Technical Report

NetApp HCI with Mellanox SN2100 and SN2700 Network Switch, Best Practice Cabling Guide

Mike Geroche, NetApp
May 2020 | TR-4836

NetApp

TABLE OF CONTENTS

1 Introduction 4

2 Prerequisites 4

3 NetApp HCI Network Topology 4

4 Mellanox Switch Overview 5

4.1 SN2010 5

4.2 SN2100 5

4.3 SN2700 6

4.4 Breakout Cable Usage 7

5 Cluster Network Planning, Connection Diagrams and Cabling 7

5.1 Minimum HCI Cluster 7

5.2 Small HCI Cluster 9

5.3 Mid-Size HCI Clusters 11

5.4 Large HCI Clusters 11

5.5 Maximum Size HCI Cluster 12

5.6 NetApp Orderable Cabling for Mellanox Switches 13

6 General Switch Configuration, Small Cluster Network Example 14

6.1 General Protocol Requirements 14

6.2 VLAN Requirements 14

6.3 Configure Switch -to-Switch Inter-Peer Link and Multi-Chassis Link Aggregation 15

6.4 Configure Uplink 15

6.5 Configure Management Ports 15

6.6 Configure Storage Ports, iSCSI Network 16

6.7 Configure VM Data, vMotion Network 16

7 Mid-Size Cluster Network Configuration 17

8 Large-Size and Max-Size Cluster Network Configuration 17

9 Sample Network Configuration for a 16 Storage / 32 Compute Node NetApp HCI Cluster 18

9.1 Initial Switch Setup and Onyx OS Update 18

9.2 Switch Programming with the CLI 19

Where to Find Additional Information 22

Acknowledgements 23

LIST OF TABLES

Table 1) Network cables for node to switch and switch to switch connections. 13

LIST OF FIGURES

Figure 1) Simplified 1 Compute / 1 Storage Node, Reference Network Topology for NetApp HCI. 5

Figure 2) SN2010 faceplate and ports. 5

Figure 3) SN2100 faceplate and ports. 6

Figure 4) SN2700 front and back. 6

Figure 5) QSFP Port Splitting on the SN2700 with corresponding port deactivation. 6

Figure 6) QSFP breakout cable for cluster node connections. 7

Figure 7) NetApp HCI minimum size 2 Storage Node / 2 Compute Node Cluster with High-speed Data, Management and OOB Networks with 1RU H610S Storage with H615C Compute nodes. 8

Figure 8) NetApp HCI minimum size 2 Storage Node / 2 Compute Node Cluster with High-speed Data, Management and OOB Networks with 2RU H410S Storage with H410C Compute nodes. 8

Figure 9) Typical 4 Storage / 4 Compute Node HCI Cluster with Data, Management and OOB Networks. 9

Figure 10) Typical 4 Storage / 4 Compute Node HCI Cluster with Data, Management and OOB Networks, 6-Cable Compute Option. 10

Figure 11) 16 Storage / 32 Compute Node HCI Cluster using one pair of SN2100 ToR switches. 11

Figure 12) Large Size HCI Cluster with Maximum number of Compute Nodes (64). 12

Figure 13) Large Size HCI Cluster with Maximum number of Storage Nodes (40). 12

Figure 14) NetApp HCI Maximum Cluster Size, 64 Compute and 40 Storage Nodes. 13

1 Introduction

This document describes the preferred network architecture for integrating the Mellanox SN2000 family of Ethernet switches into the NetApp® HCI system. This information includes network design, cabling, and switch configuration for small, medium, and large NetApp HCI cluster deployments using the Mellanox SN2010, SN2100, and SN2700 switches. Building high-speed data storage network for various NetApp HCI cluster sizes is the primary focus of this document. As such, we assume that medium size and larger HCI clusters will reside in a datacenter or IT space that has an existing 1/10GbE network for the management and out-of-band networks.

2 Prerequisites

This document assumes the following:

- You are familiar with the content of NetApp HCI Networking Quick Planning Guide.
- You have completed the NetApp HCI prerequisites checklist.
- You have not yet deployed the NetApp HCI system software with the NetApp Deployment Engine (NDE).
- You are ready to rack the Mellanox SN2100 or SN2700 switches, and, given the chosen cluster configuration, you have or are ready to purchase the NetApp HCI Mellanox networking kit cables.
- You are ready to or have already racked all NetApp HCI chassis and installed all NetApp HCI nodes.

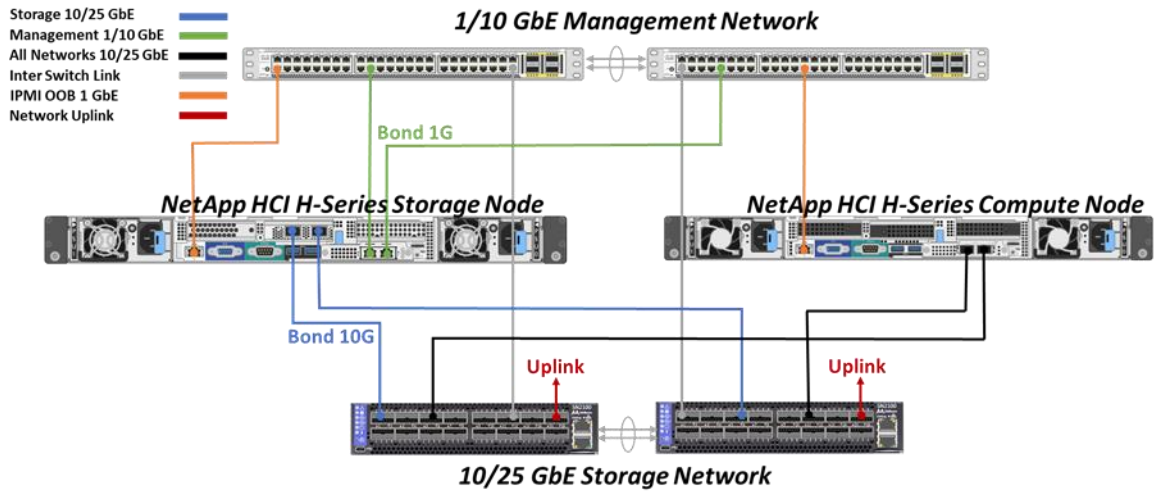
Additional resources and documentation for NetApp HCI can be found in the NetApp Product Documentation center.

For detailed information about Mellanox Ethernet switches, see the Mellanox Scale-Out SN2000 Ethernet Switch Series. This webpage includes links to SN2000 series switch family information and documentation.

3 NetApp HCI Network Topology

NetApp HCI is deployed in clusters that consist of specialized compute and storage resources that are redundantly interconnected to top-of-rack (ToR) network switches. By physically disaggregating the compute and storage nodes, NetApp HCI can independently scale out to meet the needs of the workload. Figure 1 shows a simplified reference network topology for a single storage node and a single compute node.

Figure 1) Simplified one compute and one storage node reference network topology for NetApp HCI.



Note: The minimum cluster size for NetApp HCI is two storage nodes and two compute nodes, and the maximum cluster size is 40 storage nodes and 64 compute nodes.

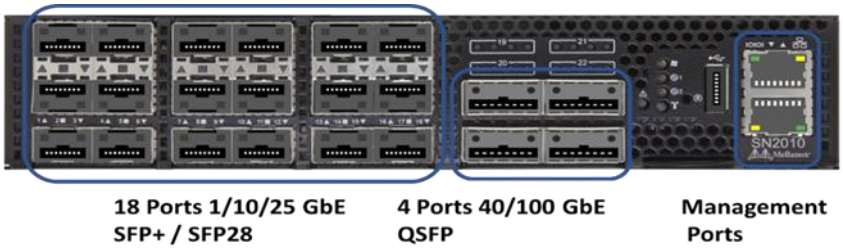
4 Mellanox Switch Overview

The Mellanox SN2000 series of Ethernet switches are ideal for ToR leaf and spine use in NetApp HCI solutions. NetApp HCI requires configuring the switches with the Mellanox Onyx operating system, and airflow is configured in the power cord-to-cable (P2C) direction. The Mellanox Spectrum 1U Switch Systems Hardware User Manual can be found here.

4.1 SN2010

The Mellanox SN2010 (NetApp PN MSN2010-CB2F-NTAP) is a 1RU-high, half-width switch with 18 ports of 1/10/25GbE and 4 ports of 10/25/40/100GbE. The ideal use case is for small cluster size deployments and the inclusion of management network ports (Figure 2).

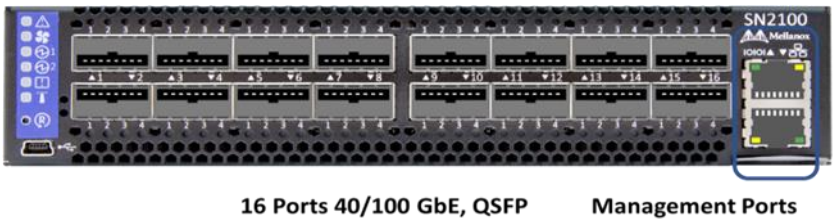
Figure 2) SN2010 faceplate and ports.



4.2 SN2100

The Mellanox SN2100 (NetApp PN MSN2100-CB2F-NTAP) is a 1RU-high, half-width switch with 16 ports of 10/25/40/100GbE in a QSFP transceiver form factor. Its use case is anywhere from small to large cluster sizes. Given that all of the ports are QSFP, 4:1 breakout cables must be used for HCI compute and storage node connections (Figure 3).

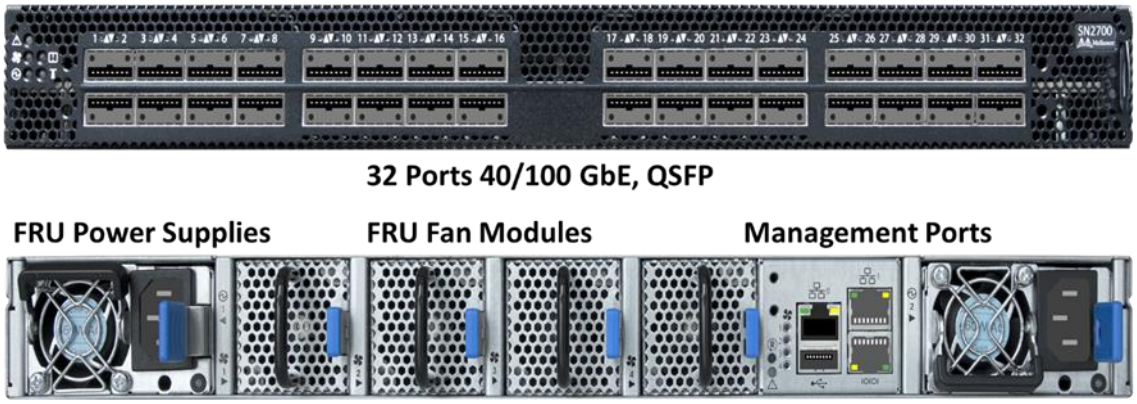
Figure 3) SN2100 faceplate and ports.



4.3 SN2700

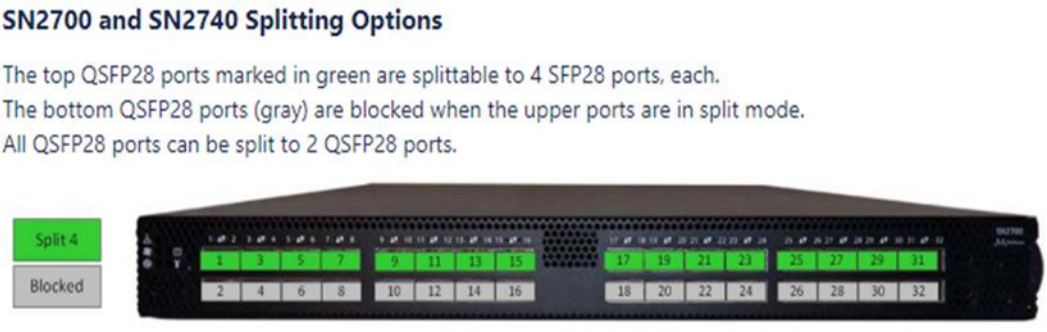
The Mellanox SN2700 (NetApp PN MSN2700-CB2F-NTAP) is a 1RU-high, full-width switch with 32 ports of 10/25/40/100GbE in a QSFP transceiver form factor (Figure 4).

Figure 4) SN2700 front and back.



Note: Note that, while it has 32 ports (double the number of the SN2100), the SN2700 switch can only support 64 connections. When using 4:1 breakout cables for the node connections, half of the QSFP ports are blocked. Therefore, it has the same number of ports for node connections as the SN2100 (Figure 5).

Figure 5) QSFP port splitting on the SN2700 with corresponding port deactivation.

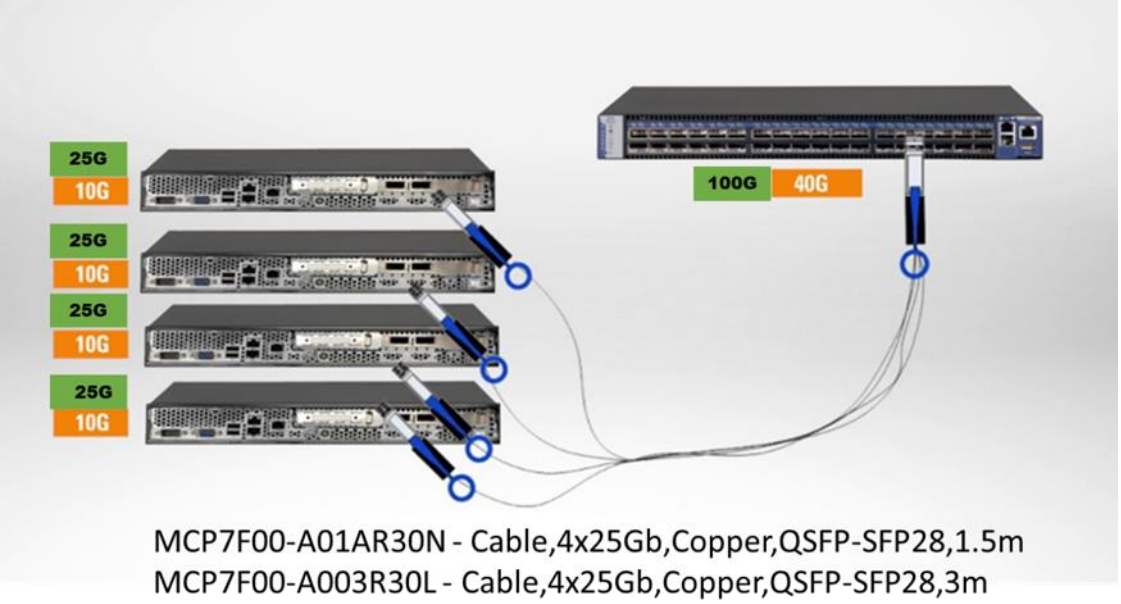


The SN2700 can be used in place of the SN2100 for spine switching, expansion, and where high serviceability is desired, because the power supplies and fan modules are field-replaceable units. If a PSU or fan module fails, it can be easily be replaced without chassis removal or uncabling while the switch is operating.

4.4 Breakout Cable Usage

The QSFP ports on SN2100 and SN2700 switches require 4:1 breakout cables to connect to cluster nodes. Each QSFP breakout cable connects to four cluster nodes as shown below (Figure 6).

Figure 6) QSFP breakout cable for cluster node connections.



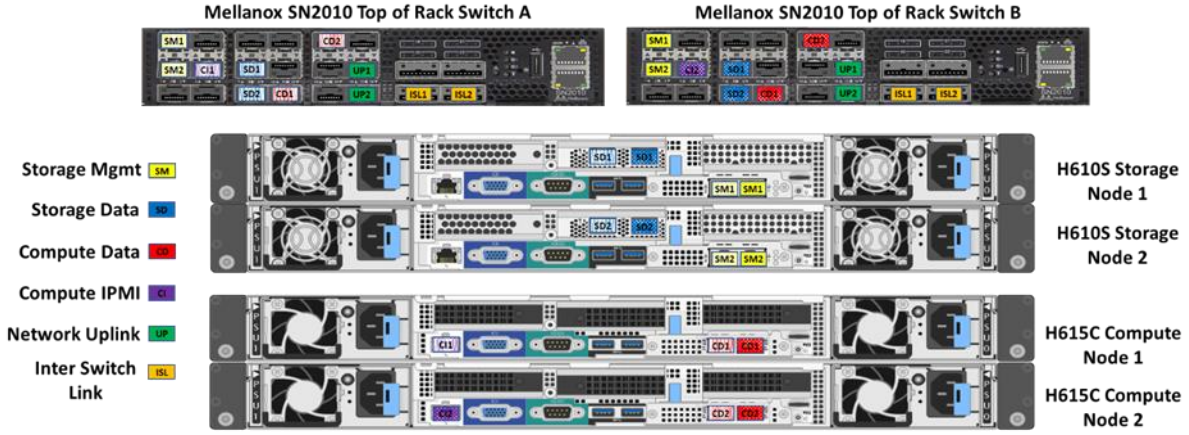
5 Cluster Network Planning, Connection Diagrams and Cabling

The planned number of compute nodes and storage nodes (cluster size) is the primary determining factor in sizing and selecting an appropriate Mellanox Ethernet switch for your HCI deployment. Other determining factors include the management network, out-of-band network, and compute node cable deployment requirements. The following sections provide various NetApp HCI cluster sizes with recommended switches and cabling diagrams.

5.1 Minimum HCI Cluster

The minimum cluster size for NetApp HCI is two storage nodes and two compute nodes. Figure 7 and Figure 8 are network connection diagrams for a minimum-size NetApp HCI system, including all high speed, management, and out of band networks.

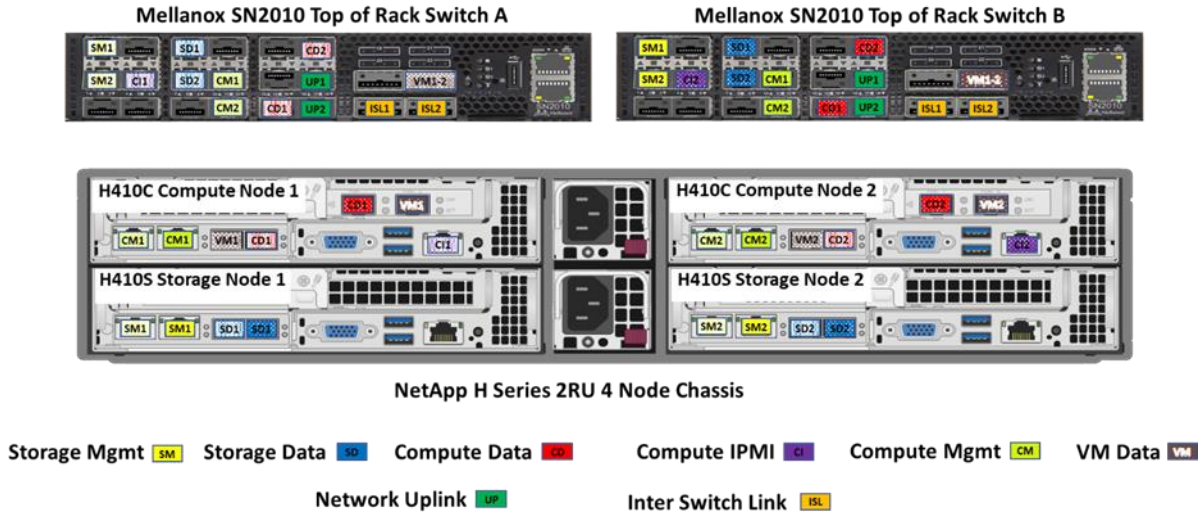
Figure 7) NetApp HCI minimum size two-storage-node / two-compute-node cluster with high-speed data, management, and OOB networks with 1RU H610S storage with H615C compute nodes.



Cabling configuration:

- High-speed node connections: (8) direct-attach copper Twinax or optical, 10/25GbE
- Management: (6) Cat 6 RJ-45 for storage management and compute OOB, 1/10GbE
- Management: (6) 1GbE SFP 1000BASE-T transceiver, adapt SFP to Base-T
- Storage OOB is assumed to be assigned to the management network. Compute management is on the high-speed data network (for example, the two-cable option)
- ISL: (2) QSFP to QSFP copper cable, 100GbE
- Uplink: (4) optical or Twinax cable, 10/25GbE, or optionally (4) QSFP 100GbE

Figure 8) NetApp HCI minimum size two-storage-node / two-compute-node cluster with high-speed data, management and OOB networks with 2RU H410S storage with H410C compute nodes.



Cabling configuration:

- High-speed node connections (12) direct-attach copper Twinax or optical, 10/25GbE
- Management: (10) Cat 6 RJ-45 for storage management, compute management, and compute OOB, 1GbE

- Management; (10) 1GbE SFP 1000BASE-T transceiver, adapt SFP to Base-T
- ISL: (2) QSFP to QSFP copper cable, 100GbE
- Uplink: (4) optical or Twinax cable, 10/25GbE

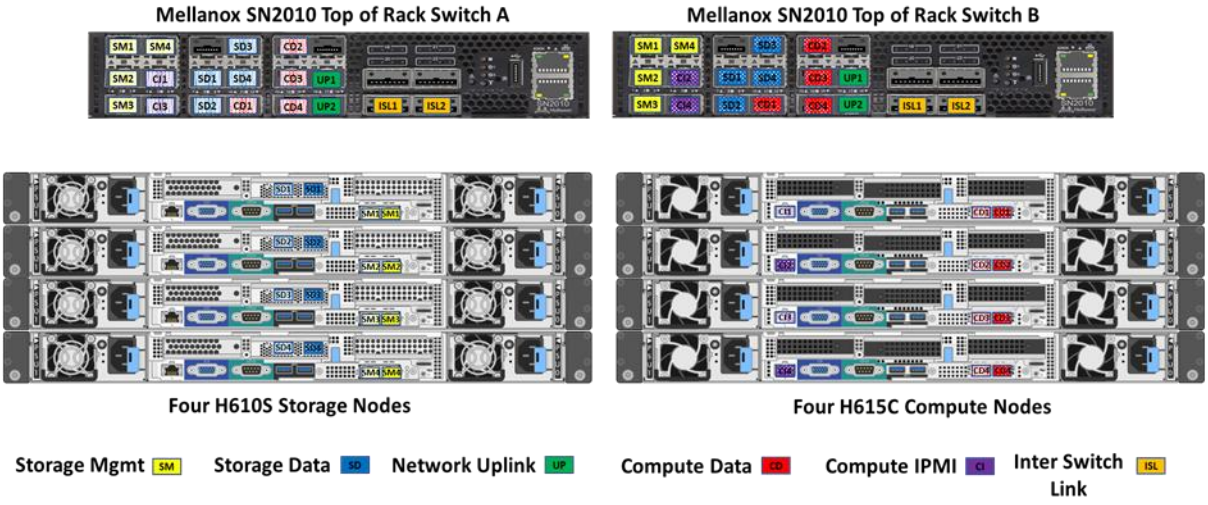
Network notes:

- In the example above, the SN2010 is connecting the high-speed data network as well as the management and out-of-band networks.
- Compute nodes in Figure 7 use the two-cable option, which combines data and management on a single interface logically segregated with VLANs.
- Compute nodes in Figure 8 use the six-cable option, which segregates the data storage, vMotion, and management networks onto separate physical interfaces.
- The IPMI OOB network is not redundant and is only connected on compute nodes.
- We assume that storage IPMI OOB network traffic is redirected to the storage management ports, and therefore they are not included in port usage.
- The network uplinks can be configured on 10/25GbE SFP ports 17 and 18 (as shown), or optionally configured on the 100GbE QSFP ports 19 and 21.
- A Multi-Chassis Link Aggregation Group (MLAG) must be established between the two ToR switches, and configuring the Link Aggregation Control Protocol (LACP) between the storage nodes and switches is a best practice.
- Jumbo frames must be configured on all high-speed interfaces within the HCI cluster network and are recommended for the network uplinks.

5.2 Small HCI Cluster

Small cluster sizes range from four storage and four compute nodes up to 12 storage and 12 compute if the management and OOB networks are not included on the Mellanox switch (Figure 9).

Figure 9) Typical four-storage-node / four-compute-node HCI cluster with data, management, and OOB networks.



Cabling configuration:

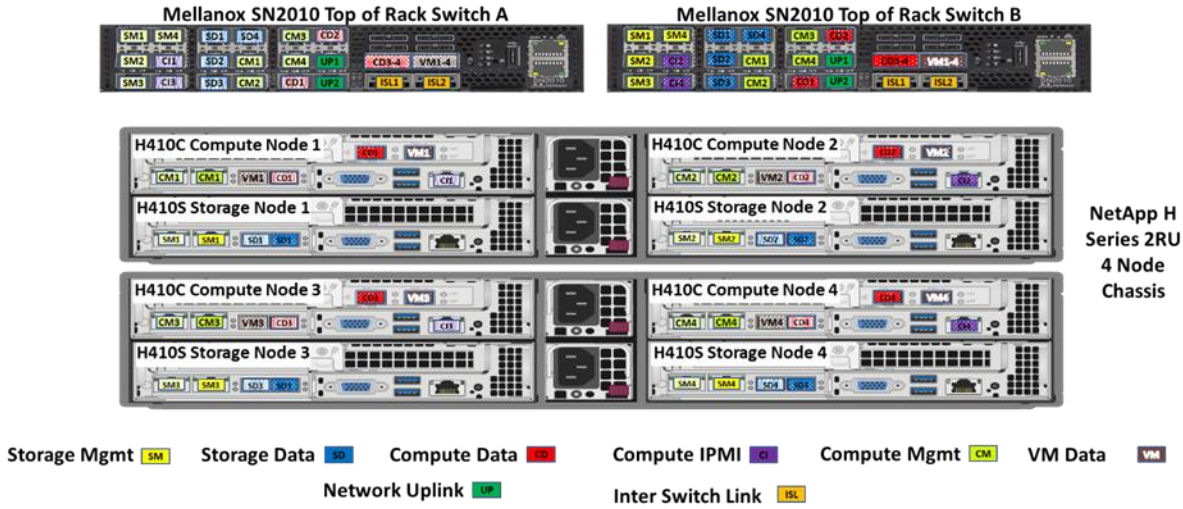
- High-speed node connections: (12) direct-attach copper Twinax or optical, 10/25GbE
- Management: (10) Cat 6 RJ-45 for storage management, compute management, and compute OOB, 1GbE

- Management: (10) 1GbE SFP 1000BASE-T transceiver, adapt SFP to Base-T
- ISL: (2) QSFP to QSFP copper cable, 100GbE
- Uplink: (4) optical or Twinax cable, 10/25GbE

Network notes:

- This HCI system example is shown using 1RU compute and storage nodes. NetApp HCI systems can also be configured in 2RU chassis as shown below in Figure 10 with four slots per chassis for compute or storage resources.
- The compute IPMI OOB network is not redundant, and connections are split between the A and B ToR switches. Optionally, these connections can be connected to a separate, dedicated IPMI network.

Figure 10) Typical four-storage-node / four-compute-node HCI cluster with data, management, and OOB networks, six-cable compute option.



Cabling configuration:

- High-speed node connections: (12) copper Twinax or optical, SFP28 / SFP28, 10/25GbE
- High-speed node connections: (4) copper Twinax, QSFP / SFP28, 4:1 breakout cables
- Management: (20) Cat 6 RJ-45 for storage management, compute management, and compute OOB, 1GbE
- Management: (20) 1GbE SFP 1000BASE-T transceiver, adapt SFP to Base-T
- ISL: (2) QSFP to QSFP copper cable, 100GbE
- Uplink: (4) optical or Twinax cable, 10/25GbE

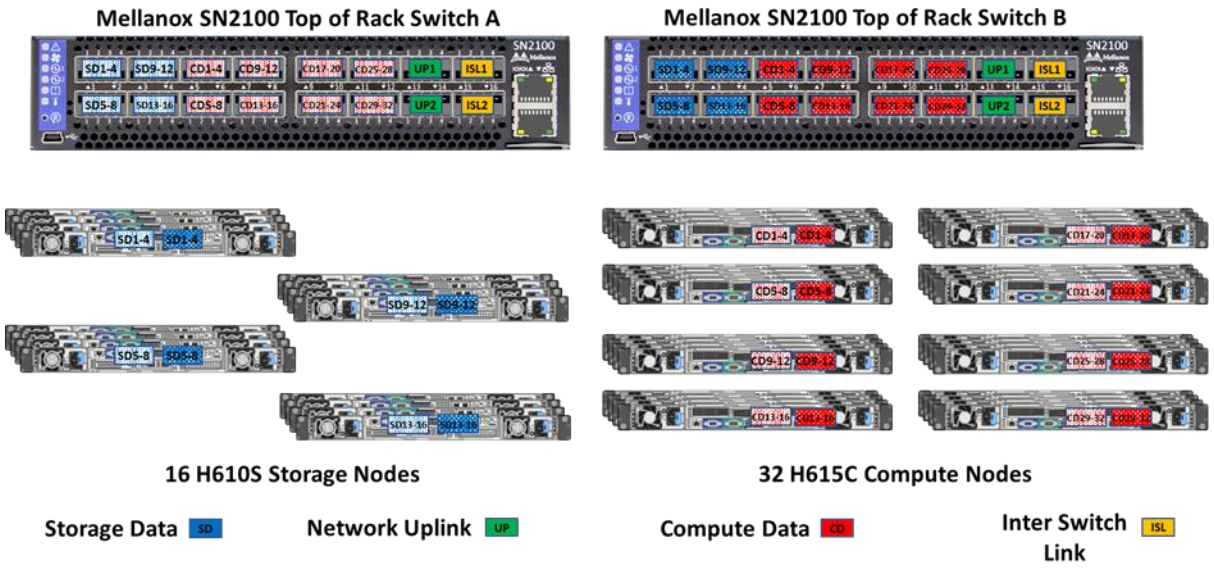
Network notes:

- With 2RU-chassis compute-node models, it is possible to unbundle the management network, the high-speed vMotion network, and the high-speed storage network (this option is shown and is called six-cable compute). When configuring networks with this cable option, compute nodes use twice the number of high-speed data ports (two network switch ports consumed per compute node), and two additional Cat 6 ports per node.

5.3 Mid-Size HCI Clusters

Medium cluster sizes for NetApp HCI start at 12 storage / 12 compute nodes and range up to 16 storage / 32 compute nodes. The SN2100 (or SN2700 if high serviceability is required) is an ideal fit here, because a single pair of ToR switches can support up to 16 storage and 32 compute node clusters (Figure 11).

Figure 11) 16-storage-node / 32-compute-node HCI cluster using one pair of SN2100 ToR switches.



Network Notes:

- Depicted above are the high-speed data network connections. All management and OOB network ports are assumed to be connected to an existing datacenter rack infrastructure.
- With medium and larger cluster sizes, multiple racks might be required for equipment install depending on the rack size and datacenter cooling and power concerns. The location of the nodes in the racks relative to the switches dictate cable length requirements. Other rack node placement concerns such as protection domains are not within the scope of this document.
- Each quad breakout cable supports four node connections. These network diagrams show each of these breakouts connected to either storage or compute nodes (not intermixed). This configuration is used for ease of switch programmability but is not a requirement.
- Uplinks to the data center network can be 10GbE or 25GbE if interfaces are split, or 100GbE natively using QSFP. It is a best practice is to use multiple uplinks configured with link aggregation.

5.4 Large HCI Clusters

Cluster sizes beyond 16 storage and 32 compute nodes require multiple pairs of ToR switches, configured in a full mesh spine-and-leaf, scale-out arrangement. Configured as such, two pairs of ToR switches can support a maximum of 64 compute and 32 storage nodes (Figure 12 and Figure 13).

Figure 12) Large-size HCI cluster with maximum number of compute nodes (64).

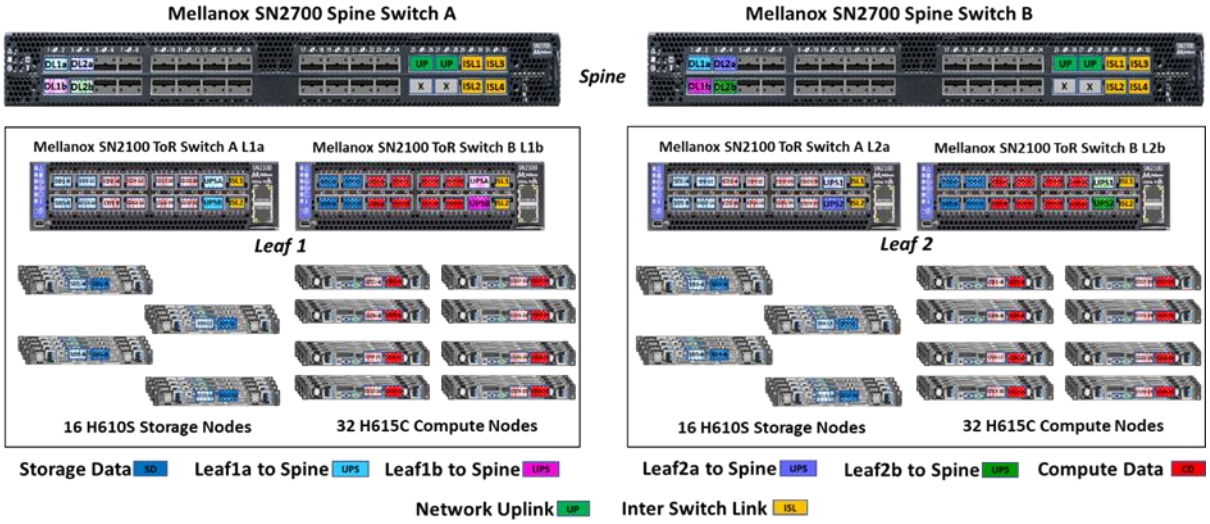
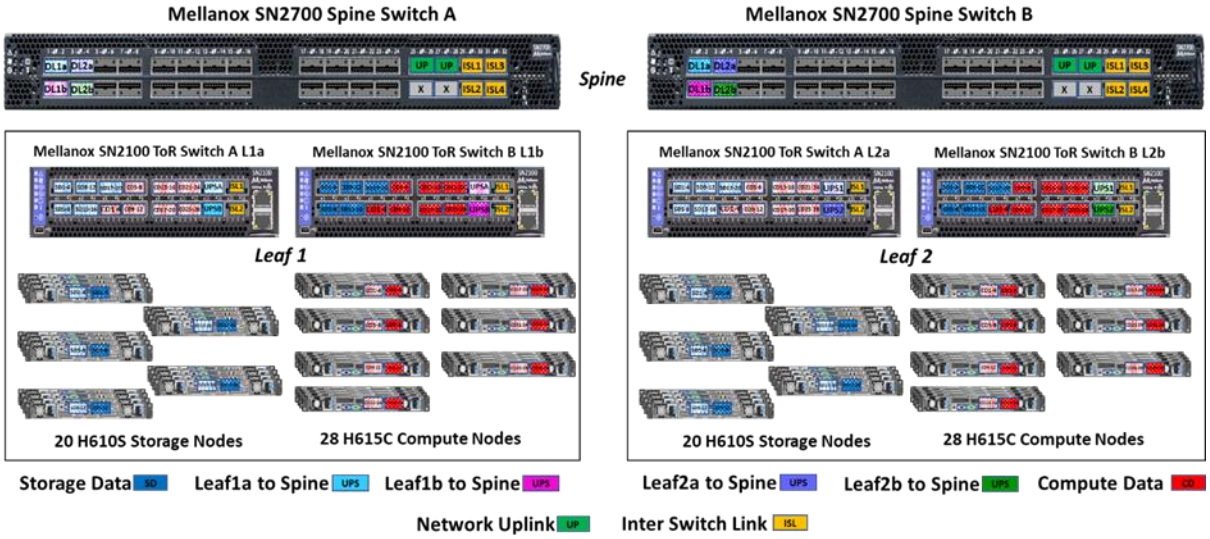


Figure 13) Large-size HCI cluster with maximum number of storage nodes (40).



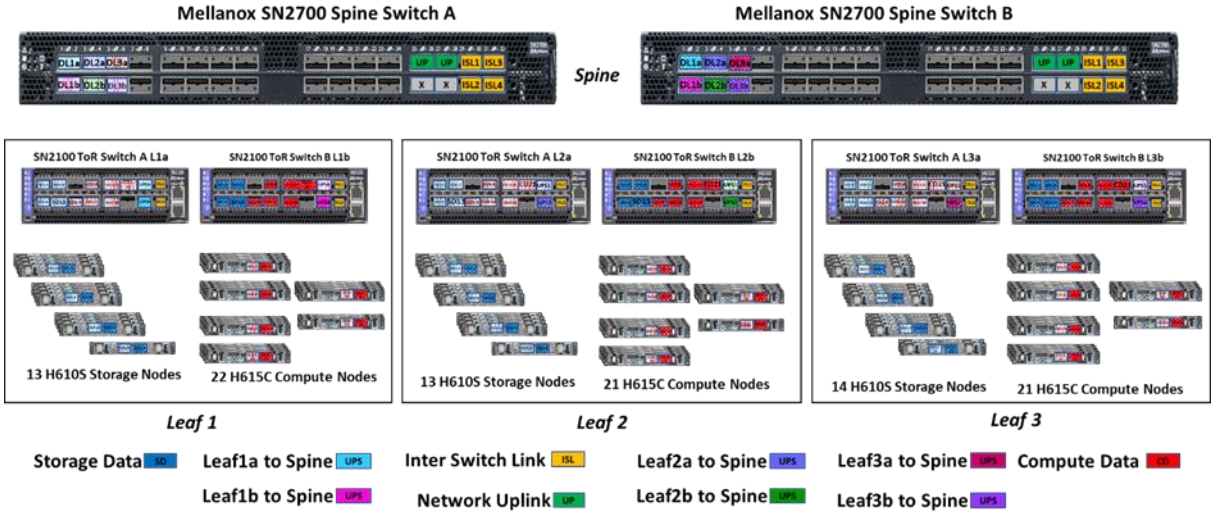
Network Notes:

- As with leaf ToR pairs, spine switch pairs are required to be interconnected with MLAG.
- Multiple network uplinks are shown at 100GbE connectivity. Configured this way, the SN2700 spine does not block the corresponding port beneath. See Section 4.3, Figure 5 for the blocking diagram.
- If your data center cannot connect to 100GbE QSFP uplink ports, you can split the uplink ports and connect at 10GbE or 25GbE. In this case, the ports beneath are blocked as shown in Section 4.3, Figure 5.

5.5 Maximum Size HCI Cluster

A maximum cluster size of 64 compute and 40 storage nodes requires three pairs of ToR leaf switches connected to a spine switch layer (Figure 14).

Figure 14) NetApp HCI maximum cluster size, 64 compute and 40 storage nodes.



Network Notes:

- Additional leaf pairs require a set of connections to each spine switch. Therefore, with a single pair of spine switches, you can support multiple clusters.

5.6 NetApp Orderable Cabling for Mellanox Switches

Table 1) Network cables for node-to-switch and switch-to-switch connections.

Marketing Part No.	Item Description	Use Case
X6584-R6	1Gb cables, 1m	Cat 6 RJ-45 Ethernet cable
X6585-R6	1Gb cables, 3m	Cat 6 RJ-45 Ethernet cable
MCP2M00-A001E30N	Cable, 25Gb, copper, SFP28-SFP28, 1m	Node to SN2010 SFP28 switch port
MCP2M00-A003E30L	Cable, 25Gb, copper, SFP28-SFP28, 3m	Node to SN2010 SFP28 switch port
MCP1600-C00AE30N	Cable, 100Gb, copper, QSFP-QSFP, 0.5m	Switch-to-switch ISL QSFP ports
MCP1600-C001E30N	Cable, 100Gb, copper, QSFP-QSFP, 1m	Switch-to-switch ISL QSFP ports
MCP1600-C003E30L	Cable, 100Gb, copper, QSFP-QSFP, 3m	Switch-to-switch ISL QSFP ports
MCP7F00-A01AR30N	Cable, 4x25Gb, copper, QSFP-SFP28, 1.5m	4:1 breakout, switch QSFP port to node
MCP7F00-A003R30L	Cable, 4x25Gb, copper, QSFP-SFP28, 3m	4:1 breakout, switch QSFP port to node
MC3208411-T	SFP optical, 1GbE Base-T, RJ45	SN2010 1GbE management connections
MFM1T02A-SR	SFP+ optical, LC, 10GbE, 300m, SR	10GbE short-reach optical transceiver
MFM1T02A-LR	SFP+ optical, LC, 10GbE, 10km, LR	10GbE long-reach optical transceiver

Marketing Part No.	Item Description	Use Case
MMA2P00-AS-SP	SFP28 optical, LC, 25GbE, 100m, SR	25GbE short-reach optical transceiver
MMA2L20-AR	SFP28 optical, LC, 25GbE, 10km, LR	25GbE long-reach optical transceiver
MMA1B00-C100D	QSFP28 optical, MPO, 100GbE, 100m, SR4	Optical transceiver for switch-to-switch ISL QSFP ports
MMA1L10-CR	QSFP28 optical, LC, 100GbE, 10km, LR4	Optical transceiver for long-range ISL QSFP ports

6 General Switch Configuration, Small Cluster Network Example

6.1 General Protocol Requirements

Adherence to the cabling diagrams above is essential to availability, performance, and extensibility of NetApp HCI. A fully redundant and fault-tolerant network is required for proper operation. Switch protocols such as LACP and MLAG are required to make sure that redundant links are readily available and performance is optimized.

Note: The following Mellanox switch CLI code examples are not necessarily tied to any specific network configuration or drawing above. Use these as examples on how to program the switch for your specific network connection diagram. Consult NetApp Professional Services for additional specifics.

```
#Example for Mellanox SN2000 Series Switches Protocol Services
#Configuration steps are in order they should be entered

#Enable Services
   protocol mlag
   lacp
   lldp
   ip routing vrf default
   dcb priority-flow-control enable force
   what-just-happened all enable
#Allow these commands to fully complete before continuing
```

Note: These commands must be executed in order first and allowed to complete before continuing.

6.2 VLAN Requirements

NetApp HCI requires logically separate networks (VLANs) for leaf MLAG domains, spine MLAG domains, management, iSCSI storage, virtual machine (VM) migration (vMotion), and VM data.

```
#Configure VLANs
   vlan 4000 name “Leaf 1 IPL VLAN”
   exit
   vlan 100 name “HCI Management”
   exit
   vlan 101 name "HCI iSCSI Storage"
   exit
   vlan 102 name "HCI vMotion"
   exit
   vlan 103 name “HCI Guest VM1 Network”
   exit
```

6.3 Configure Switch -to-Switch Inter-Peer Link and Multi-Chassis Link Aggregation

MLAG configuration is required between pairs of ToR leaf or spine switches. IP addresses for Inter-Peer Links (IPLs) are private between the MLAG pair and can be the same as other switch pairs.

```
#Configure IPL
   interface port-channel 100
   description IPL
   exit
   interface ethernet 1/20 channel-group 100 mode active
   interface ethernet 1/20 description ISL to partner ToR Switch
   interface ethernet 1/22 channel-group 100 mode active
   interface ethernet 1/22 description ISL to partner ToR Switch
   interface port-channel 100 ipl 1
   interface port-channel 100 dcb priority-flow-control mode on force
   interface vlan 4000
   exit
   #IPL Ip should *not* be in the management network
   interface vlan 4000 ip address 10.255.255.1 255.255.255.252
   interface vlan 4000 ipl 1 peer-address 10.255.255.2

#Configure MLAG VIP
#Both MLAG name and ip must be unique
#MLAG IP *should* be in the management network
   mlag-vip MLAG-VIP-DOM ip 10.xxx.xxx.174 /24 force
   #mlag system mac can be any unicast MAC
   mlag system-mac 00:00:5E:00:AA:01
   no mlag shutdown
```

Note: The previous example assumes the IPLs between the two ToR switches are on ports 20 and 22. Also note that the corresponding partner ToR switch interchanges the IP addresses in the peer-address command (highlighted).

6.4 Configure Uplink

The network uplinks should be configured to match the speed of the corresponding network data center port capability. The following example assumes a 10GbE connection speed, and network uplinks on ports 17 and 18. It is also preferred (but not required) to configure jumbo frames on the uplinks.

```
#Configure Uplink
   interface mlag-port-channel 101
   description Uplink to Datacenter
   exit
   interface ethernet 1/17 description Network Uplink A
   interface ethernet 1/17 speed 10000 force
   interface ethernet 1/18 description Network Uplink B
   interface ethernet 1/18 speed 10000 force
   interface mlag-port-channel 101 mtu 9216 force
   interface ethernet 1/17 mlag-channel-group 101 mode active
   interface ethernet 1/18 mlag-channel-group 101 mode active
   interface mlag-port-channel 101 switchport mode trunk
   interface mlag-port-channel 101 switchport trunk allowed-vlan all
   interface mlag-port-channel 101 no shutdown
```

6.5 Configure Management Ports

This section is included for smaller installations that include management ports in the switch setup. All node ports should have a spanning tree configured to enter a forwarding state immediately (Rapid Spanning Tree Protocol). These commands can be used as an example on how to configure the management network on an existing datacenter 1GbE rack infrastructure.

```
#Configure Management Ports (Storage Node Management)
   interface ethernet 1/1-1/4 spanning-tree bpdufilter enable
   interface ethernet 1/1-1/4 spanning-tree port type edge
```

```
   interface ethernet 1/1-1/4 spanning-tree bpduguard enable
   interface ethernet 1/1-1/4 speed 1000 force
   interface ethernet 1/1-1/4 switchport mode access
   interface ethernet 1/1 description HCI-Storage Node 1 Management Port
   interface ethernet 1/2 description HCI-Storage Node 2 Management Port
   interface ethernet 1/3 description HCI-Storage Node 3 Management Port
   interface ethernet 1/4 description HCI-Storage Node 4 Management Port
   interface ethernet 1/1 switchport access vlan 100
   interface ethernet 1/2 switchport access vlan 100
   interface ethernet 1/3 switchport access vlan 100
   interface ethernet 1/4 switchport access vlan 100
#Configure Out-of-Band Ports (Storage Node OOB for ToR "A" switch, Compute Node OOB for "B")
   interface ethernet 1/13-1/16 spanning-tree bpdufilter enable
   interface ethernet 1/13-1/16 spanning-tree port type edge
   interface ethernet 1/13-1/16 spanning-tree bpduguard enable
   interface ethernet 1/13-1/16 speed 1000 force
   interface ethernet 1/13-1/16 switchport mode access
   interface ethernet 1/13 description HCI-Storage/Compute Node 1 IPMI OOB Management Port
   interface ethernet 1/14 description HCI-Storage/Compute Node 2 IPMI OOB Management Port
   interface ethernet 1/15 description HCI-Storage/Compute Node 3 IPMI OOB Management Port
   interface ethernet 1/16 description HCI-Storage/Compute Node 4 IPMI OOB Management Port
   interface ethernet 1/13 switchport access vlan 100
   interface ethernet 1/14 switchport access vlan 100
   interface ethernet 1/15 switchport access vlan 100
   interface ethernet 1/16 switchport access vlan 100
```

Note: When using the compute node two-cable option, the management network for compute nodes is combined with vMotion and guest VM data traffic on the high-speed data network and logically separated with VLANs. Also, you can optionally assign OOB network access to the management network on the storage nodes, which does not require separate connections.

6.6 Configure Storage Ports, iSCSI Network

```
#Configure iSCSI Ports (storage nodes)
   interface ethernet 1/5-1/8 spanning-tree port type edge
   interface ethernet 1/5-1/8 spanning-tree bpduguard enable
   interface ethernet 1/5-1/8 lacp port-priority 10
   interface ethernet 1/5-1/8 lacp rate fast
   interface ethernet 1/5 description STORAGE Node 1
   interface ethernet 1/6 description STORAGE Node 2
   interface ethernet 1/7 description STORAGE Node 3
   interface ethernet 1/8 description STORAGE Node 4
   interface mlag-port-channel 201-204
   exit
   interface ethernet 1/5 mlag-channel-group 201 mode active
   interface ethernet 1/6 mlag-channel-group 202 mode active
   interface ethernet 1/7 mlag-channel-group 203 mode active
   interface ethernet 1/8 mlag-channel-group 204 mode active
   interface mlag-port-channel 201-204 mtu 9216 force
   interface mlag-port-channel 201-204 lacp-individual enable force
   interface mlag-port-channel 201-204 no shutdown
   interface mlag-port-channel 201-204 switchport mode hybrid
   interface mlag-port-channel 201 switchport hybrid allowed-vlan all
   interface mlag-port-channel 202 switchport hybrid allowed-vlan all
   interface mlag-port-channel 203 switchport hybrid allowed-vlan all
   interface mlag-port-channel 204 switchport hybrid allowed-vlan all
   interface mlag-port-channel 201 switchport access vlan 100
   interface mlag-port-channel 202 switchport access vlan 100
   interface mlag-port-channel 203 switchport access vlan 100
   interface mlag-port-channel 204 switchport access vlan 100
```

Note: Note that LACP and data port nodes, along with multi-chassis link aggregation, are set up on the storage nodes.

6.7 Configure VM Data, vMotion Network

```
#Configure data ports (compute nodes vMotion and VM Data Network)
   interface ethernet 1/9-1/12 spanning-tree port type edge
   interface ethernet 1/9-1/12 spanning-tree bpduguard enable
```

```
   interface ethernet 1/9-1/12 switchport mode hybrid
   interface ethernet 1/9-1/12 mtu 9216 force
   interface ethernet 1/9 description COMPUTE Node 1
   interface ethernet 1/10 description COMPUTE Node 2
   interface ethernet 1/11 description COMPUTE Node 3
   interface ethernet 1/12 description COMPUTE Node 4
   interface ethernet 1/9 switchport hybrid allowed-vlan all
   interface ethernet 1/10 switchport hybrid allowed-vlan all
   interface ethernet 1/11 switchport hybrid allowed-vlan all
   interface ethernet 1/12 switchport hybrid allowed-vlan all
   interface ethernet 1/9 switchport access vlan 100
   interface ethernet 1/10 switchport access vlan 100
   interface ethernet 1/11 switchport access vlan 100
   interface ethernet 1/12 switchport access vlan 100
```

7 Mid-Size Cluster Network Configuration

There are some differences when configuring the network for mid-size HCI clusters as shown in Section 5.3 Figure 11 when compared to the small clusters shown in Figure 10:

- With SN2100 and SN2700 switches, all ports are type QSFP. Therefore, any network connections to compute or storage nodes must be made through a QSFP to 4X SFP breakout cable as outlined in Section 4.4, Figure 6.
- Because the SN2100 and SN2700 switches have the QSFP port type, the management network connections and configuration are not included here.

```
# Configure data Ports (shortened for brevity)
# Example of port splitting
   interface ethernet 1/5-1/12 module-type qsfp-split-4 force
   interface ethernet 1/5/1-1/12/4 spanning-tree port type edge
   interface ethernet 1/5/1-1/12/4 spanning-tree bpduguard enable
   interface ethernet 1/5/1-1/12/4 switchport mode hybrid
   interface ethernet 1/5/1-1/12/4 mtu 9216 force
   interface ethernet 1/5/1 description COMPUTE Node 1
   interface ethernet 1/5/2 description COMPUTE Node 2
   interface ethernet 1/5/3 description COMPUTE Node 3
   interface ethernet 1/5/4 description COMPUTE Node 4
# … skip ahead to Node 32
   interface ethernet 1/12/4 description Compute Node 32
   interface ethernet 1/5/1 switchport hybrid allowed-vlan all
# … skip ahead to Node 32
   interface ethernet 1/12/4 switchport hybrid allowed-vlan all
   interface ethernet 1/5/1 switchport access vlan 100
# … skip ahead to Node 32
   interface ethernet 1/12/4 switchport access vlan 100
```

8 Large-Size and Max-Size Cluster Network Configuration

Large-size and max-size clusters with SN2100 and SN2700 require a spine-leaf network architecture as shown in Figures Figure 12Figure 13Figure 14. This network architecture includes multiple pairs of ToR switches (leafs) connected in a full mesh to the spine switch layer, which uplinks to the data center core.

- Each switch in the leaf layer is redundantly connected to its ToR partner switch using MLAG. There is no change from small-size or mid-size cluster configurations.
- The switches in the spine layer are also redundantly connected together with MLAG protocol.
- Each switch in the leaf layer is connected to every spine switch (full mesh).

```
#Configure IPL between ToR Leaf Switches
   interface port-channel 100
   description IPL
   exit
   interface ethernet 1/15 channel-group 100 mode active
   interface ethernet 1/15 description ISL to partner ToR Switch
```

```
   interface ethernet 1/16 channel-group 100 mode active
   interface ethernet 1/16 description ISL to partner ToR Switch
   interface port-channel 100 ipl 1
   interface port-channel 100 dcb priority-flow-control mode on force
   interface vlan 4000
   exit
#IPL IP should *not* be in the management network
   interface vlan 4000 ip address 10.255.255.1 255.255.255.252
   interface vlan 4000 ipl 1 peer-address 10.255.255.2

#Configure MLAG VIP for Leaf
#Both MLAG name and IP must be unique
#MLAG IP *should* be in the management network
   mlag-vip MLAG-VIP-DOM ip 10.xxx.xxx.174 /24 force
#mlag system mac can be any unicast MAC
   mlag system-mac 00:00:5E:00:AA:01
   no mlag shutdown

#Configure ISL between Leaf and Spine Switches
   interface port-channel 101
   description ISL Leaf Switch 1 to Spine Switch 1
   exit
   interface ethernet 1/13 channel-group 100 mode active
   interface ethernet 1/15 description ISL to partner ToR Switch
   interface ethernet 1/16 channel-group 100 mode active
   interface ethernet 1/16 description ISL to partner ToR Switch
   interface port-channel 100 ipl 1
   interface port-channel 100 dcb priority-flow-control mode on force
   interface vlan 4000
   exit
#IPL IP should *not* be in the management network
   interface vlan 4000 ip address 10.255.255.1 255.255.255.252
   interface vlan 4000 ipl 1 peer-address 10.255.255.2

#Configure IPL between Spine Switches, use example from Leaf Switches
```

9 Sample Network Configuration for a 16 Storage / 32 Compute Node NetApp HCI Cluster

9.1 Initial Switch Setup and Onyx OS Update

The following programming for the Mellanox SN2100 switch CLI is used to setup a mid-size network, as shown in Section 5.3, Figure 11. See the following notes to set up this Mellanox switch configuration:

- You must modify both the A side and the B side ToR leaf and spine switches.
- For instructions on how to connect and configure the switches for the first time, you must establish an account at Mellanox if you do not already have one.
- Navigate to the Download Center and type the switch model number in the search bar. Click the Documentation button and download the latest version of the Onyx Users Manual. Follow the instructions in Section 2.1, Configuring the Switch for the First Time.
- Set up an Ethernet connection between the switch and a local network machine with a standard RJ-45 connector.

Important: Do **not** modify both switches at the same time. Wait for switch A to come up completely before you modify switch B.

- NetApp highly recommends checking the OS version on the switches and upgrading to the latest system software if necessary. NetApp also highly recommends running the same version of Mellanox Onyx on both switches in a ToR leaf or spine pair.
- Download the latest Onyx switch OS from the Download Center.

- Follow the instructions in Section 4.4 of the Onyx Users Manual to upgrade switch OS. Do this for both switches.

After you have input general configuration information; updated the switch to the latest Onyx OS version; and named, saved, and enabled the configuration, the switchports are ready to be configured according to the NetApp HCI cabling diagram.

9.2 Switch Programming with the CLI

The following console output is a sample of the configuration CLI commands to run and deploy on the Mellanox SN2100 switch for a two-cable compute in this example. NetApp HCI cabling configuration consisting of 32 compute nodes and 16 storage nodes. This is the maximum number of nodes for a single ToR pair of Mellanox SN2100 switches.

Note: Shown in this example are hybrid ports with all VLANs accessible on all data and storage ports, with access ports for management and normal trunk ports for the uplinks.

Note: In this example, switchports 13 and 14 on both switch A and switch B are configured as 100GbE uplinks. If your data center cannot connect to 100GbE QSFP, you can optionally split these ports and connect at 10GbE or 25GbE to the data center.

```
#Example for Mellanox SN2100 Series Switches Protocol Services
#Configuration steps are in order they should be entered

#Enable Services
   protocol mlag
   lacp
   lldp
   ip routing vrf default
   dcb priority-flow-control enable force
   what-just-happened all enable

#Allow these commands to fully complete before continuing

#Configure VLANs
   vlan 4000 name “Leaf 1 IPL VLAN”
   exit
   vlan 100 name “HCI Management”
   exit
   vlan 101 name "HCI iSCSI Storage"
   exit
   vlan 102 name "HCI vMotion"
   exit
   vlan 103 name “HCI Guest VM Network”
   exit

#Configure IPL
   interface port-channel 100
   description IPL
   exit
   interface ethernet 1/15 channel-group 100 mode active
   interface ethernet 1/15 description ISL to partner ToR Switch
   interface ethernet 1/16 channel-group 100 mode active
   interface ethernet 1/16 description ISL to partner ToR Switch
   interface port-channel 100 ipl 1
   interface port-channel 100 dcb priority-flow-control mode on force
   interface vlan 4000
   exit
   #IPL Ip should *not* be in the management network
   interface vlan 4000 ip address 10.255.255.1 255.255.255.252
   interface vlan 4000 ip1 1 peer-address 10.255.255.2

#Configure MLAG VIP
#Both MLAG name and ip must be unique
#MLAG IP *should* be in the management network
   mlag-vip MLAG-VIP-DOM ip 10.xxx.xxx.174 /24 force
#mlag system mac can be any unicast MAC
   mlag system-mac 00:00:5E:00:AA:01
```

```
   no mlag shutdown

#Configure Uplink
   interface mlag-port-channel 101
   description Uplink to Datacenter
   exit
   interface ethernet 1/13 description Network Uplink 1
   interface ethernet 1/14 description Network Uplink 2
   interface mlag-port-channel 101 mtu 9216 force
   interface ethernet 1/17 mlag-channel-group 101 mode active
   interface ethernet 1/18 mlag-channel-group 101 mode active
   interface mlag-port-channel 101 switchport mode trunk
   interface mlag-port-channel 101 switchport trunk allowed-vlan all
   interface mlag-port-channel 101 no shutdown

#Split Ports (Storage nodes)
   interface ethernet 1/1-1/4 module-type qsfp-split-4 force

##Configure iSCSI Ports (Storage nodes)
   interface ethernet 1/1/1-1/4/4 spanning-tree port type edge
   interface ethernet 1/1/1-1/4/4 spanning-tree bpduguard enable
   interface ethernet 1/1/1-1/4/4 lacp port-priority 10
   interface ethernet 1/1/1-1/4/4 lacp rate fast
   interface ethernet 1/1/1 description STORAGE Node 1
   interface ethernet 1/1/2 description STORAGE Node 2
   interface ethernet 1/1/3 description STORAGE Node 3
   interface ethernet 1/1/4 description STORAGE Node 4
   interface ethernet 1/2/1 description STORAGE Node 5
   interface ethernet 1/2/2 description STORAGE Node 6
   interface ethernet 1/2/3 description STORAGE Node 7
   interface ethernet 1/2/4 description STORAGE Node 8
   interface ethernet 1/3/1 description STORAGE Node 9
   interface ethernet 1/3/2 description STORAGE Node 10
   interface ethernet 1/3/3 description STORAGE Node 11
   interface ethernet 1/3/4 description STORAGE Node 12
   interface ethernet 1/4/1 description STORAGE Node 13
   interface ethernet 1/4/2 description STORAGE Node 14
   interface ethernet 1/4/3 description STORAGE Node 15
   interface ethernet 1/4/4 description STORAGE Node 16
   interface mlag-port-channel 201-216
   exit
   interface ethernet 1/1/1 mlag-channel-group 201 mode active
   interface ethernet 1/1/2 mlag-channel-group 202 mode active
   interface ethernet 1/1/3 mlag-channel-group 203 mode active
   interface ethernet 1/1/4 mlag-channel-group 204 mode active
   interface ethernet 1/2/1 mlag-channel-group 205 mode active
   interface ethernet 1/2/2 mlag-channel-group 206 mode active
   interface ethernet 1/2/3 mlag-channel-group 207 mode active
   interface ethernet 1/2/4 mlag-channel-group 208 mode active
   interface ethernet 1/3/1 mlag-channel-group 209 mode active
   interface ethernet 1/3/2 mlag-channel-group 210 mode active
   interface ethernet 1/3/3 mlag-channel-group 211 mode active
   interface ethernet 1/3/4 mlag-channel-group 212 mode active
   interface ethernet 1/4/1 mlag-channel-group 213 mode active
   interface ethernet 1/4/2 mlag-channel-group 214 mode active
   interface ethernet 1/4/3 mlag-channel-group 215 mode active
   interface ethernet 1/4/4 mlag-channel-group 216 mode active
   interface mlag-port-channel 201-216 mtu 9216 force
   interface mlag-port-channel 201-216 lacp-individual enable force
   interface mlag-port-channel 201-216 no shutdown
   interface mlag-port-channel 201-216 switchport mode hybrid
   interface mlag-port-channel 201 switchport hybrid allowed-vlan all
   interface mlag-port-channel 202 switchport hybrid allowed-vlan all
   interface mlag-port-channel 203 switchport hybrid allowed-vlan all
   interface mlag-port-channel 204 switchport hybrid allowed-vlan all
   interface mlag-port-channel 205 switchport hybrid allowed-vlan all
   interface mlag-port-channel 206 switchport hybrid allowed-vlan all
   interface mlag-port-channel 207 switchport hybrid allowed-vlan all
   interface mlag-port-channel 208 switchport hybrid allowed-vlan all
   interface mlag-port-channel 209 switchport hybrid allowed-vlan all
   interface mlag-port-channel 210 switchport hybrid allowed-vlan all
```

```
   interface mlag-port-channel 211 switchport hybrid allowed-vlan all
   interface mlag-port-channel 212 switchport hybrid allowed-vlan all
   interface mlag-port-channel 213 switchport hybrid allowed-vlan all
   interface mlag-port-channel 214 switchport hybrid allowed-vlan all
   interface mlag-port-channel 215 switchport hybrid allowed-vlan all
   interface mlag-port-channel 216 switchport hybrid allowed-vlan all
   interface mlag-port-channel 201 switchport access vlan 100
   interface mlag-port-channel 202 switchport access vlan 100
   interface mlag-port-channel 203 switchport access vlan 100
   interface mlag-port-channel 204 switchport access vlan 100
   interface mlag-port-channel 205 switchport access vlan 100
   interface mlag-port-channel 206 switchport access vlan 100
   interface mlag-port-channel 207 switchport access vlan 100
   interface mlag-port-channel 208 switchport access vlan 100
   interface mlag-port-channel 209 switchport access vlan 100
   interface mlag-port-channel 210 switchport access vlan 100
   interface mlag-port-channel 211 switchport access vlan 100
   interface mlag-port-channel 212 switchport access vlan 100
   interface mlag-port-channel 213 switchport access vlan 100
   interface mlag-port-channel 214 switchport access vlan 100
   interface mlag-port-channel 215 switchport access vlan 100
   interface mlag-port-channel 216 switchport access vlan 100

# Split Ports (Compute nodes)
   interface ethernet 1/5-1/12 module-type qsfp-split-4 force
   interface ethernet 1/5/1-1/12/4 spanning-tree port type edge
   interface ethernet 1/5/1-1/12/4 spanning-tree bpduguard enable
   interface ethernet 1/5/1-1/12/4 switchport mode hybrid
   interface ethernet 1/5/1-1/12/4 mtu 9216 force
   interface ethernet 1/5/1 description COMPUTE Node 1
   interface ethernet 1/5/2 description COMPUTE Node 2
   interface ethernet 1/5/3 description COMPUTE Node 3
   interface ethernet 1/5/4 description COMPUTE Node 4
   interface ethernet 1/6/1 description COMPUTE Node 5
   interface ethernet 1/6/2 description COMPUTE Node 6
   interface ethernet 1/6/3 description COMPUTE Node 7
   interface ethernet 1/6/4 description COMPUTE Node 8
   interface ethernet 1/7/1 description COMPUTE Node 9
   interface ethernet 1/7/2 description COMPUTE Node 10
   interface ethernet 1/7/3 description COMPUTE Node 11
   interface ethernet 1/7/4 description COMPUTE Node 12
   interface ethernet 1/8/1 description COMPUTE Node 13
   interface ethernet 1/8/2 description COMPUTE Node 14
   interface ethernet 1/8/3 description COMPUTE Node 15
   interface ethernet 1/8/4 description COMPUTE Node 16
   interface ethernet 1/9/1 description COMPUTE Node 17
   interface ethernet 1/9/2 description COMPUTE Node 18
   interface ethernet 1/9/3 description COMPUTE Node 19
   interface ethernet 1/9/4 description COMPUTE Node 20
   interface ethernet 1/10/1 description COMPUTE Node 21
   interface ethernet 1/10/2 description COMPUTE Node 22
   interface ethernet 1/10/3 description COMPUTE Node 23
   interface ethernet 1/10/4 description COMPUTE Node 24
   interface ethernet 1/11/1 description COMPUTE Node 25
   interface ethernet 1/11/2 description COMPUTE Node 26
   interface ethernet 1/11/3 description COMPUTE Node 27
   interface ethernet 1/11/4 description COMPUTE Node 28
   interface ethernet 1/12/1 description COMPUTE Node 29
   interface ethernet 1/12/2 description COMPUTE Node 30
   interface ethernet 1/12/3 description COMPUTE Node 31
   interface ethernet 1/12/4 description COMPUTE Node 32
   interface ethernet 1/5/1 switchport hybrid allowed-vlan all
   interface ethernet 1/5/2 switchport hybrid allowed-vlan all
   interface ethernet 1/5/3 switchport hybrid allowed-vlan all
   interface ethernet 1/5/4 switchport hybrid allowed-vlan all
   interface ethernet 1/6/1 switchport hybrid allowed-vlan all
   interface ethernet 1/6/2 switchport hybrid allowed-vlan all
   interface ethernet 1/6/3 switchport hybrid allowed-vlan all
   interface ethernet 1/6/4 switchport hybrid allowed-vlan all
   interface ethernet 1/7/1 switchport hybrid allowed-vlan all
   interface ethernet 1/7/2 switchport hybrid allowed-vlan all
```

```
interface ethernet 1/7/3 switchport hybrid allowed-vlan all
interface ethernet 1/7/4 switchport hybrid allowed-vlan all
interface ethernet 1/8/1 switchport hybrid allowed-vlan all
interface ethernet 1/8/2 switchport hybrid allowed-vlan all
interface ethernet 1/8/3 switchport hybrid allowed-vlan all
interface ethernet 1/8/4 switchport hybrid allowed-vlan all
interface ethernet 1/9/1 switchport hybrid allowed-vlan all
interface ethernet 1/9/2 switchport hybrid allowed-vlan all
interface ethernet 1/9/3 switchport hybrid allowed-vlan all
interface ethernet 1/9/4 switchport hybrid allowed-vlan all
interface ethernet 1/10/1 switchport hybrid allowed-vlan all
interface ethernet 1/10/2 switchport hybrid allowed-vlan all
interface ethernet 1/10/3 switchport hybrid allowed-vlan all
interface ethernet 1/10/4 switchport hybrid allowed-vlan all
interface ethernet 1/11/1 switchport hybrid allowed-vlan all
interface ethernet 1/11/2 switchport hybrid allowed-vlan all
interface ethernet 1/11/3 switchport hybrid allowed-vlan all
interface ethernet 1/11/4 switchport hybrid allowed-vlan all
interface ethernet 1/12/1 switchport hybrid allowed-vlan all
interface ethernet 1/12/2 switchport hybrid allowed-vlan all
interface ethernet 1/12/3 switchport hybrid allowed-vlan all
interface ethernet 1/12/4 switchport hybrid allowed-vlan all
interface ethernet 1/5/1 switchport access vlan 100
interface ethernet 1/5/2 switchport access vlan 100
interface ethernet 1/5/3 switchport access vlan 100
interface ethernet 1/5/4 switchport access vlan 100
interface ethernet 1/6/1 switchport access vlan 100
interface ethernet 1/6/2 switchport access vlan 100
interface ethernet 1/6/3 switchport access vlan 100
interface ethernet 1/6/4 switchport access vlan 100
interface ethernet 1/7/1 switchport access vlan 100
interface ethernet 1/7/2 switchport access vlan 100
interface ethernet 1/7/3 switchport access vlan 100
interface ethernet 1/7/4 switchport access vlan 100
interface ethernet 1/8/1 switchport access vlan 100
interface ethernet 1/8/2 switchport access vlan 100
interface ethernet 1/8/3 switchport access vlan 100
interface ethernet 1/8/4 switchport access vlan 100
interface ethernet 1/9/1 switchport access vlan 100
interface ethernet 1/9/2 switchport access vlan 100
interface ethernet 1/9/3 switchport access vlan 100
interface ethernet 1/9/4 switchport access vlan 100
interface ethernet 1/10/1 switchport access vlan 100
interface ethernet 1/10/2 switchport access vlan 100
interface ethernet 1/10/3 switchport access vlan 100
interface ethernet 1/10/4 switchport access vlan 100
interface ethernet 1/11/1 switchport access vlan 100
interface ethernet 1/11/2 switchport access vlan 100
interface ethernet 1/11/3 switchport access vlan 100
interface ethernet 1/11/4 switchport access vlan 100
interface ethernet 1/12/1 switchport access vlan 100
interface ethernet 1/12/2 switchport access vlan 100
interface ethernet 1/12/3 switchport access vlan 100
interface ethernet 1/12/4 switchport access vlan 100
```

Where to Find Additional Information

This document provides general and overview information about setting up NetApp HCI with Mellanox SN2010 Ethernet switches. It also provides the VLAN information that you need before you run the NetApp Deployment Engine (NDE) for a completed system setup.

For more detailed information about how to set up the network for NetApp HCI before you run NDE and to learn more about the information that is described in this document, see the following:

- NetApp HCI Networking Quick Planning Guide
  https://www.netapp.com/us/media/tr-4820.pdf

- Setting Up and Using NetApp HCI 1.7P1
  https://docs.netapp.com/hci/topic/com.netapp.nav.hsu-17P1/home.html?cp=2
- NetApp HCI Documentation
  http://docs.netapp.com/hci/index.jsp
- NetApp HCI Resources
  http://docs.netapp.com/hci/topic/com.netapp.ndc.hci-home/GUID-F8FE50BE-9B4F-45E4-B884-1FA7EC8FD4AA.html

Acknowledgements

The author would like to thank Travis Stone and Stephen Carl for contributing to this technical report.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

Trademark Information