	Yes (operating system level)	

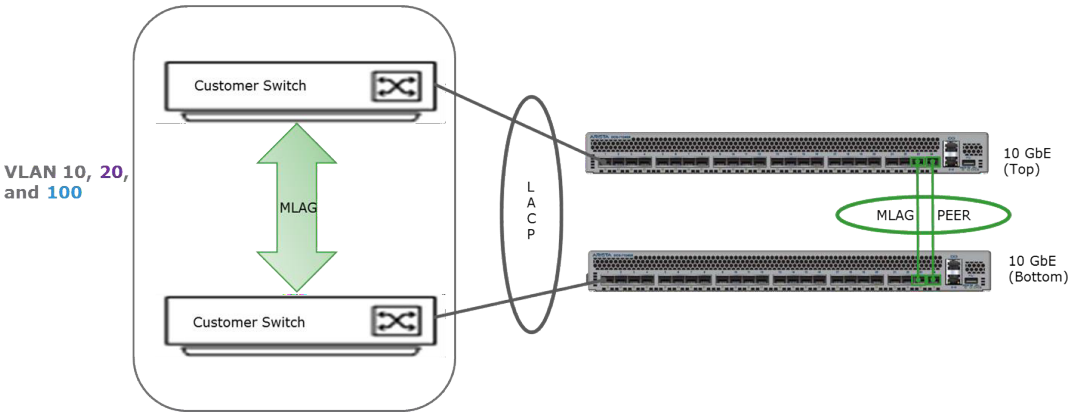

Figure 22. An example of single domain public VLAN switch setup

The settings within the configuration files of the data switches must be changed to include all the VLANs specified for network separation. As shown in the following terminal output, the native VLAN is updated to match the customer VLAN for public. In this example, the public VLAN is identified as VLAN 100.

Here is an example of code that shows a single domain with two VLANs and public VLAN settings for public switches:

```
vlan 10, 20, 100
interface po1-12
switchport trunk native vlan 100
switchport mode trunk
switchport trunk allowed vlan 10,20,100
interface po100
switchport mode trunk
switchport trunk allowed vlan 10,20,100
```

Physical separation

For physical separation, a setup might include multiple domains on the customer network defined for each type of traffic. The following table and figure provide an example of the setup and details. As shown in the table, the public network is not tagged and will be on port channel 100; data traffic will be on VLAN 10, tagged, and on port channel 101; and replication traffic will be on VLAN 20, tagged, and on port channel 102. The three domains are not MLAG together.

Table 5. An example of physical separation configuration

Interface	VLAN ID	Tagged	Uplink connection
Public	None	No	MLAG:po100
Data	10	Yes	MLAG:po101
Repl	20	Yes	MLAG:po102

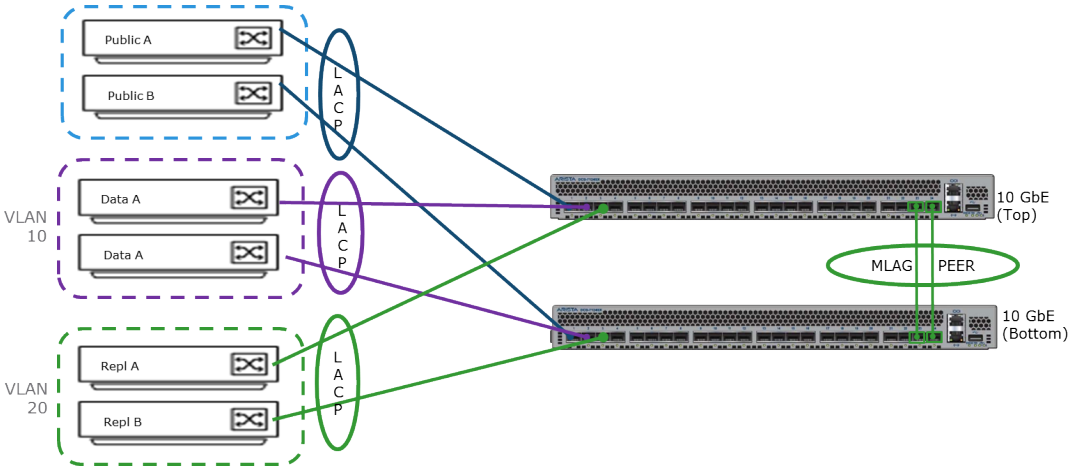

Figure 23. An example of physical separation setup

The following terminal output shows what the settings would be on the data switches for this configuration on Arista switches. Port channel 100 is set up to remove uplink ports 2 through 8, leaving only the first uplink for the public network. Port channel 101 defines the settings for the data traffic, and port channel 102 is for the replication traffic where the corresponding VLANs are allowed and switchport is set to *trunk*. Connections to the data nodes are defined by interface po1-12.

For situations where customers would want the public network on a VLAN, the following table and the subsequent terminal output provide details of a configuration example. In this case, all traffic is tagged. Public is tagged with ID 100, data traffic is tagged with 10, and replication is tagged with 20. Uplink connections and port channel 100 are set up as the trunk, and VLAN 10, 20, and 100 are allowed. The connections to the nodes defined in interface po1-12 are also set accordingly.

Table 6. Physical separation with public VLAN example

Interface	VLAN ID	Tagged	Uplink connection
Public	100	Yes (switch)	MLAG:po100 All traffic tagged
Data	10	Yes	
Repl	20	Yes	

Best practices:

- Network separation is optional. If network separation is used, it is important to determine fit and best configuration.
- Keep the management traffic within the public to reduce the number of static routes.
- Although public can have only the default gateway, have at least one of the traffic types in the public network.
- Do not use virtual IP or secondary addresses for network isolation.

ECS network performance

Network performance is a major factor that can affect the ability of any cloud storage platform to serve data. When architecting or designing the customer network to connect with ECS data switches, there are some considerations to maintain optimal performance. Data, replication, internode traffic, and management traffic (ECS portal, Rest APIs, and traffic to network services such as DNS, and AD) flows through the data switches. A reliable and highly available network is also important.

For production network, a minimum of one uplink per switch to customer switch is required. However, one per switch might not be enough to handle the performance necessary for all traffic, specifically in a multirack and single-site deployment or when one switch fails. Internode traffic in a single-site multirack deployment traverses through one rack, up to the customer network and down to the next rack of switches, in addition to handling traffic associated with data, replication, and management. At a minimum, four uplinks per rack (two links per switch) are recommended for performance and HA. Since both the data switches are peers, if link to either switch is broken, one of the other switches is available to handle the traffic.

Network latency is one of the considerations in multisite or geo-replicated environments. In a multisite configuration, 1000 ms is the recommended maximum latency between two sites.

Understanding workload, deployment, current network infrastructure, requirements, and expected performance is fundamental in architecting ECS network connections. Some additional areas to understand include:

- Multirack ECS deployment
- Multisite or geo-replicated deployment
- Rate of data ingress, average size of objects, and expected throughput per location, if applicable
- Read/write ratio
- Customer network infrastructure such as VLANs, specific switch requirements, traffic isolation requirements, known throughput, or latency requirements

Note: Network performance is only one aspect of overall ECS performance. The software and hardware stack both contribute as well.

Best practices:

- A minimum of four uplinks per rack (two links per switch) is recommended to maintain optimal performance in case one of the switches fails.
- Use enough uplinks to meet any performance requirements.
- Get a good understanding of workloads, requirements, deployment, current network infrastructure, and expected performance.
- When replicating two EXF900 systems across sites, consider the potential performance impacts over the WAN. A large ingest might put a high load on the link, causing saturation or delayed RPO. Also, a user or application might experience higher latency times on remote reads and writes as compared to local requests.

Tools

Introduction

This section describes tools available for planning, troubleshooting, and monitoring of ECS networking.

ECS portal

The WebUI provides a view of network metrics within ECS. For instance, the average bandwidth of the network interfaces on the nodes can be viewed on the Node and Process Health page. The Traffic Metrics page provides read and write metrics at the site level and individual node level. It shows the read and write latency in milliseconds, the read and write bandwidth in bytes/second, and read and write transactions per second.

The Geo-Replication monitor page shows information relating to geo-replication occurring between sites. For instance, the rates and chunks pages provide the current read and write rates for geo-replication and the chunks broken down by user data, metadata, and XOR data pending for replication, by replication group or remote site. The ECS portal also provides a way to filter data based on timeframe to get a historical view of traffic. Note that updates to any rate information in the ECS portal can take some time. For more information about the ECS portal, see the most recent *ECS Administration Guide*.

ECS Designer and planning guide

The ECS Designer is an internal tool to help with streamlining the planning and deployment of ECS. It integrates the ECS Configuration Guide with the internal validation process. The tool is in spreadsheet format, and inputs are color-coded to indicate which fields require customer information. The sheets are ordered in a workflow to guide the architects in the planning.

Note: Contact your account team if you cannot obtain the installation scripts from the ECS Designer team.

Also available is an ECS Planning Guide that provides information about planning an ECS installation and site preparation. It also provides an ECS installation readiness checklist and echoes the considerations discussed in this white paper.

Secure Remote Services

Secure Remote Services (SRS) provides secure two-way communication between customer equipment and Dell Support. It leads to faster problem resolution with proactive remote monitoring and repair. SRS traffic goes through the ECS public network and not the RMM access ports on the ECS private network. SRS can enhance customer experience by streamlining the identification, troubleshooting, and resolution of issues.

Linux or HAL tools

ECS software runs on a Linux operating system. Common Linux tools can be used to validate or get information about ECS network configurations. Some tools useful for this include ifconfig, netstat, and route. HAL tools such as getrackinfo are also useful.

For instance, to validate if network separation configuration is working, running, and filtering, run the `netstat` command for processes that are part of the *object-main* container. The following truncated output of `netstat` shows the open ports and processes using it, such as the *georeceiver* used by *object-main* container to pass around the data and *nginx* directs requests for the user interfaces.

The following code shows an example of truncated output of `netstat` to validate network separation:

```
admin@memphis-pansy:/opt/emc/caspian/fabric/agent> sudo netstat -nap | grep
georeceiver | head -n 3

tcp 0      0 10.10.10.55:9098   :::*       LISTEN     40339/georeceiver
tcp 0      0 10.10.30.55:9094   :::*       LISTEN     40339/georeceiver
tcp 0      0 10.10.30.55:9095   :::*       LISTEN     40339/georeceiver

admin@memphis-pansy:/opt/emc/caspian/fabric/agent> sudo netstat -nap | grep
nginx | grep tcp
```

Another tool that can validate the setup of network separation is the `domulti wicked ifstatus public.<traffic type>` command, which shows the state of the network interfaces. The state of each interface should be *up*. Here is the command being used to check the public.data interface.

```
admin@boston-pansy:~> domulti wicked ifstatus public.data

192.168.219.9
======================================
public.data up
    link:    #14, state up, mtu 1500
    type:    vlan public[1000], hwaddr 00:1e:67:e3:1c:46
    config:  compat:suse:/etc/sysconfig/network/ifcfg-public.data
    leases:  ipv4 static granted
    addr:    ipv4 10.10.10.35/24 [static]

192.168.219.10
======================================
public.data up
    link:    #13, state up, mtu 1500
    type:    vlan public[1000], hwaddr 00:1e:67:e3:28:72
    config:  compat:suse:/etc/sysconfig/network/ifcfg-public.data
    leases:  ipv4 static granted
    addr:    ipv4 10.10.10.36/24 [static]

192.168.219.11
======================================
public.data up
    link:    #13, state up, mtu 1500
    type:    vlan public[1000], hwaddr 00:1e:67:e3:29:7e
    config:  compat:suse:/etc/sysconfig/network/ifcfg-public.data
    leases:  ipv4 static granted
    addr:    ipv4 10.10.10.37/24 [static]

192.168.219.12
======================================
public.data up
    link:    #11, state up, mtu 1500
    type:    vlan public[1000], hwaddr 00:1e:67:e3:12:be
    config:  compat:suse:/etc/sysconfig/network/ifcfg-public.data
    leases:  ipv4 static granted
    addr:    ipv4 10.10.10.38/24 [static]
```

Some of the HAL tools were discussed in ECS network separation; however, here is an output of `getrackinfo -a` that lists the IP addresses, RMM MAC, and public MAC across nodes within an ECS rack.

```
admin@hop-u300-12-pub-01:~> getrackinfo -a
Node private      Node              Public                                  RMM
Ip Address        Id       Status   Mac                 Ip Address          Mac
Ip Address          Node Name
===============   ======   ======   =================   =================   =================
=================   =========
192.168.219.1     1        MA       00:1e:67:93:c6:1c   10.246.150.179      00:1e:67:4d:ae:f4
10.246.150.155      provo-green
192.168.219.2     2        SA       00:1e:67:93:ca:ac   10.246.150.180      00:1e:67:4d:a3:1e
10.246.150.156      sandy-green
192.168.219.3     3        SA       00:1e:67:93:c2:5c   10.246.150.181      00:1e:67:4f:aa:b8
10.246.150.157      orem-green
192.168.219.4     4        SA       00:1e:67:93:c7:c4   10.246.150.182      00:1e:67:4f:ab:76
10.246.150.158      ogden-green
192.168.219.5     N/A      noLink   N/A                 N/A                 N/A
N/A                 N/A
192.168.219.6     N/A      noLink   N/A                 N/A                 N/A
N/A                 N/A
192.168.219.7     N/A      noLink   N/A                 N/A                 N/A
N/A                 N/A
```

Best practices:

- Use ECS Designer to help with the planning of the ECS network with the customer network.
- Use the ECS portal to monitor traffic and alerts.
- Set up and enable SRS for streamlining of issues to the Dell Support team.

Network services

Certain external network services need to be reachable by the ECS system:

- **Authentication Providers (optional):** System and namespace administrative users can be authenticated using Active Directory and LDAP. Swift object users can be authenticated using Keystone. Authentication providers are not required for ECS. ECS has integrated local user management. Local users on ECS are not replicated between VDCs.
- **DNS Server (required):** Domain Name server or forwarder.
- **NTP Server (required):** Network Time Protocol server. For guidance on optimum configuration, see NTP best practices.
- **SMTP Server (optional):** Simple Mail Transfer Protocol Server is used for sending alerts and reporting from the ECS rack.
- **DHCP server (optional):** Only necessary if assigning IP addresses through DHCP.
- **Load Balancer** (optional but **highly** recommended): Evenly distributes client and application load across all available ECS nodes.

Also, the data switch uplinks must reside in the same network or be accessible by the ECS system.

The *ECS General Best Practices* white paper provides additional information about these network services. White papers that discuss how to deploy ECS with vendor-specific load balancers are also available.

Conclusion

ECS supports specific network hardware and configurations plus customer variations and requirements. The switches used as part of the ECS hardware infrastructure provide the backbone for the ECS communication paths to the customer network, node-to-node communication, and node- and cluster-wide management. It is a best practice to architect ECS networking to be reliable, highly available, and performant. There are tools to help with planning, monitoring, and diagnosing the ECS network. Customers are encouraged to work closely with Dell Technologies personnel to obtain the optimal ECS network configuration to meet their requirements.

Technical support and resources

Dell.com/support is focused on meeting customer needs with proven services and support.

The Dell Technologies Info Hub provides expertise that helps to ensure customer success on Dell storage platforms.