BCM957608

Cable Solutions Guide

Application Note

Broadcom reserves the right to make changes without further notice to any products or data herein to improve reliability, function, or design. Information furnished by Broadcom is believed to be accurate and reliable. However, Broadcom does not assume any liability arising out of the application or use of this information, nor the application or use of any product or circuit described herein, neither does it convey any license under its patent rights nor the rights of others.

Table of Contents

1 Introduction

This document provides insight into typical deployment scenarios for AI/ML and HPC applications, with a focus on Network Interface Card (NIC) and cable solutions.

NIC and cable selections play a key role in designing efficient AI and HPC systems. There are multiple cable options available today including:

- Direct Attached Copper Cables (DAC)
- Active Electrical Cables (AEC)
- Active Optical Cables (AOC)
- Optical transceivers combined with standalone optical cables

Each of these options offers different performances and maximum cable lengths with associated differences in power, cost, and availability.

The following sections provide an overview of common AI cluster configurations and how these configurations impact the requirements for cable selection. An introduction to available connectors and cables is provided with recommendations based on the target system configuration.

The document centers on Broadcom's BCM957608 NIC with high-level features shown in Table 1.

Table 1: High-Level Features

Performance	■ 400Gbps ■ 250Mpps
PCIe	■ Gen5
8x 112G SerDes	■ Linear Pluggable Optics (LPO) and 4 meter DAC
Ports	■ 1, 2, or 4 ports

2 Deployment Scenarios

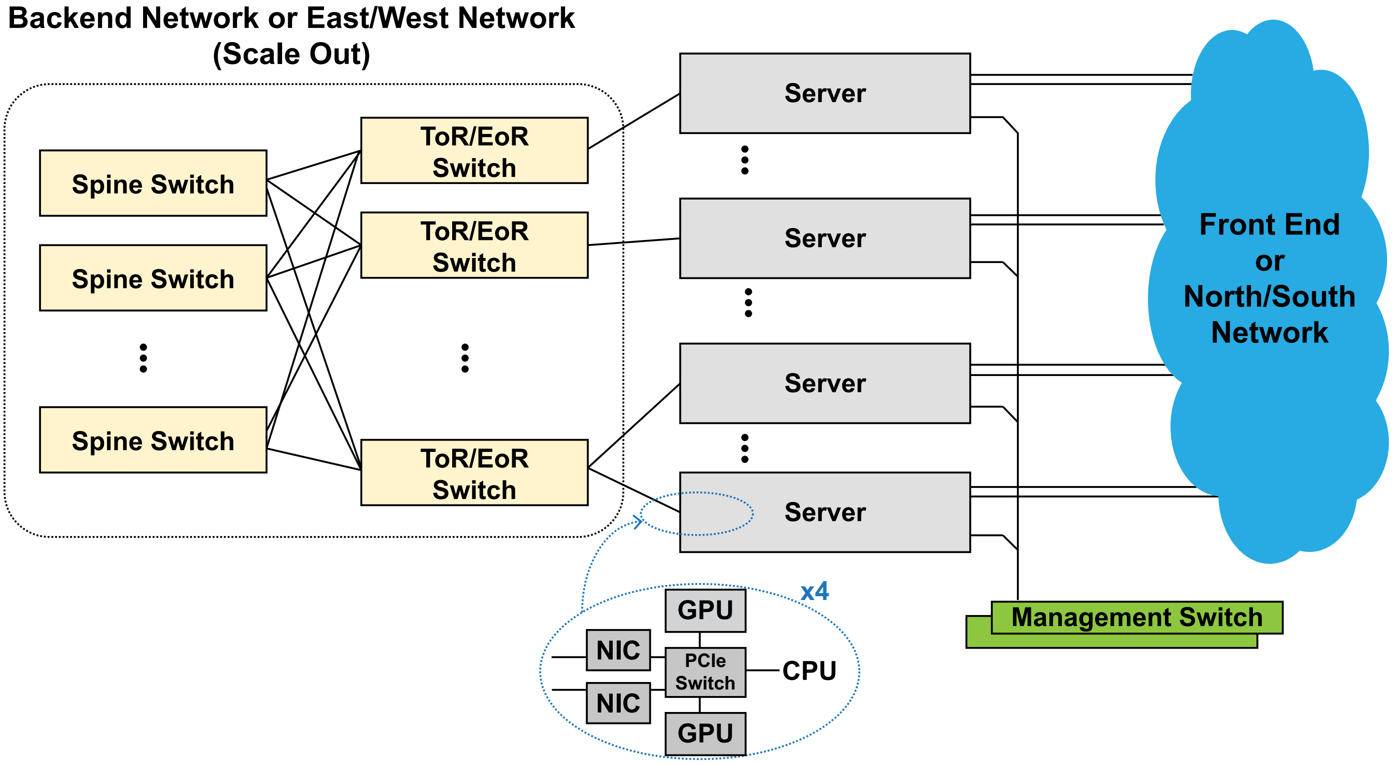
AI clusters scale from relatively low GPU count in the 16 – 64 range up to multiple thousands. The cluster configurations can be split into two categories:

- Compact systems requiring a single Top of Rack (ToR) switch
  These systems can often be optimized to fit in one or two racks.
- Scalable systems using two or three layers of network switching
  These systems can be built by replicating a base configuration and connecting the associated ToR into the larger network.

Cluster scale and rack-level modularity impact connector and connectivity choices. This document explores rack configurations and cabling from the NIC to ToR. Typically, connectivity between switches uses optics to support relatively longer distances between ToR and leaf and leaf and spine. This is not always the case as switches may be grouped together physically, allowing shorter cables between switches and longer cables between the NIC and switch.

The configurations shown in Figure 1 assume one Ethernet NIC per GPU with additional NICs for storage and management. The connectivity described focuses on the high-bandwidth backend network connecting the GPUs. Storage uses a separate network which may use the same or lower bandwidth connections. Management connectivity is provided with an independent, lower bandwidth network which may be redundant.

Figure 1: AI Cluster Overview



Servers used in AI clusters are often designed with eight GPUs, allowing efficient scale-up connectivity within the server enclosure. This document refers to each server as a node, so a single node has eight GPUs and eight backend NICs, plus storage and management NICs. The power consumed by this server varies. Considering a range of 10kW – 15kW, the example configurations use 12.5kW per server.

Modern high-performance Ethernet switches provide 128 to 256 high-speed Ethernet ports. Some deployments choose to configure the switch using Ethernet ports where each port is ½ the highest speed port. For example, a switch with 64 x 800G ports may be configured with 128 x 400G ports, or a combination of mixed port speeds.

The following sections discuss compact AI clusters which use a single Ethernet switch (also known as Top-of-Rack or ToR) for the backend network and scalable AI clusters which may use two or three tiers of switching. Compact AI clusters use all ports on the single switch to connect to servers. Scalable AI clusters use some ports to the servers and some ports to the next tier of switches. These examples are not oversubscribed, dedicating half of the ports to servers and half to the next tier of switches.

2.1 Compact AI Clusters

An overriding factor in designing an AI cluster is determining the maximum power supported by a single rack. This defines the maximum number of servers and switches that may be installed in the rack. Often power and cooling requirements limit rack capacity to well below the available physical area.

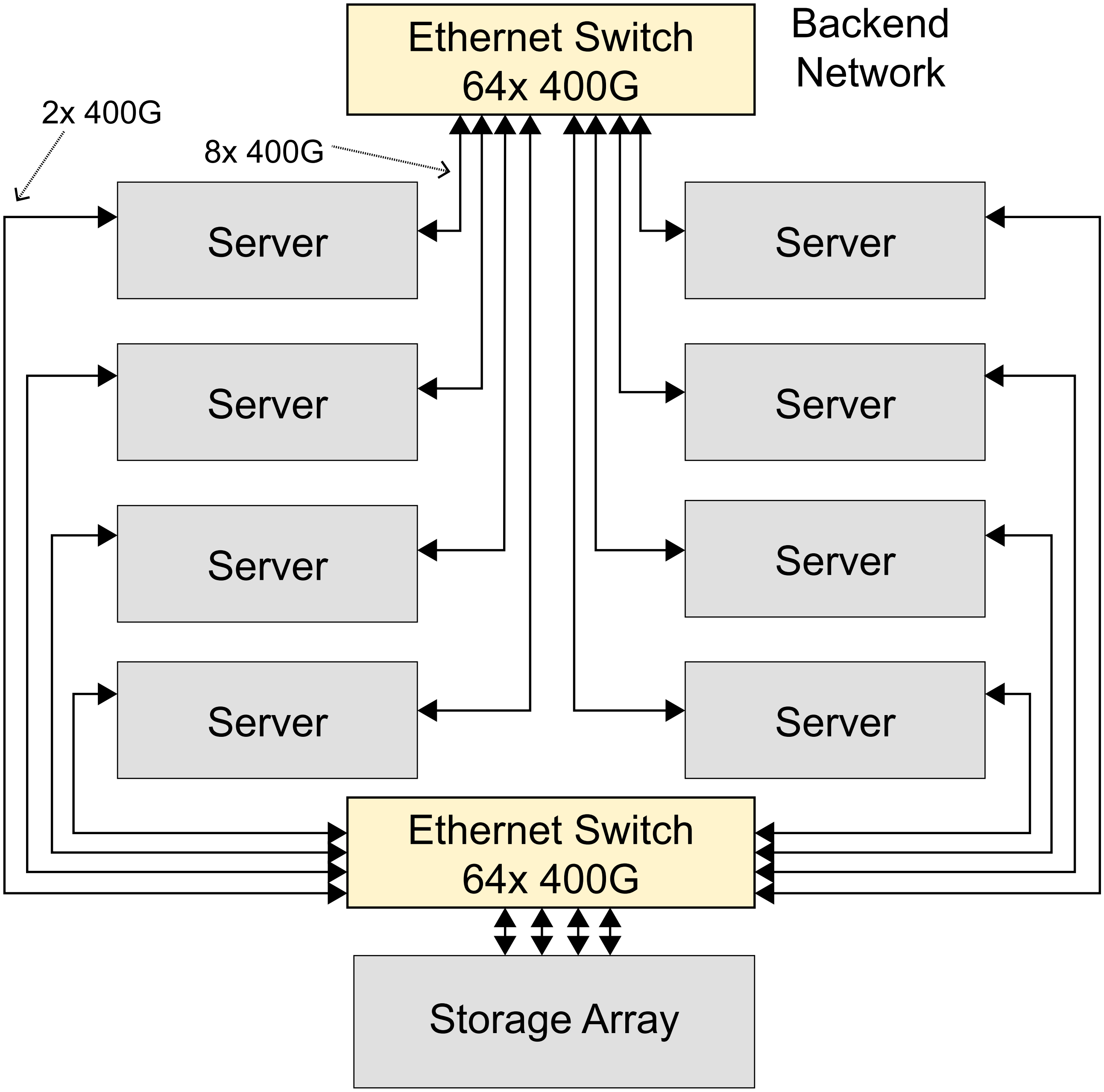
The total number of servers in the cluster and the maximum number of servers per rack determines the number of racks required. These decisions drive the distance between the NIC and switch and must be considered when choosing the connectivity technology.

In addition, the connectivity choice impacts the rack power. Active cables may contribute significantly to rack power.

The following examples use 40kW, 60kW, and 100kW available power and cooling per rack. This allows two, four, or six servers within a single rack, with four servers per rack being a common design point.

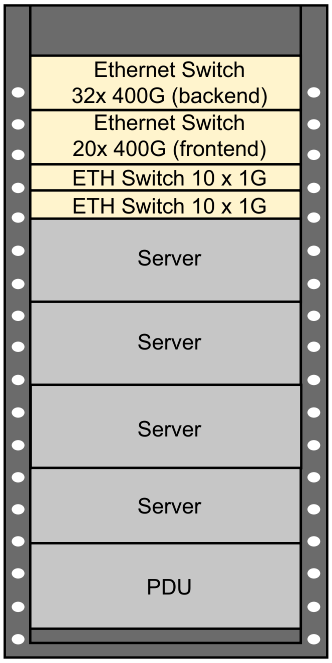
A compact cluster using a single 64-port switch for the backend network supports up to eight servers for a total of 64 GPUs, as shown in Figure 2. A second switch may be used to provide connectivity to a storage array. Storage may use the same Ethernet port speed as the backend, or may be half speed. The management switch is not shown.

Figure 2: Compact AI Cluster Connectivity – 64 Port ToR



The following figures show example system configurations based on cluster size and power per rack.

Figure 3: Rack View for 32 GPU Cluster, 60kW Rack Power



4 Node Cluster (32 GPUs)

60kW Rack Power

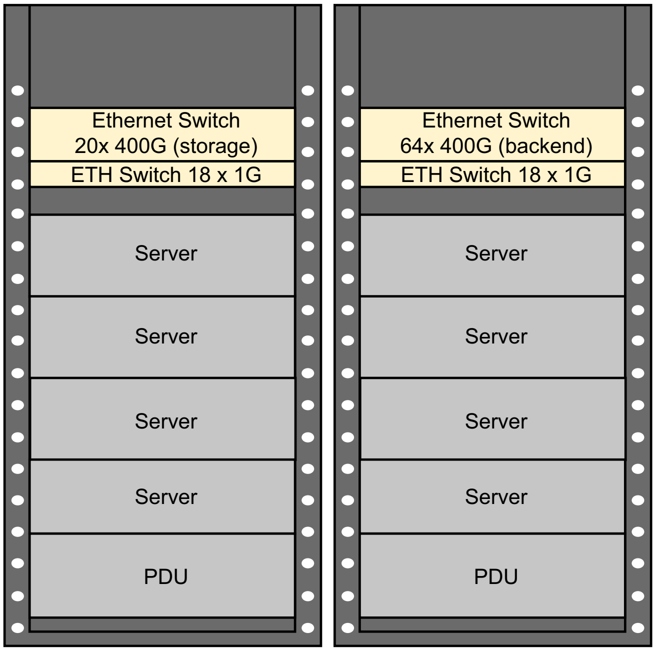
- 50kW Server Power (4 servers)
- 3-5kW Switch Power (4 switches)
- 3-5kW Power Supplies

Connectivity

- Backend
  - Each server has 8x 400G
- Frontend/Storage
  - Each server has 2x 400G
- Management (redundant)
  - Each server has 2x 1G
  - Each management switch has two ports to central management network

Intra-rack connection length: < 2 meters
Number of high-speed, intra-rack connections: 32 NIC to Switch + 8 Storage = 40

Figure 4: Rack View for 64 GPU Cluster, 60kW Rack Power



8 Node Cluster (64 GPUs)

60kW Rack Power

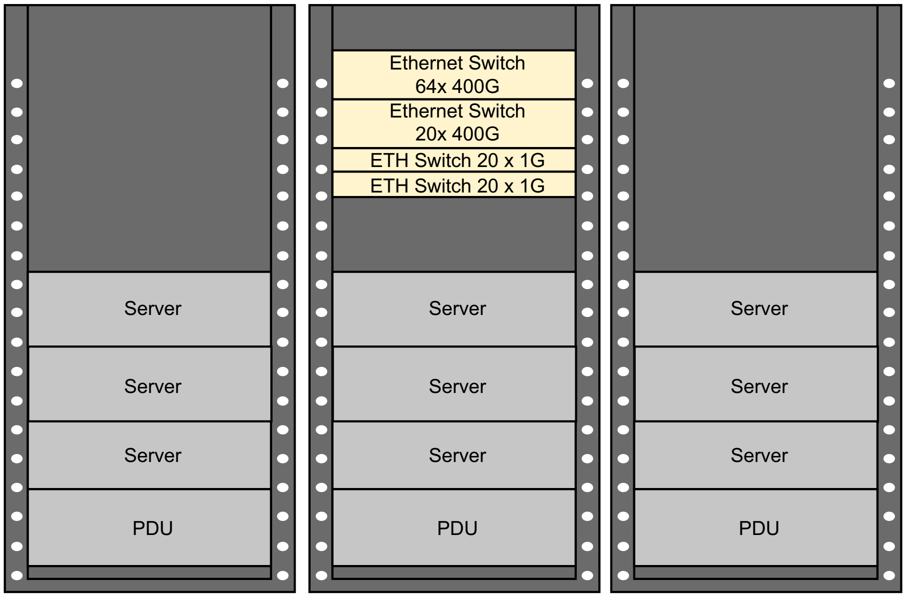
- 50kW Server Power (4 servers)
- 3-5kW Switch Power (2 switches)
- 3-5kW Power Supplies

Connectivity

- Backend
  - Each server has 8x 400G
- Frontend/Storage
  - Each server has 2x 400G
- Management (redundant)
  - Each server has 2x 1G
  - Each management switch has two ports to central management network

Intra-rack connection length: < 4 meters
Number of high-speed, intra-rack connections: 64 NIC to Switch + 16 Storage = 80

Figure 5: Rack View for 64 GPU Cluster, 40kW Rack Power



8 Node Cluster (64 GPUs)

40kW Rack Power
- 37.5kW for 3 servers

Connectivity
- Backend
  - Each server has 8x 400G
- Frontend/Storage
  - Each server has 2x 400G
  - Storage array has 4x 400G
- Management (redundant)
  - Each server has 2x 1G
  - 2 ports to DC management hub

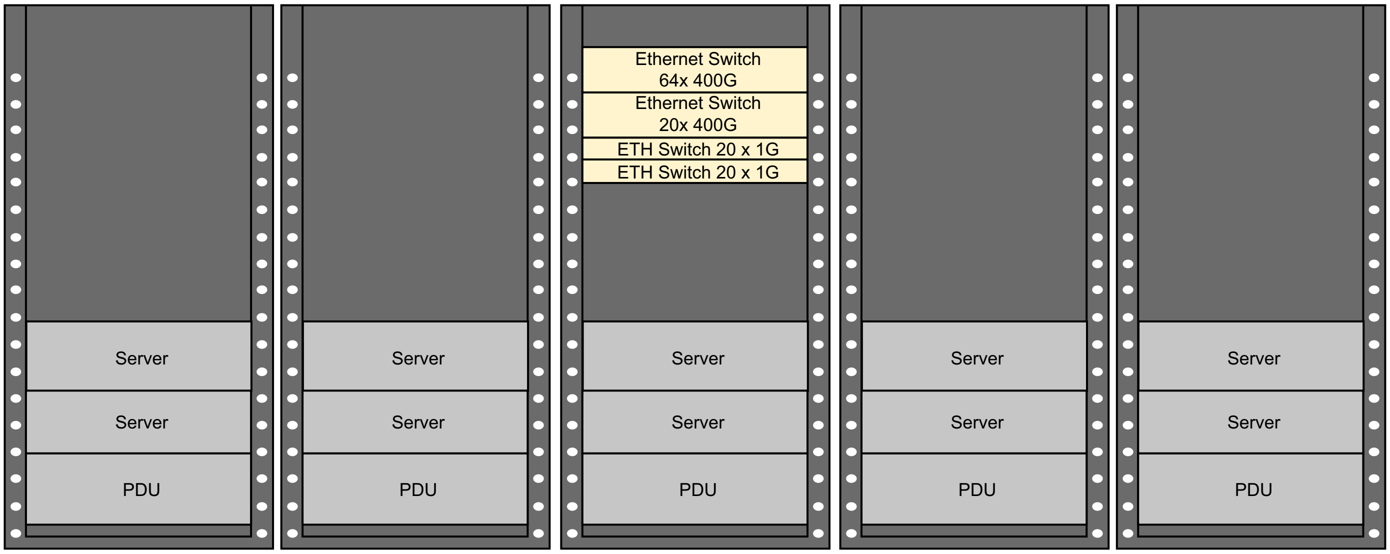
NIC - Switch connection length: < 4 meters
Number of high-speed, intra-rack connections: 64 NIC to Switch + 16 Storage = 80

Figure 6: Rack View for 64 GPU Cluster, 30kW Rack Power

8 Node Cluster (64 GPUs)

8 Node Cluster (64 GPUs)
- 25kW for 2 servers

Connectivity
- Cable lengths are longer



NIC - Switch connection length: 7 - 8 meters from outer rack to switches
Number of high-speed, intra-rack connections: 64 NIC to Switch + 16 Storage = 80

An alternative system configuration using redundancy is shown in Figure 7. The system uses two Ethernet switches in the backend network with traffic load balanced across both switches. This tolerates link failures without requiring the job to halt.

Figure 7: Compact AI Cluster Connectivity – 64 port ToR with Redundancy

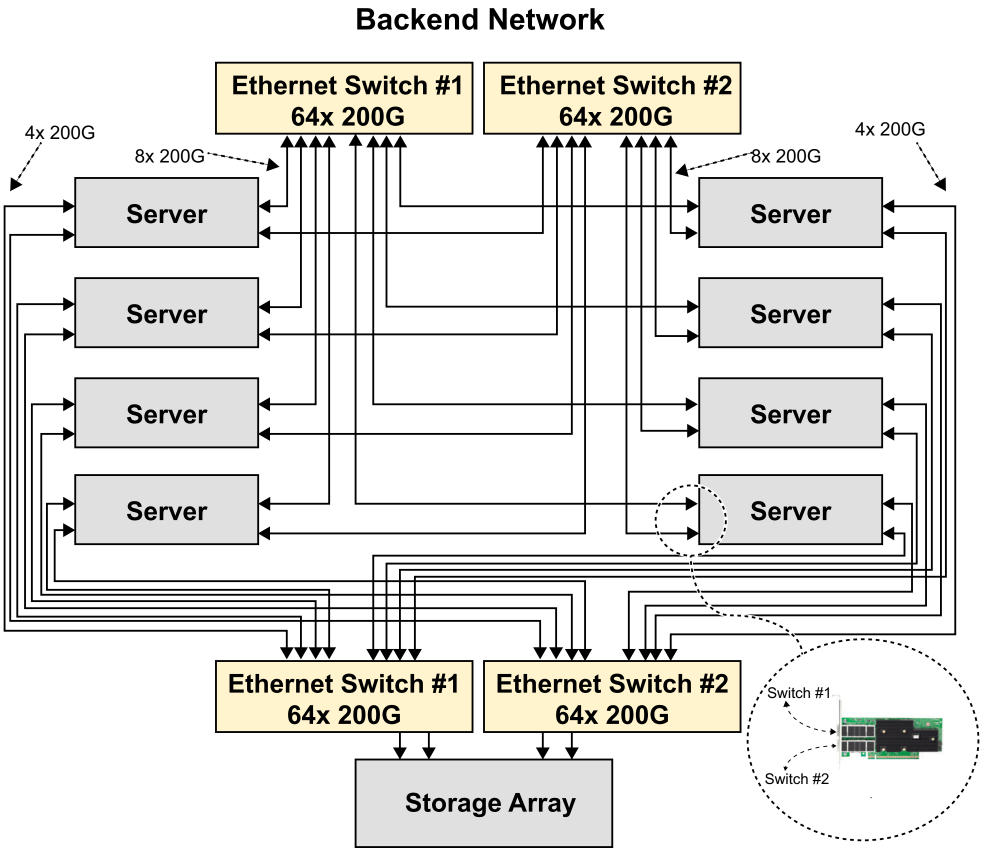
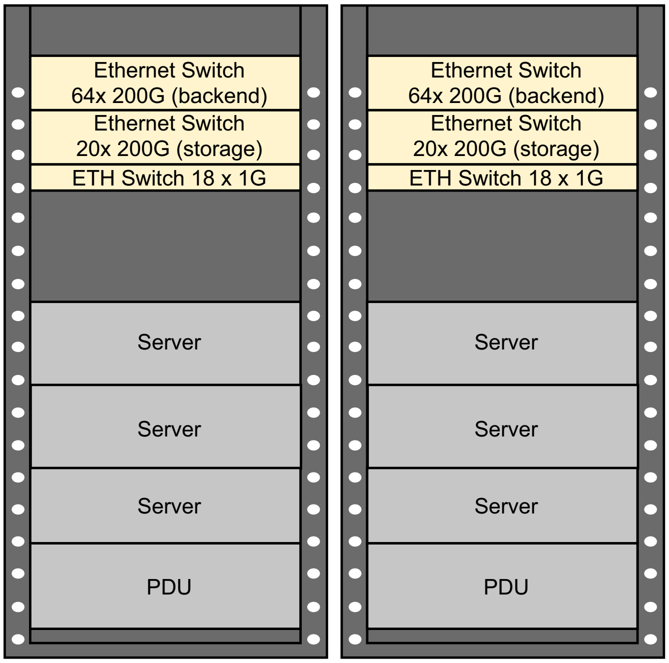


Figure 8: Rack View for 64 GPU Cluster, 60kW Rack Power – Redundant


Ethernet Switch 64x 200G (backend)
Ethernet Switch 20x 200G (storage)
ETH Switch 18 x 1G
Server
Server
Server
PDU
Ethernet Switch 64x 200G (backend)
Ethernet Switch 20x 200G (storage)
ETH Switch 18 x 1G
Server
Server
Server
PDU

8 Node Cluster (64 GPUs)
Redundant Network

60kW Rack Power

- 4 servers per rack

Connectivity

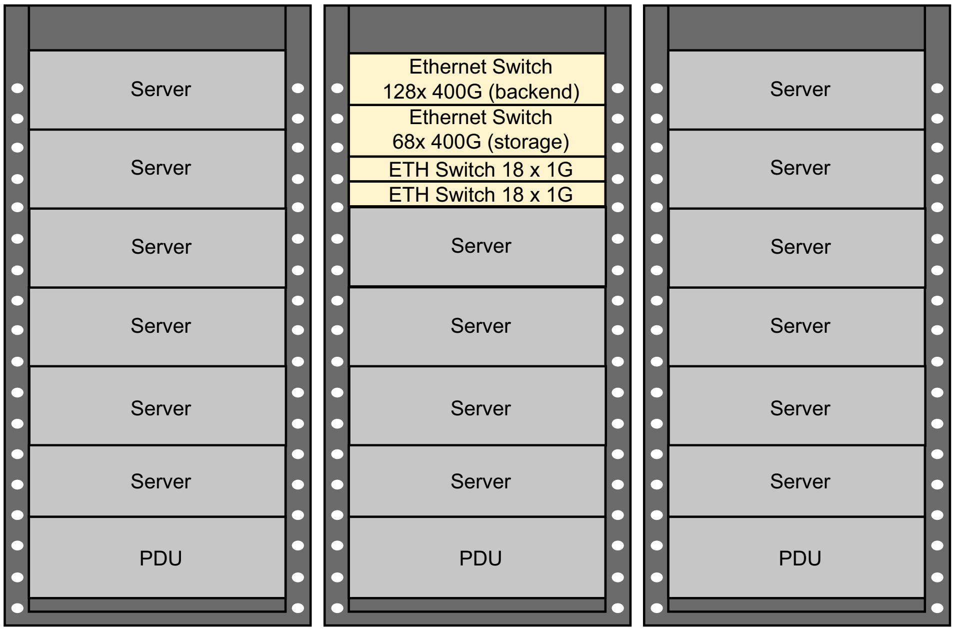
- Backend
  - Each server has 16x 200G
  - Each GPU is connected to both backend switches
- Storage
  - Each server has 4x 200G
  - 2x 200G to each storage switch
- Management (redundant)
  - Each server has 2x 1G
  - 1x 1G to each management switch

NIC - Switch connection length: < 4 meters
Number of high-speed, intra-rack connections: 128 NIC to Switch + 32 Storage = 160

A larger radix switch with 128 ports enables a larger compact cluster with up to 16 servers. When 100kW per rack is available, this can be built with as few as 3 racks as shown in Figure 9. Racks with 60kW power and cooling can be built with 4 or 5 racks.

Figure 9: Rack View for 128 GPU Cluster, 100kW Rack Power – 128 port ToR Switch

160 Node Cluster (128 GPUs) Compact Cluster



100kW Rack Power

- 75kW for 6 servers
- 60kW for 4 servers
- 3-5kW for switches
- 3-5kW for power

Connectivity

- Backend
  - Each server has 8x 400G
- Storage
  - Each server has 2x 400G
- Management (redundant)
  - 2x 1G - one to each management switch

NIC - Switch connection length: < 4 meters
Number of high-speed, intra-rack connections: 128 NIC to Switch + 32 Storage = 160

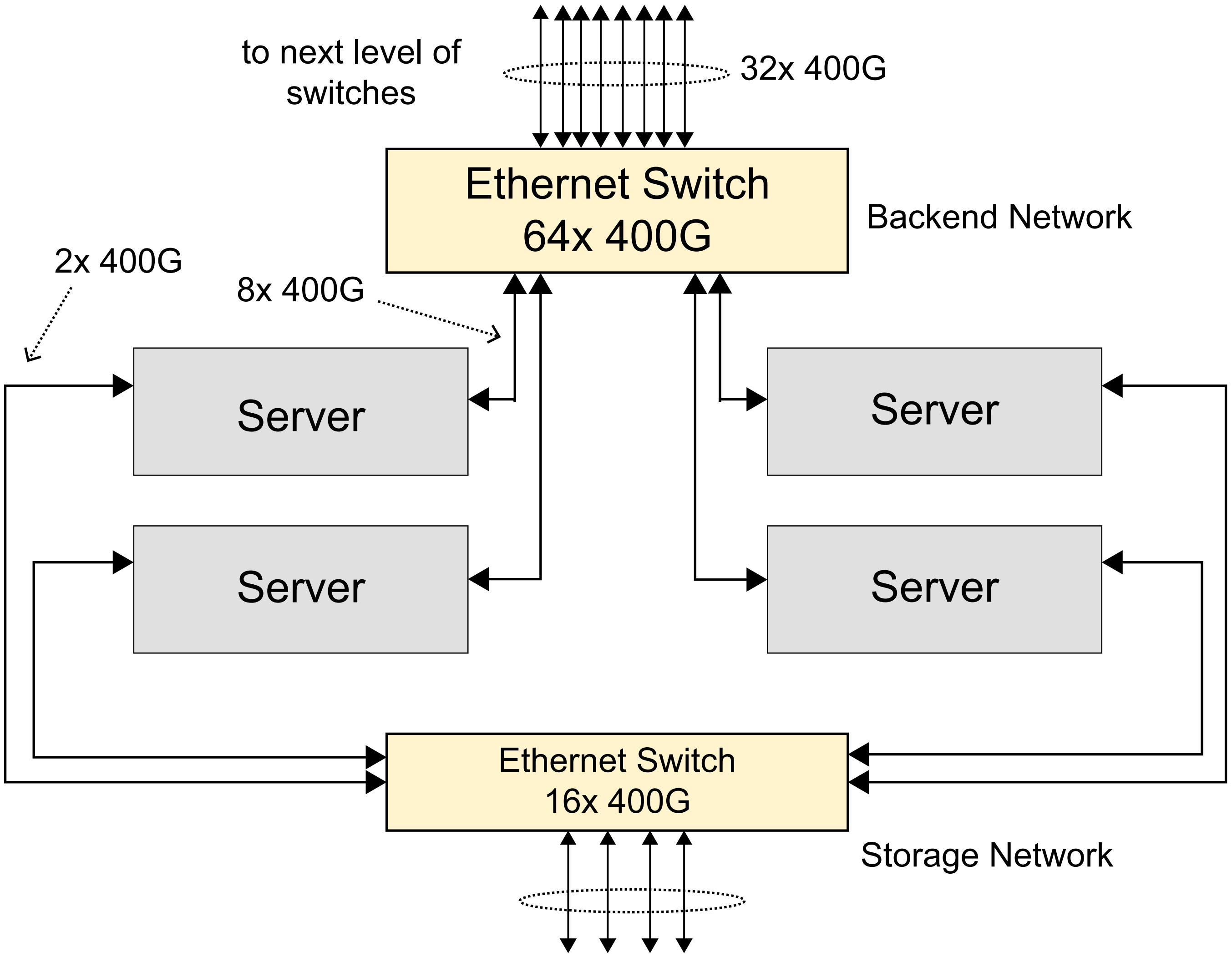
2.2 Scalable AI Clusters

Larger AI clusters require multiple layers of network switching in the backend network. These are generally built by defining a compact cluster and using that configuration as a building block. This section provides example building blocks and shows one example of a larger scale 1K cluster. Architecture of multi-level backend networks, in general, is not covered in this document.

It is common for backend networks to be full bandwidth, meaning these networks are not oversubscribed. One-half of the network switch ports connect to the servers and the other half of the network ports connect to the next level of switches.

A common building block using 64-port Ethernet switches consists of four servers, one backend network switch, one storage network switch, and one or two management switches. The connectivity for this building block is shown in Figure 10.

Figure 10: Scalable AI Cluster Connectivity – 64 port ToR

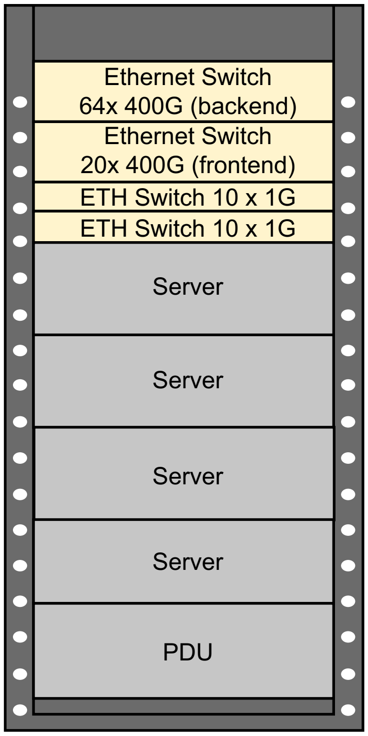


Similar to the compact AI cluster in Figure 10, the configuration of the building block is influenced by the available rack power. Deployments supporting 60kW or higher may fit a full building block in a single rack. This provides the most flexibility in connectivity options as it leads to the shortest distance between NICs and switches.

Some deployments may require two racks to support four servers if the available power is in the 30kW – 40kW range. When rack power is limited to < 20kW, a deployment may use a single server per rack.

Figure 11: Rack View for 32 GPUs, 60kW Rack Power – 64 port ToR Switch

4 Node Building Block Using 64-port Ethernet Switches



~60kW Rack Power

- 50kW - 4 servers per rack
- 3-5kW - switches
- 3-5kW - power, and so forth

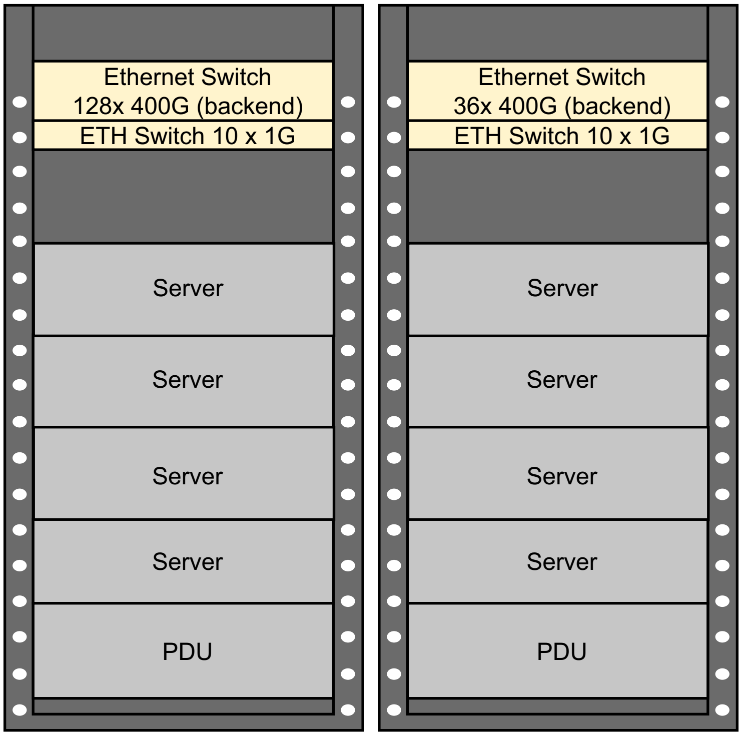
Connectivity

- Backend
  - Each server has 8x 400G
- Frontend/Storage
  - Each server has 2x 400G
- Management (redundant)
  - Each server has 2x 1G

NIC - Switch connection length: < 2 meters
Number of high-speed, intra-rack connections: 32 NIC to Switch + 8 Storage = 40

Figure 12: Rack View for 64 GPUs, 60kW Rack Power – 128 port ToR Switch

8 Node Building Block Using 128 Port Ethernet Switches



60kW Rack Power

- 50kW - 2 servers per rack
- 3-5kW - switches
- 3.5kW - power, and so forth

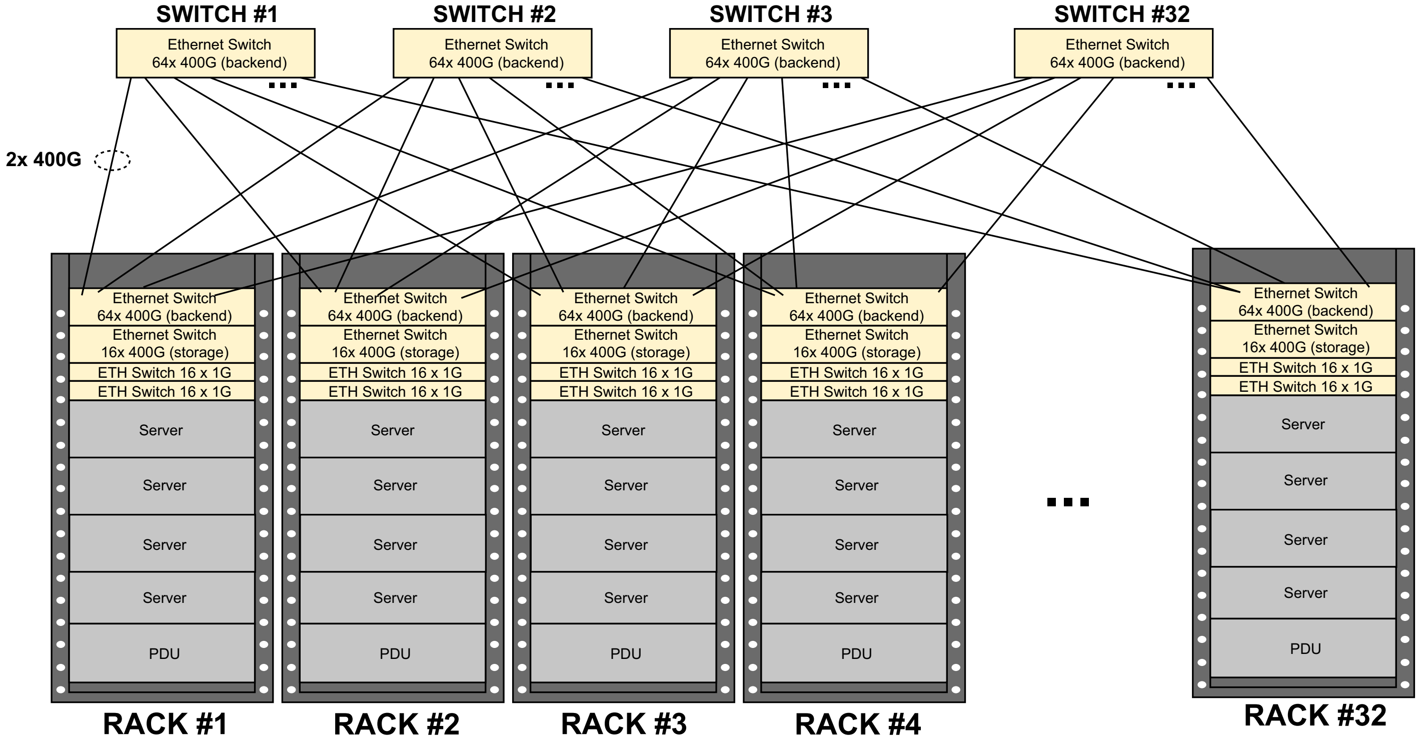
Connectivity within rack

- Backend
  - Each server has 8x 400G
  - 64x 400G to next switching tier
- Storage
  - Each server has 2x 400G
  - 4x 400G to storage array
- Management (redundant)
  - Each server has 2x 1G
  - 2x 1G redundant connect to central management network

NIC - Switch connection length: < 4 meters
Number of high-speed, intra-rack connections: backend: 64 + storage: 16 = 80

The building blocks are interconnected using one or more layers of switches. There are many topologies available for this network. Figure 13 uses a full Clos/fat-tree scaling to 1K nodes with 8K GPUs.

Figure 13: 1K GPU Cluster using 2 Tiers of Switching


SWITCH #1
Ethernet Switch 64x 400G (backend)
SWITCH #2
Ethernet Switch 64x 400G (backend)
SWITCH #3
Ethernet Switch 64x 400G (backend)
SWITCH #32
Ethernet Switch 64x 400G (backend)
2x 400G
Ethernet Switch 64x 400G (backend)
Ethernet Switch 16x 400G (storage)
ETH Switch 16 x 1G
ETH Switch 16 x 1G
Server
Server
Server
Server
PDU
RACK #1
RACK #2
RACK #3
RACK #4
RACK #32

3 Introduction to Cable Types

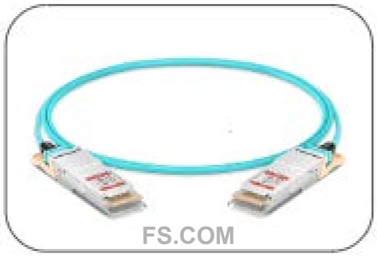
NIC-to-switch cables can be split into two categories: copper and optical.

Table 2: Cable Types

Cable type	Typical length for 400G	Characteristics
Direct Attached Copper (DAC)	Up to 5m	■ Short-range, high-speed connectivity ■ No active components ■ Low power (<0.1W) ■ Thicker cables → larger bend radius
Active Electrical (AEC)	Up to 7m	■ Additional circuitry to improve the signal ■ Reduced diameter compared to DACs ■ Slightly longer range than DACs ■ 5W at 400G typical power
Active Optical (AOC)	Up to 100s of m	■ Use transceivers on both ends to convert electrical to optical signal ■ Support long distances ■ Lighter weight with small bend radius ■ Up to 10W per termination, or 20W total at 400G
Pluggable Optical Transceivers	10s of m or km	■ Standalone modules that go into switch or NIC with separate optical cable that connects the transceivers ■ More flexible than AOCs when length needs to change ■ Similar power to AOCs
Linear Pluggable Optics (LPOs)	10s of m	■ Eliminates DSP from the transceivers ■ 40% less power than transceivers ■ Lower cost than AOCs/Transceivers

For additional information, see Appendix A, Cable Types.

DAC



AOC

Pluggable Optical Transceivers with cable

4 Cable Module and Transceiver Form Factors

DAC modules and optical transceivers plug into the connector cage of the NIC or switch. Both follow the same mechanical and electrical standards. The most prominent standards are QSFP56, QSFP56-DD, QSFP112, QSFP112-DD, and OSFP for Ethernet ports operating at 200G or higher. All use PAM-4 modulation.

- QSFP stands for Quad Small Form-factor Pluggable where quad indicates the module supports 4 electrical lanes. The number next to QSFP indicates the maximum speed of the SerDes supported per lane. For example, QSFP56 indicates support for 4 lanes of 56Gbps SerDes which supports a total bandwidth sufficient for 200G NIC.
- QSFP-DD adds Double Density, indicating the number of lanes is doubled. For example, QSFP56-DD supports 8 electrical lanes of 56G each.
- OSFP stands for Octal Small Form-factor Pluggable where octal indicates the module supports 8 electrical lanes.

Table 3 shows the differences between these modules/transceivers:

Table 3: QSFP-based and OSFP lane/BW characteristics

Connector Cage	Max bps/lane	# of lanes	Max BW (bps)	Modules Compatible with Connector Cage
QSFP56	50G	4	200G	QSFP56
QSFP56-DD	50G	8	400G	QSFP56 QSFP56-DD
QSFP112	100G	4	400G	QSFP56 QSFP112
QSFP112-DD	100G	8	800G	QSFP56 QSFP56-DD QSFP112 QSFP112-DD
OSFP	100G	8	800G	OSFP

Important Notes:

- QSFP cages and connectors are compatible with 4-lane modules for lower speeds. For example, QSFP56 modules can be plugged into QSFP56-DD or QSFP112 cage.
- OSFP cages and connectors are different and **NOT compatible** with any of the QSFP module variants, neither physically nor electrically. OSFP cages on the switch require an OSFP transceiver or a cable with OSFP connector.

5 BCM957608 Ethernet NICs and Connectors

The BCM957608 Ethernet NIC solutions come in two form factors: OCP v3.0 and PCIe Low Profile (LP). For each of these form factors, Broadcom offers 1-port and 2-port solutions. Table 4 shows the available cable types.

Table 4: Connector Cage Types

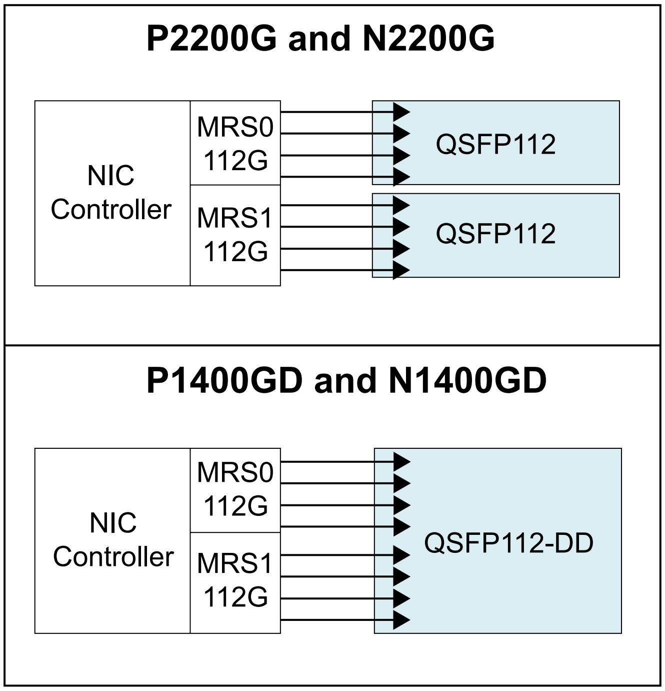
Product	**N1400GD**	**N2200G**	**P1400GD**	**P2200G**
Form Factor	OCP v3.0 SFF		PCIe HHHL	
PCIe Host	x16 PCIe Gen5			
Network SerDes	112G PAM-4			
Card image				
Connector cage	QSFP112-DD	QSFP112	QSFP112-DD	QSFP112
MPN	BCM957608- N1400GDP00	BCM957608- N2200GQP00	BCM957608-P1400GDF00	BCM957608- P2200GQF00
Ethernet Ports	1	2	1	2
Bandwidth/Port	400G	200G	400G	200G

- Front end connectivity: all of these cards are well suited for 1x400G or 2x200G
- Backend connectivity: 1x 400G is commonly used to support maximum bandwidth flows, 2x 200G is available for systems with redundant switches and load balancing

5.1 Multi-rate SerDes to Connector Cage Connectivity

As shown in Table 4, BCM957608 boards use either QSFP112 or QSFP112-DD connectors. Figure 14 shows how the connector cage is connected to the 8 lanes of the internal Multi-rate SerDes (MRS) blocks.

Figure 14: Connector Cage Connecting to 8 Lanes



For the NICs with 4 lanes per port (P220G, N2200G), the maximum network speed per port can be configured as:

- 200G (4 lanes at 56G/lane or 2 lanes at 112G/lane)
- 400G (4 lanes at 112G/lane)

For the NIC with 8 lanes per port (P1400GD, N1400GD), the maximum network speed per port can be configured as:

- 400G (8 lanes at 56G per lane) or
- 400G (4 lanes at 112G per lane)

6 Cable Configurations for Broadcom BCM957608 Ethernet NICs

Connectivity selection to connect a Broadcom BCM957608 Ethernet NIC with a network switch depends on the following variables:

1. Distance between NIC and switch
   – DAC is an option for up to 4 meters
   – AEC is an option for ~7m
2. Power, cooling, and cable routing
   – DAC and AEC are lower power than optics
   – Bend radius should be considered for lower gauge DAC
3. Cost considerations
   – DAC and AEC are lower cost than optics
   – LPO is lower cost than non-LPO optics
4. Connector type on the network switch:

Switch Type	Max SerDes Speed	Max Speed Per Port	Cage Connector
Tomahawk 4-based	56G	400G (8 x 50)	QSFP56-DD
Tomahawk 5-based	112G	800G (8 x 100)	QSFP112-DD OSFP

5. Connector type on the Broadcom BCM957608 Ethernet NIC:

NIC	Max Speed Per Port	Cage Connector
P2200G	200G (4 x 50) 400G (4 x 100)	QSFP112
N2200G		
P1400GD	400G (8 x 50 or 4 x 100)	QSFP112-DD
N1400GD		

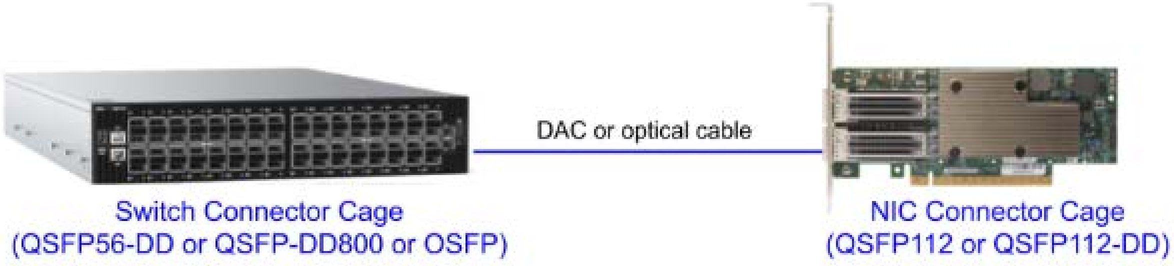
NOTE: Only the N1400GD and P1400GD NICs can achieve 400G (in 8 x 50 configuration) if the switch is based on Tomahawk 4 since the switch SerDes supports up to 56G per lane.

The connectivity options are shown in Table 5.

Table 5: Cable Configurations Based on Switch/NIC Connectors and Port Configuration

Switch	Connectivity Type	Power	Cable Configuration Name	Switch Port Configuration	Switch side	NIC side
					Cable Connector	Cable Connector
Tomahawk 4 or other 56G per-lane switches	DAC (1-5m)	~0W	D1 (1:2 split)	2x 200G	QSFP56-DD	2xQSFP56
			D2 (1:2 split)	1x 400G	QSFP56-DD	2xQSFP56
			D3	1x 400G	QSFP56-DD	1xQSFP56-DD
			D4	1x 200G	QSFP56	1xQSFP56
			D5	2x 200G	QSFP56-DD	QSFP56-DD
	Optics non-LPO (1m - kms)	10-15W per transceiver	O1 ([a]Gearbox)	1x 400G	QSFP56-DD	1xQSFP112
			O2 (1:2 split)	2x 200G	QSFP56-DD	2xQSFP56
			O3	1x 400G	QSFP56-DD	1xQSFP56-DD
Tomahawk 5 or other 112G per-lane switches	DAC (1-5m)	~0W	D21	2x 400G	OSFP112	1xQSFP112
			D22	1x 400G	QSFP112	1xQSFP112
	Optics non-LPO (1m - kms)	10-15W per transceiver	O21	2x 400G	OSFP112	1xQSFP112
			O22	1x 400G	QSFP112	1xQSFP112
	LPO (1m - kms)	50-70% less than standard transceivers	L21	2x 400G	OSFP112	1xQSFP112

[a] Using Gearbox splitter cable (1x 400G SR8 → 4x 112G SR4)



7 Cable Selection Process

The choice of DACs vs. AOC/transceivers depends on multiple factors as shown in Table 6.

Table 6: Cable Selection Factors

Decision Factor	Notes
Cost	Advantage: DACs DACs have much lower cost than AOCs and/or transceivers.
Power	Advantage: DACs Passive DACs consume negligible power while AOCs and/or transceivers are usually in the 10-15W range.
Length	Advantage: AOCs/transceivers Intra-rack: Both DACs or AOCs/transceivers are a good option. Inter-rack or longer distances (5m+): Up to 7m AEC cables can be an option, for longer distances AOCs/transceivers have an advantage.
Cabling organization and maintenance	Advantage: AOCs/transceivers Optical cables are much thinner, and have a small bend radius, thus a large number of cables can be neatly organized in the rack.

Consider the AI Cluster shown in Figure 12 with eight servers in two racks. Using a 128 x 400G Tomahawk 5 switch (with QSFP112-DD connectors) and 1x 400G BCM957608 NIC (P1400G2 with QSFP112), a Y-cable would be used.

The distance between the NIC and switch ranges up to between 2.5 – 3 meters from the lower server in one rack to the switch in the top of the other rack. A total of 80 high-speed cables are needed. Table 7 provides example options for this rack configuration.

Table 7: Examples Options

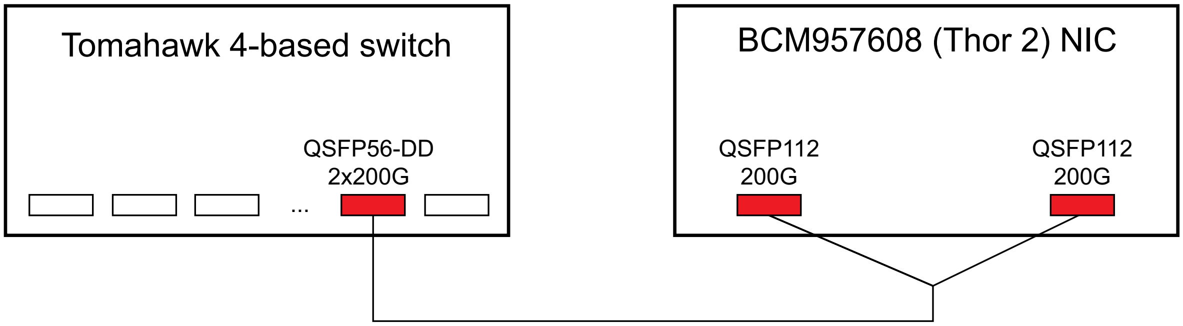
	28AWG DAC	AEC	AOC
Example Part Number	Amphenol NDYRYH-0003	TBD	Accelink RTXM500-1XX
Diameter	Largest	Medium	Smallest
Bend radius	Largest	Medium	Smallest
Power – per cable	0W	4.5W	10W
Power – 80 cables	0W	360W	800W
Cost – per cable	Lowest	Medium	High

A more detailed example of the cable selection process is provided in Appendix B, Example for Cable Selection Process.

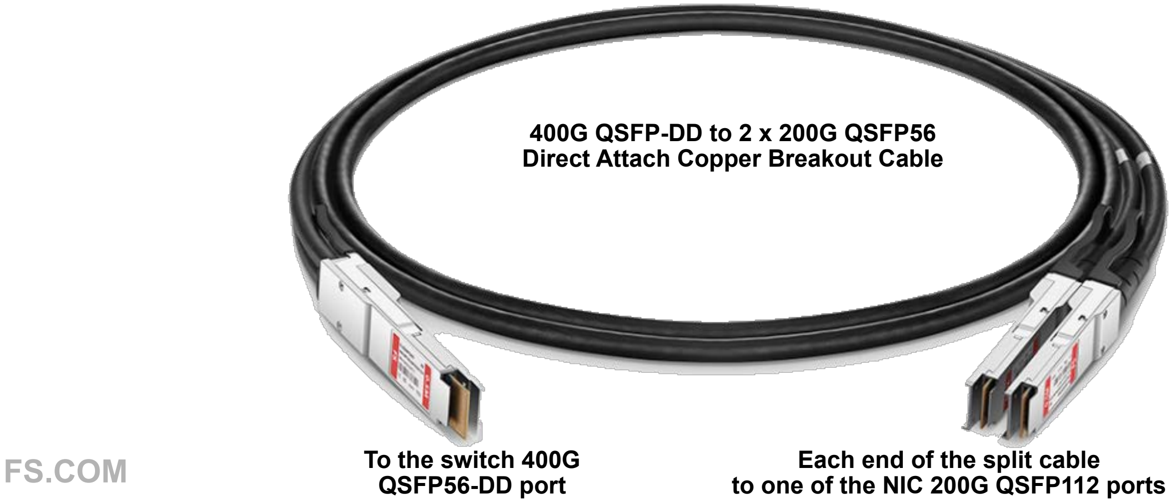
8 Example Configurations for DAC Cables

This section contains example configurations for DAC cables.

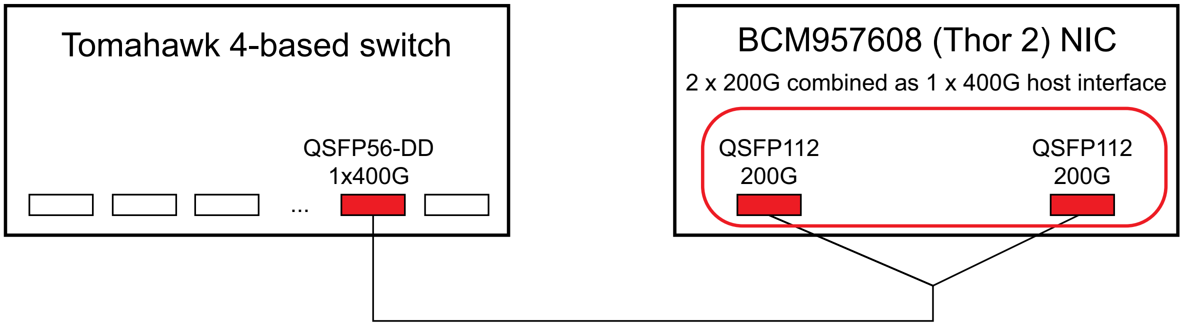
8.1 DAC Cable Configuration D1 – Y Cable



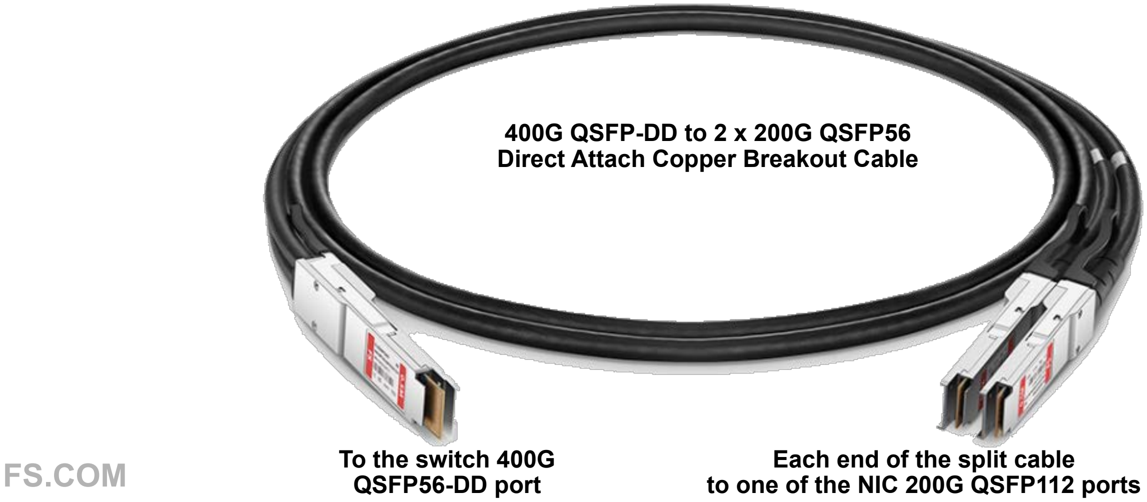
- Switch port configuration is 2 x 200G (8 lanes of 50G) using QSFP56-DD cage.
- NIC port configuration is 4 lanes of 50G using QSFP112 cage.
- The required cable has to be a splitter/breakout cable (Y cable). One end with QSFP56-DD module towards the switch and two ends with QSFP56 modules towards the NIC.


400G QSFP-DD to 2 x 200G QSFP56
Direct Attach Copper Breakout Cable
FS.COM
To the switch 400G
QSFP56-DD port
Each end of the split cable
to one of the NIC 200G QSFP112 ports

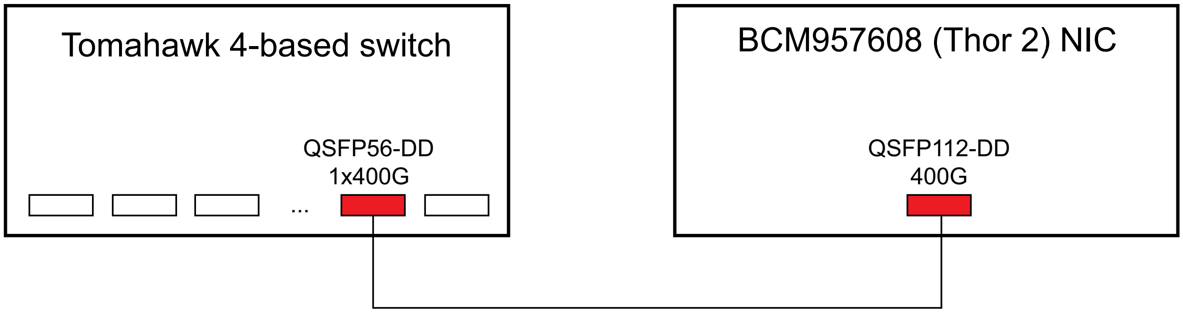
8.2 DAC Cable Configuration D2 – Y Cable



- Switch port configuration is 400G (8 lanes of 50G) using QSFP56-DD cage.
- The two ports on the NIC are set as 200G (4 lanes of 50G) using QSFP112 cages, but presented by the NIC FW as a single 400G interface to the host.
- The required cable has to be a splitter/breakout cable (Y cable). One end with QSFP56-DD module towards the switch and two ends with QSFP56 modules towards the NIC.

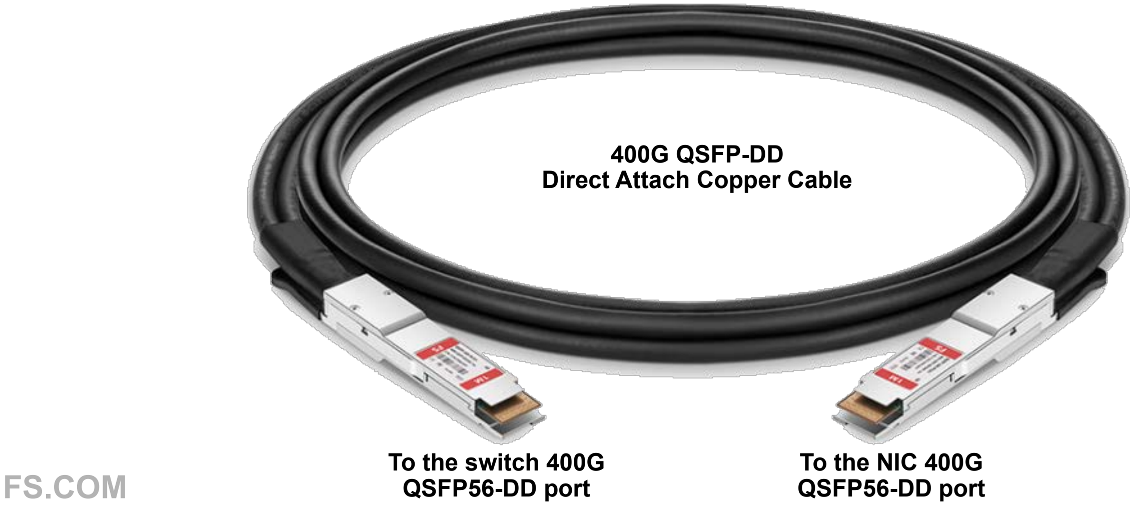

400G QSFP-DD to 2 x 200G QSFP56
Direct Attach Copper Breakout Cable
FS.COM
To the switch 400G
QSFP56-DD port
Each end of the split cable
to one of the NIC 200G QSFP112 ports

8.3 DAC Cable Configuration D3

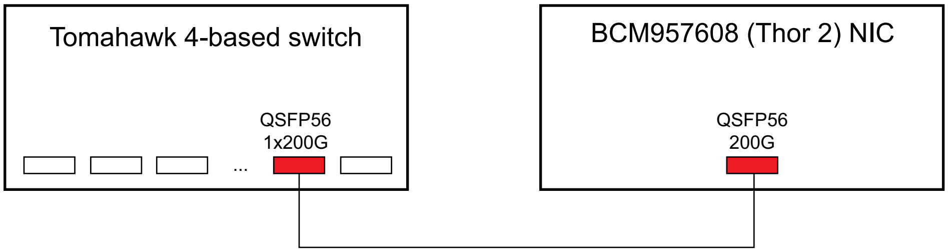


- Switch port configuration is 400G (8 lanes of 50G) using QSFP112-DD cage.
- NIC port configuration is 400G (8 lanes of 50G) using QSFP112-DD cage.
- The required cable needs to have QSFP56-DD on both ends.

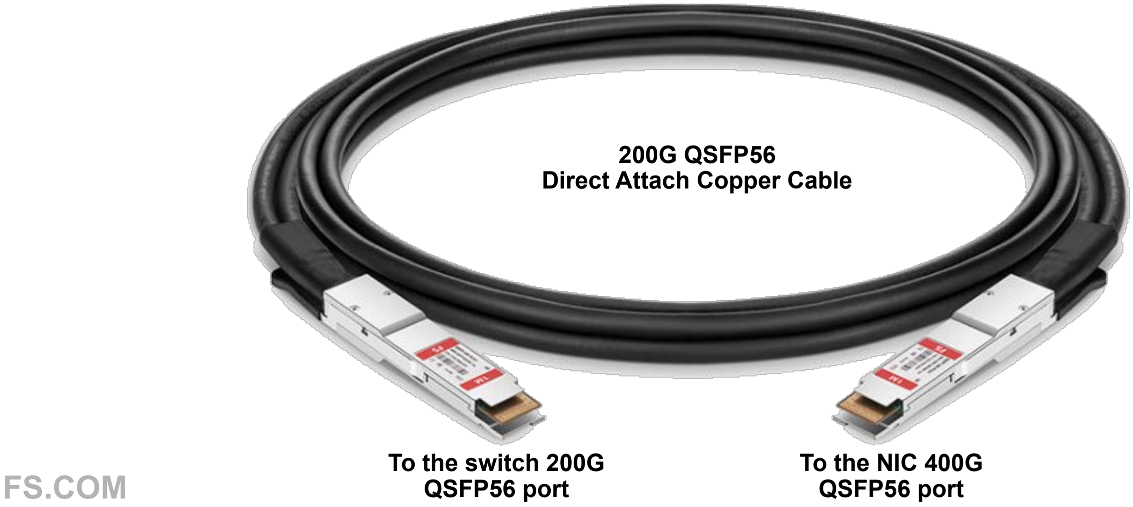
NOTE: QSFP112-DD connector on the NIC is compatible with QSFP56-DD cable module.



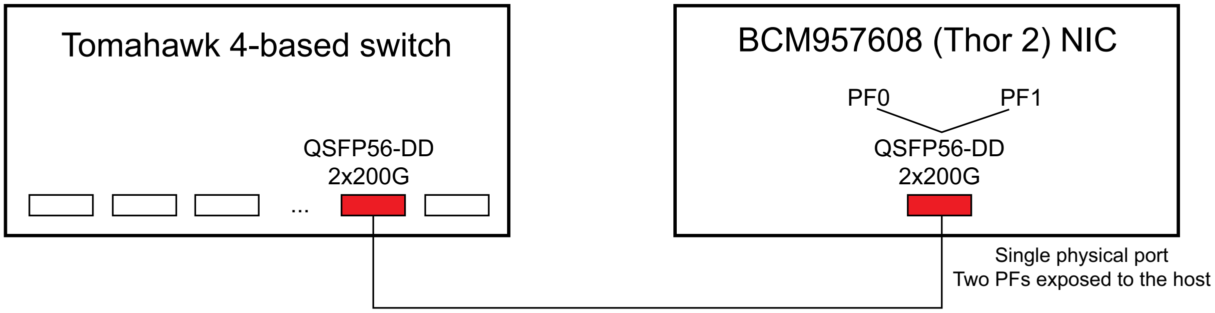
8.4 DAC Cable Configuration D4



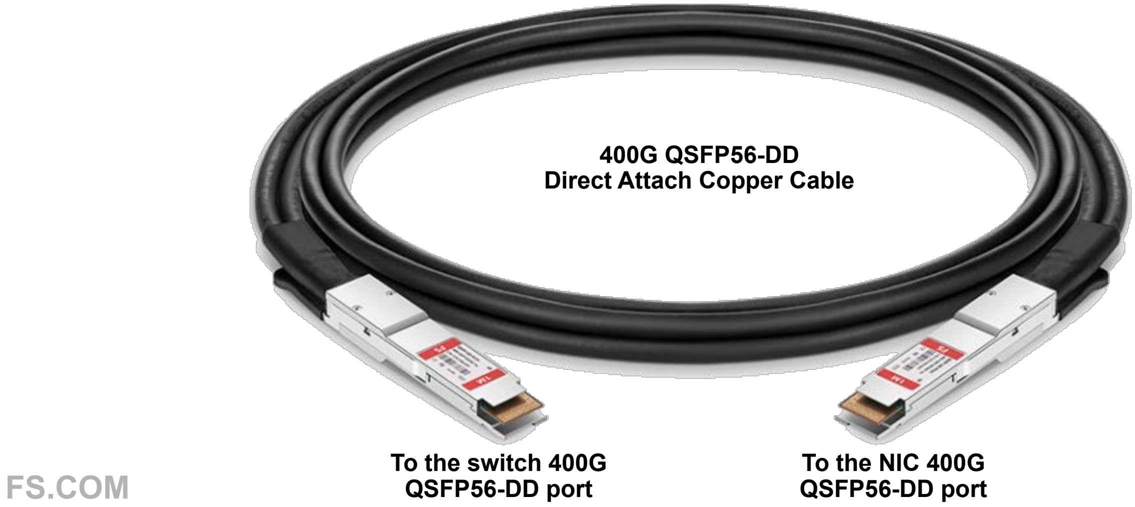
- Switch port configuration is 200G (4 lanes of 50G) using QSFP56 cage.
- NIC port configuration is 200G (4 lanes of 50G) using QSFP56 cage.
- The required cable needs to have QSFP56 on both ends.



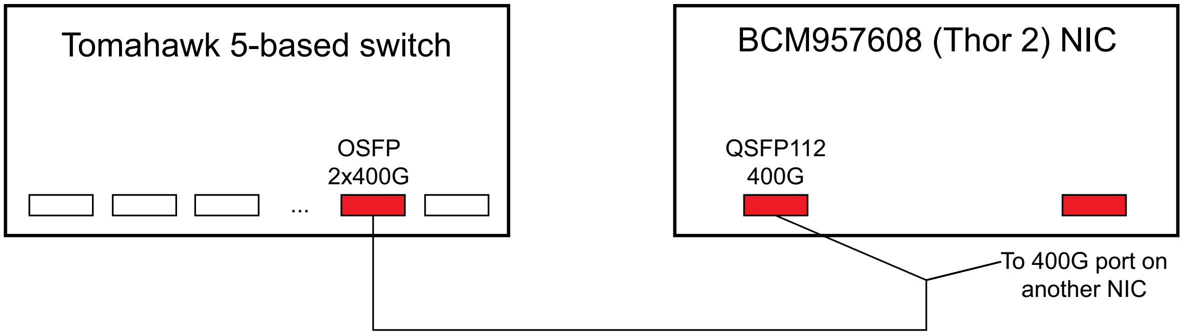
8.5 DAC Cable Configuration D5



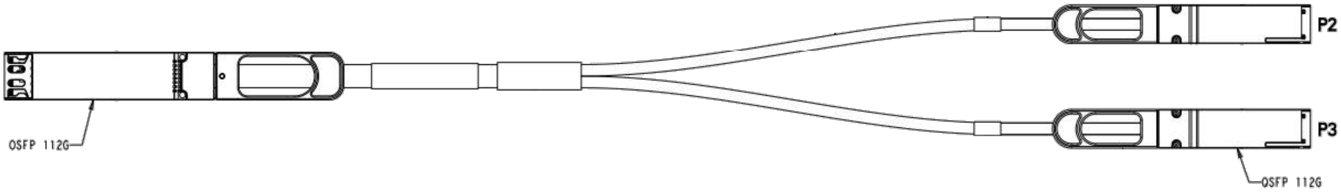
- Switch port configuration is 2x200G (on a QSFP56-DD port on the switch).
- NIC port configuration is 2x200G (on a QSFP56-DD port on the NIC).
- The required cable needs to have QSFP56-DD on both ends. The physical cable is identical to cable configuration D3.



8.6 DAC Cable Configuration D21 – Y Cable (800G to 2x400G)



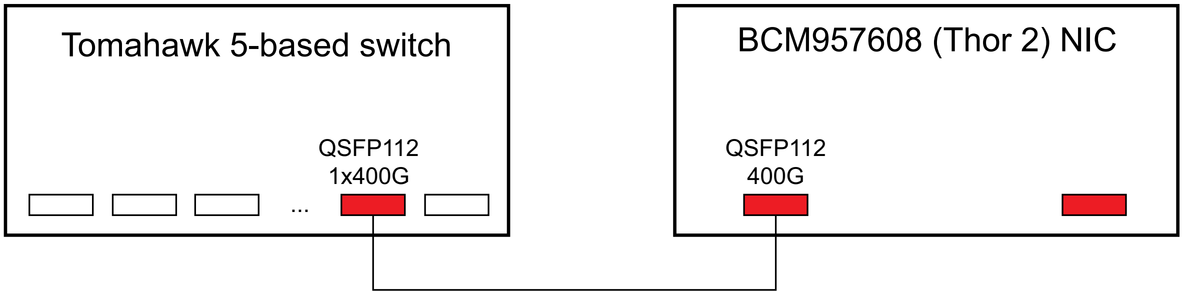
- Switch port configuration is 800G (8 lanes of 100G) using OSFP cage.
- NIC port configuration is 400G (4 lanes of 100G).
- The required cable needs to have OSFP112 on one end and QSFP112 on the other.



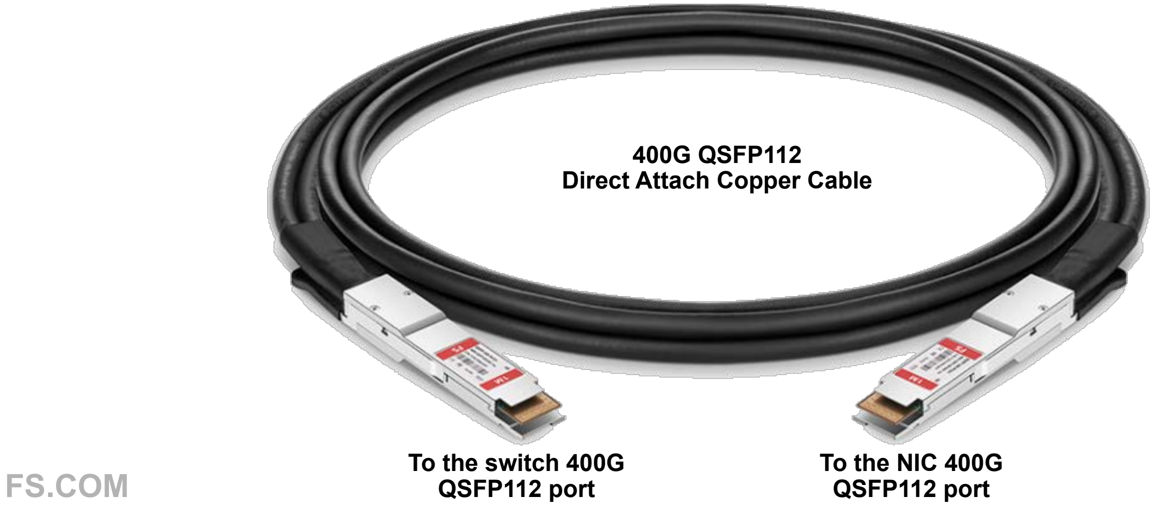
To the switch 800G OSFP port

To the NIC 400G QSFP112 port

8.7 DAC Cable Configuration D22



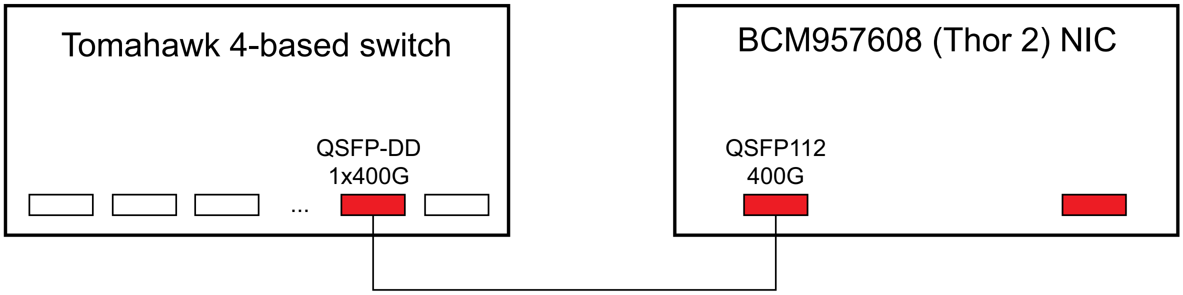
- Switch port configuration is 400G (4 lanes of 100G) using QSFP112 cage.
- NIC port configuration is 400G (4 lanes of 100G) using QSFP112 cage.
- The required cable needs to have QSFP112 on both ends.



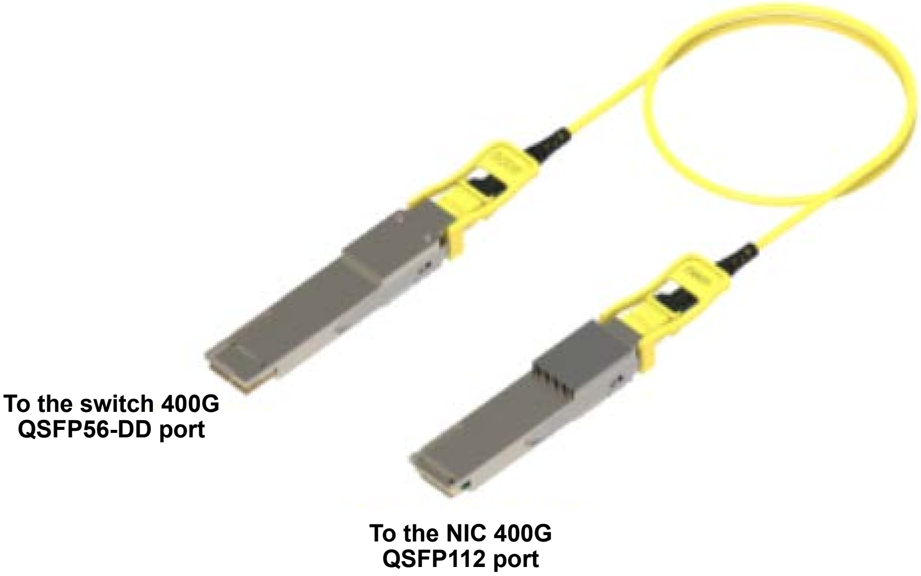
9 Examples for AOC Cables and Optical Transceivers

When using transceivers + optical cables, it is important to ensure the transceivers on both the switch and the NIC side operate at the same center wavelength. For AOC cables, the switch/NIC transceivers and the actual cable are packaged by the vendor and this condition is already correct.

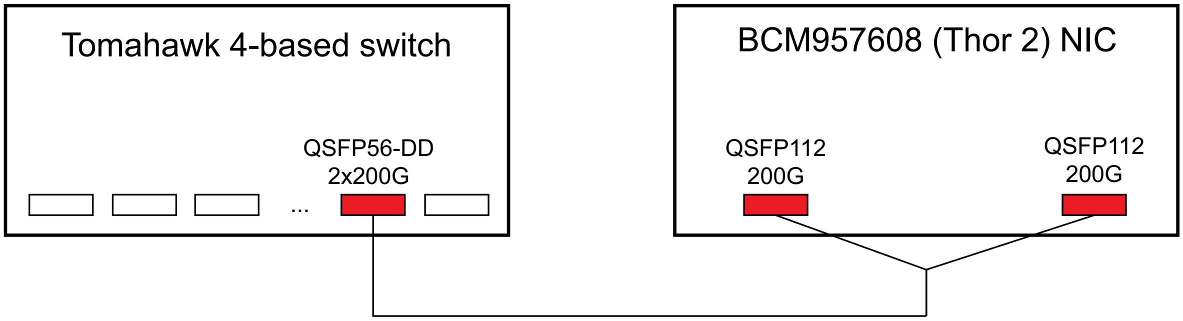
9.1 Cable Configuration O1



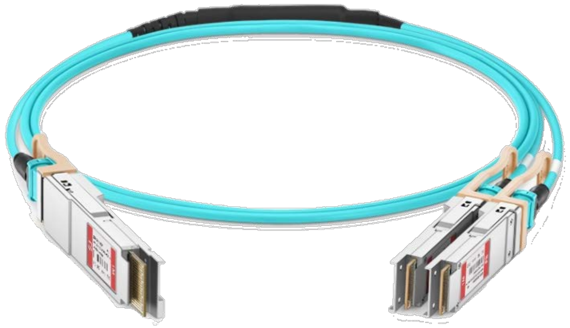
- Switch port is 400G QSFP-DD (8 lanes of 50G).
- NIC port is 400G QSFP112 (4 lanes of 100G).
- The required cable needs to have a QSFP56-DD transceiver on one end and QSFP112 transceiver on the other end.



9.2 Cable Configuration O2

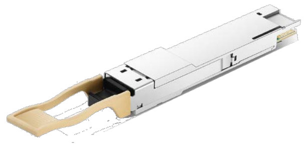
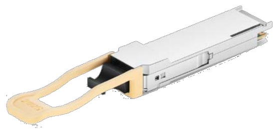


- Switch port configuration is 2 x 200G (8 lanes of 50G).
- NIC port configuration is 200G (4 lanes of 50G).
- If using AOC cable, it will need to be a splitter/breakout cable (Y cable).

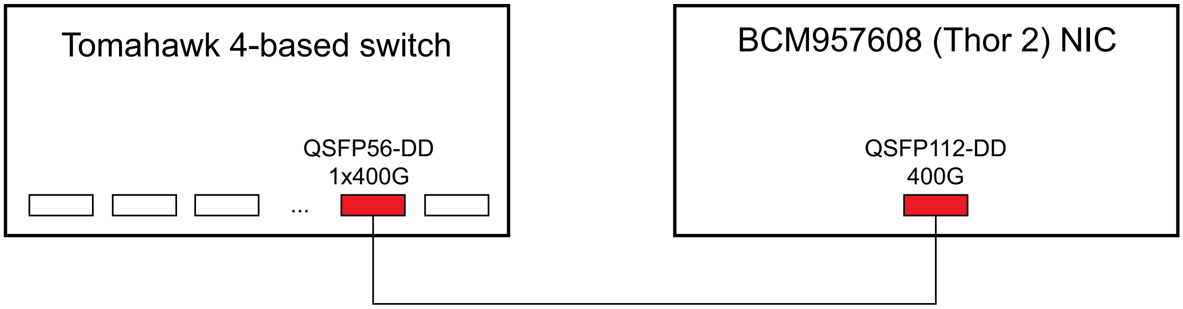


To switch QSFFP56-DD port **To the NIC QSFP112 ports**

- If using transceivers:
  - QSFP56-DD transceiver is needed on the switch side.
  - QSFP56 transceivers are needed for each port on the NIC side.
  - Split optical cable is needed to split the 8 lanes from the QSFP-DD transceiver to 2x4 lanes going to each transceiver on the NIC.

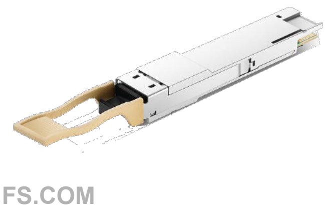
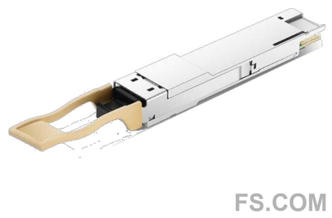
Switch side: QSFP56-DD SR8	Split optical cable (8 lanes to 2 x 4)	NIC side: QSFP56 SR4 (one transceiver for each port)
FS.COM 	 FS.COM	 FS.COM

9.3 Cable Configuration O3

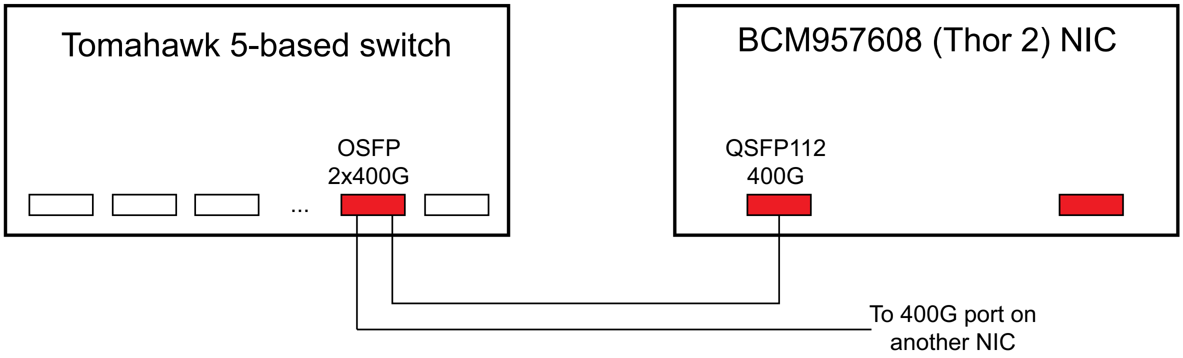


- Switch port configuration is 400G (8 lanes of 50G).
- NIC port configuration is 400G (8 lanes of 50G).
- If using AOC cable, QSFP56-DD is needed on both ends of the cable.

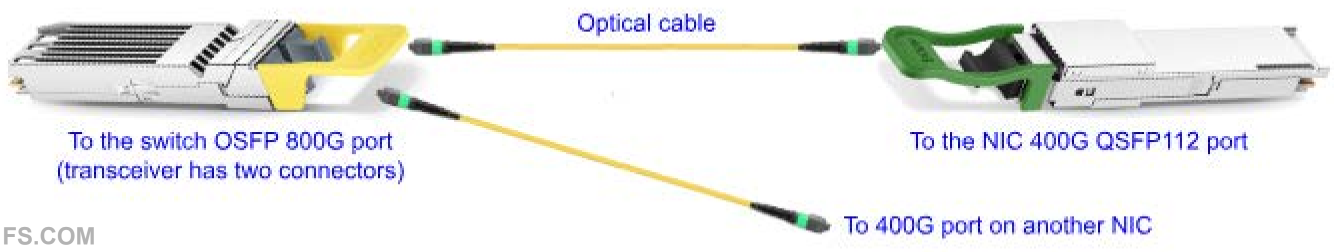
- If using transceivers, both the switch and the NIC transceiver need to be QSFP56-DD.

Switch side: QSFP56-DD SR8	Optical cable (8 lanes)	NIC side: QSFP56-DD SR8
		

9.4 Cable Configuration O21

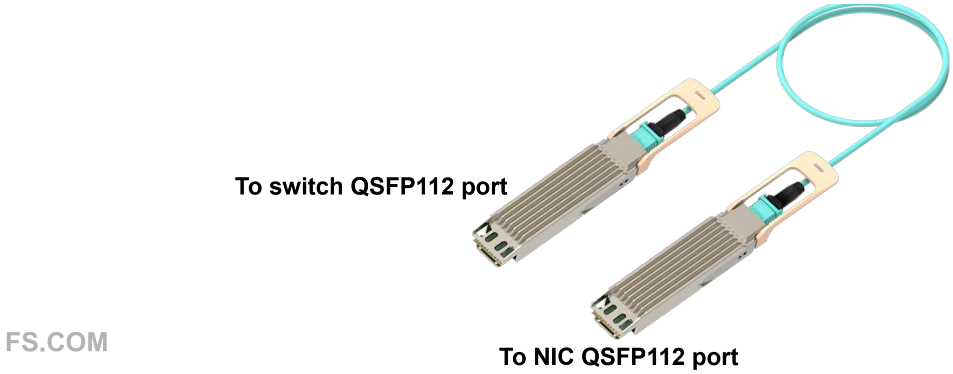


- Switch port configuration is 800G OSFP112 (8 lanes of 100G).
- NIC port configuration is 400G QSFP112 (4 lanes of 100G).
- The required cable should have an OSFP112 transceiver on one end and a QSFP112 transceiver on the other end.
- Make sure the center wavelength on both transceivers is the same (for example, 1310 nm).

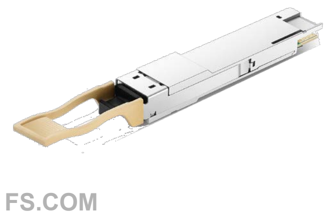
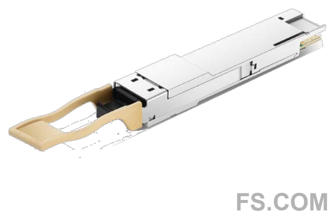

Optical cable
To the switch OSFP 800G port
(transceiver has two connectors)
To the NIC 400G QSFP112 port
To 400G port on another NIC
FS.COM

9.5 Cable Configuration O22

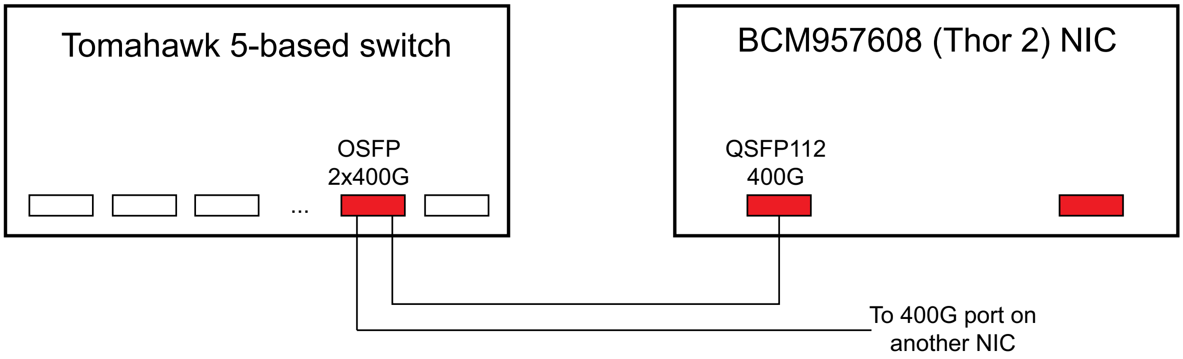
- Switch port configuration is 400G QSFP112 (4 lanes of 100G).
- NIC port configuration is 400G QSFP112 (4 lanes of 100G).
- If using AOC cable, QSFP112 is needed on both ends of the cable.



If using transceivers, both the switch and the NIC transceiver need to be QSFP112.

Switch side: QSFP56-DD SR4	Optical cable	NIC side: QSFP112 SR4
		

9.6 Cable Configuration L21 (LPOs)



- Switch port configuration is 800G OSFP112 (8 lanes of 100G)
- NIC port configuration is 400G QSFP112 (4 lanes of 100G)

Instead of using AOC cable or transceivers, this cable uses an LPO cable. This is functionally equivalent to Cable Configuration O21.

10 Rack Cabling

The connectivity between the servers in the rack with the top-of-rack (ToR) switch (intra-rack connectivity) and the connectivity between ToR switches and the next layer of leaf or spring switches (inter-rack connectivity) can be implemented with either DAC or AOC cables.

AOCs are thinner and have a smaller bend radius (for example, they are more bendable). However, they also come with higher power for the transceivers on both ends and higher cost than passive DAC cables.

The choice of the intra-rack cables depends on the size of the rack, the number of servers in the rack, the number of NICs in each rack, the distance to the ToR switch, and the number of ports (radix) in the ToR switch.

The following figures show several examples of intra-rack connectivity. In the first example, the switch is at the top of the rack and in the second example the switch is in the middle. The second approach allows a shorter distance to the servers at the top or the bottom allowing for shorter cables. Shorter cable distances allow the use of DACs, which in turn means lower cabling cost and power.

Figure 15: Server/Switch Rack Configuration 1 (Switch is at the Top of the Rack)

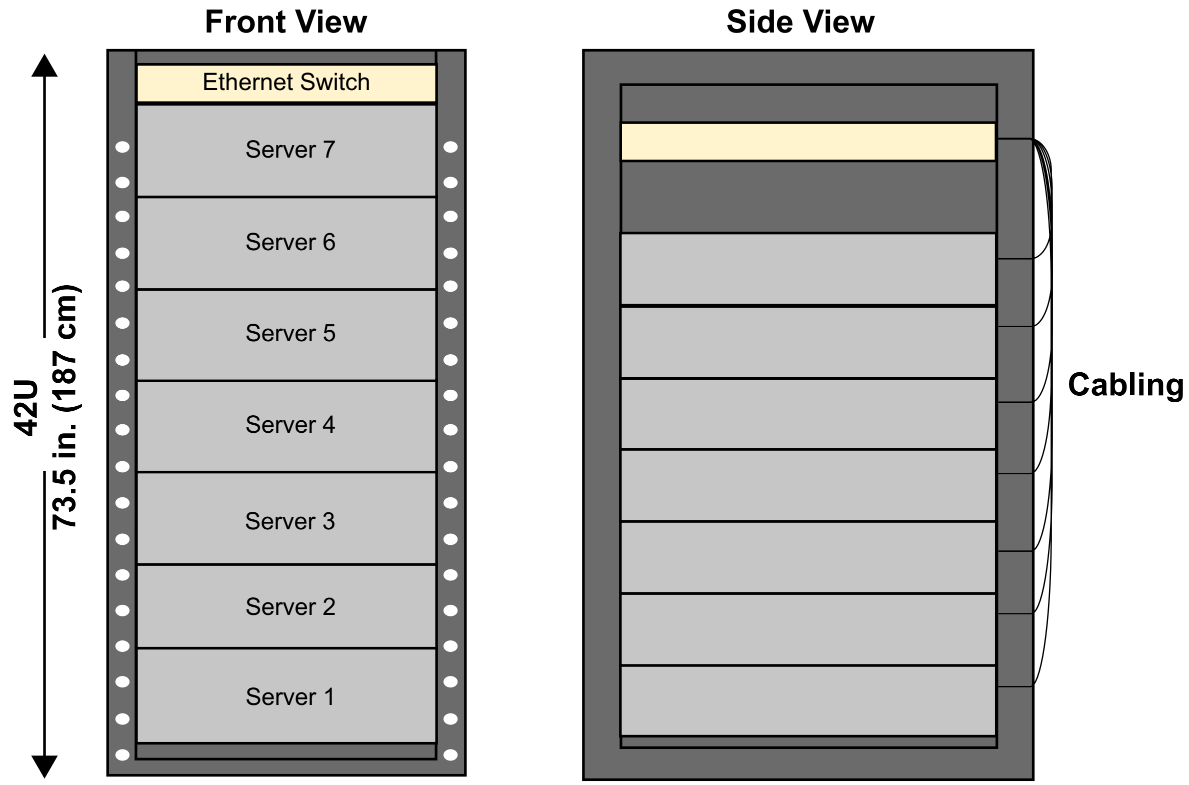
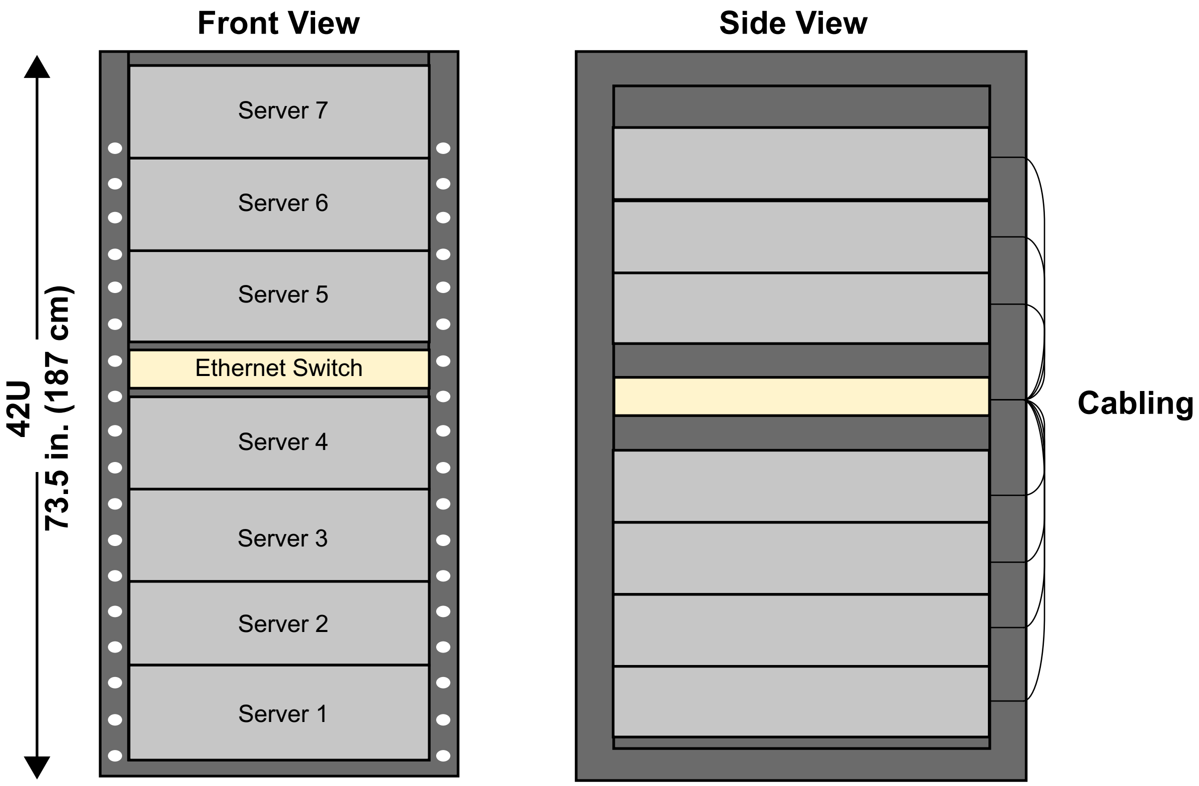


Figure 16: Server/switch rack configuration 2 (Switch is in the Middle of the Rack)



Appendix A: Cable Types

A.1 DAC – Direct Attached Copper Cable

DAC cables provide short-range, high-speed connectivity using copper cables. Passive DACs have minimal electronics and therefore draw very low power (typically less than 0.1W since there are no active components to boost the signal) and are lower cost. At higher speeds, the cable diameter limits the bend radius, which must be considered along with the cable weight. While these cables generally support lower distances than optical, DAC can reach distances up to 5m for 400Gbps (longer for lower speeds).

A.2 AEC – Active Electrical Cable

AEC cables are a version of DACs, which include additional circuitry in the connector module that amplifies the signal or enhances its integrity. These typically draw less than 3W and can reach longer distances than passive DACs. For 400G speeds, AEC cables can reach distances of about 5-7m (longer for lower speeds). Cable diameter can also be reduced compared to passive DAC, which allows for smaller bend radius and easier organization within the rack.

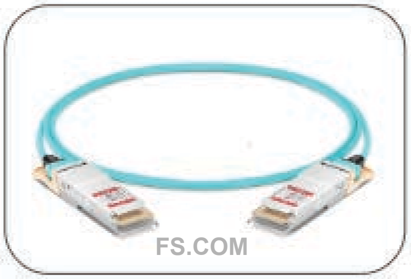
A.3 AOC – Active Optical Cable

AOC cables are high-speed cables that use optical fiber for transmission of data. AOCs have transceivers at both ends of the cable that convert electrical to optical signals and vice versa. AOC cables are of fixed length since the two transceivers and the optical cable that connects the transceivers are one integrated unit. These cables are more expensive and higher power, but are lighter weight and support longer distances.

A.4 Pluggable Optical Transceivers

Pluggable optical transceivers are standalone modules that go into the switch or NIC and convert electrical to optical signals and vice versa. A separate optical cable is plugged into both transceivers. The advantage of separate transceivers is that they can be reused in a different environment by only changing the optical cable that connects the transceivers. In essence, two optical transceivers plus an optical cable are functionally equivalent to an AOC.

DAC



AOC

Pluggable Optical Transceivers
with Cable

A.5 LPO – Linear Pluggable Optics

LPOs eliminate a DSP that is normally present in the optical transceivers and AOC cables. The elimination of the DSP means that no signal processing or modulation is performed and the output signal is a linear representation of the input signal. Any required signal correction is handled by the NIC SerDes and/or the remote network switch SerDes.

The elimination of DSP processing in the transceiver leads to lower cost and lower power dissipation. The LPO technology allows for 40% lower power compared to DSP-based optical transceivers, and lower latency. LPO modules also run at lower temperatures, which significantly improves reliability.

At the time of this writing, the LPO standard is in development but it is not finalized. Different aspects of the functionality are provided in CEI-112-Linear-PAM4, IEEE CL 124, 140, 151, 167. Until the standard is finalized, customers should stay within the approved vendor list (AVL) boundaries to make sure the LPO is compatible with the NIC and switch.

A.6 Optical Cables

The following section refers to both optical cables that constitute a part of AOC cables as well as standalone optical cables that connect two optical transceivers. Optical cables come in a variety of sizes or characteristics shown in Table 8.

Table 8: Optical Cable Sizes and Characteristics

Categorization	
Light mode	Single mode (SM) or Multimode (MM)
Optical cables	OS1/OS2 (SM), OM1/OM2/OM3/OM4/OM5 (MM)
Cable coloring	Yellow (OS2), Aqua (OM3), Aqua or Violet (OM4), Lime Green (OM5)
Cable connectors	LC, SC, ST, FC, MTP, and MPO
Cable polish	PC, UPC, APC

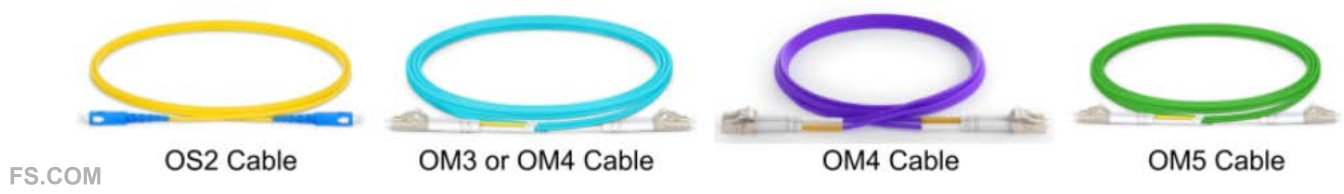
Single Mode vs. Multimode: Single mode fiber supports higher bandwidth than multimode. Multimode cables have a larger core diameter that limits the maximum length due to modal dispersion. SM is generally used for long distances as these support distances greater than 500m. Multimode is less expensive than single mode fiber.

Optical Cables: OS1/OS2 is used for single mode fiber and OM1/OM2/OM3/OM4/OM5 for multimode fiber. OM1/OM2 use LED transmitters while OM3/OM4/OM5 use laser multimode transmitters.

For 200G or 400G Ethernet applications, only OS2 is recommended for single fiber. OM3, OM4, and OM5 cables are recommended for multimode. The cables can be recognized by the following industry-set colors shown in Table 9.

Table 9: Optical Cable Colors

Fiber type	Fiber Cable Color	
Single-mode OS2	Yellow	
Multi-mode OM3	Aqua	
Multi-mode OM4	Aqua	Violet
Multi-mode OM5	Lime Green	



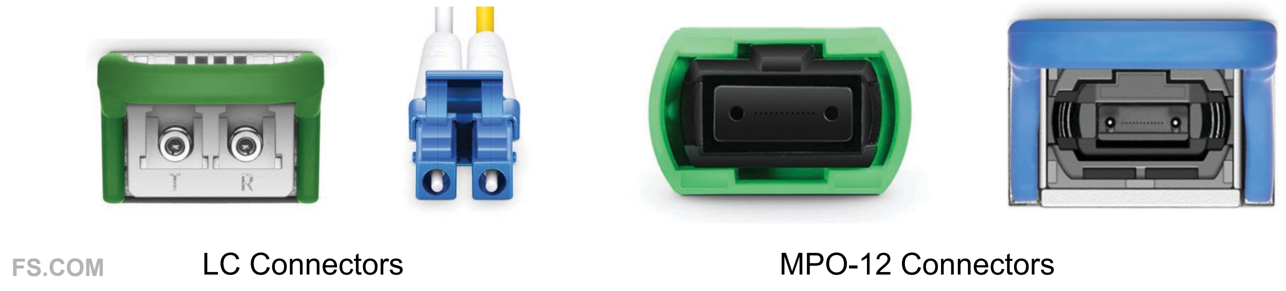
Optical Cable Connectors – Optical cable connectors have a variety of form factors, some of which are listed in Table 10. Figure 17 provides examples of LC and MPO-12 connectors.

Table 10: Optical Connectors

Acronym	Name	Notes
LC	Lucent Connector	■ Smaller (~½ size of SC) ■ Latch, push/pull
SC	Subscriber Connector (aka Standard or Square)	■ Larger, less expensive than LC ■ Latch, push on/pull off
ST	Straight Tip	■ Easy to use, keyed ■ Push-in and twist
FC	Ferrule Connector	■ Older ■ Threaded container
MPO MTP[a]	Multi-fiber Push-On Multi-fiber Termination Push-on	Multiple optical fibers using ribbon cable; snap, push/pull

[a] MTP is a registered trademark for an MPO connector from US Conec.

Figure 17: LC and MPO-12 Connector Examples



Cable Connector Polish Style: Connectors can be divided into PC (Physical Contact), UPC (Ultra Physical Contact), and APC (Angled Physical Contact). For high-bandwidth, only UPC and APC are used. APC connectors feature a fiber end face that is polished at an eight-degree angle, thus improving return loss and minimizing optical reflection. UPC connectors are polished with no angle.

A.7 Optical Transceiver Range

Each optical transceiver is designed to operate over a specific distance and bandwidth. The main factors impacting the range are the transceiver design, type of fiber, and the data rate. The range can be designated as SR (short range), DR (data center range), XDR (extended DR), LR (long range), and others. This designation is followed by the number of optical channels (usually 4 or 8). For example, SR8 indicates a short range optical cable with 8 channels.

For more information on optical cables and transceivers, see the FS Community.

Appendix B: Example for Cable Selection Process

The following major steps can be utilized when selecting a cable for switch-to-NIC connectivity:

1. Determine the switch port type and speed.
2. Determine the NIC port type and speed for the specific BCM957608 NIC.
3. Determine the choice of DAC vs. AOC cable.
   a. For up to 5m, choose a DAC solution for lowest power/cost.
   b. For up to 7m, AEC cable might be the next option to consider.
   c. For longer distances spanning 10s of meters or more, AOC will be needed - choose the AOC cable or transceivers that satisfies the port type on both ends.

B.1 Determine the Switch Port Type and Port Speed

For most 400G/800G switches this is either OSFP or QSFP-DD or QSFP-DD800.

Table 11 provides examples of commercial switches and associated switch ports.

Table 11: Commercial Switches and Switch Ports

Switch model	Port type	Port Speed
Arista 7060X	QSFP56-DD	400G
Arista 7280R3	OSFP or QSFP56-DD	400G
Arista 7358X4	OSFP or QSFP56-DD	400G
Arista 7388X5	QSFP56-DD	400G
Arista 7500R3	OSFP or QSFP56-DD	400G
Arista 7800R3	OSFP or QSFP56-DD	400G
Dell Z9664F-ON	QSFP56-DD	400G
Dell Z9432F-ON	QSFP56-DD	400G
Dell Z9332F-ON	QSFP56-DD	400G
Ruijie RG-S6980-64QC	QSFP56-DD	400G

B.2 Determine the NIC Port Type for the Specific BCM957608 NIC

This can be either QSFP112-DD or QSFP112 as shown in Table 12.

Table 12: NIC Port Types

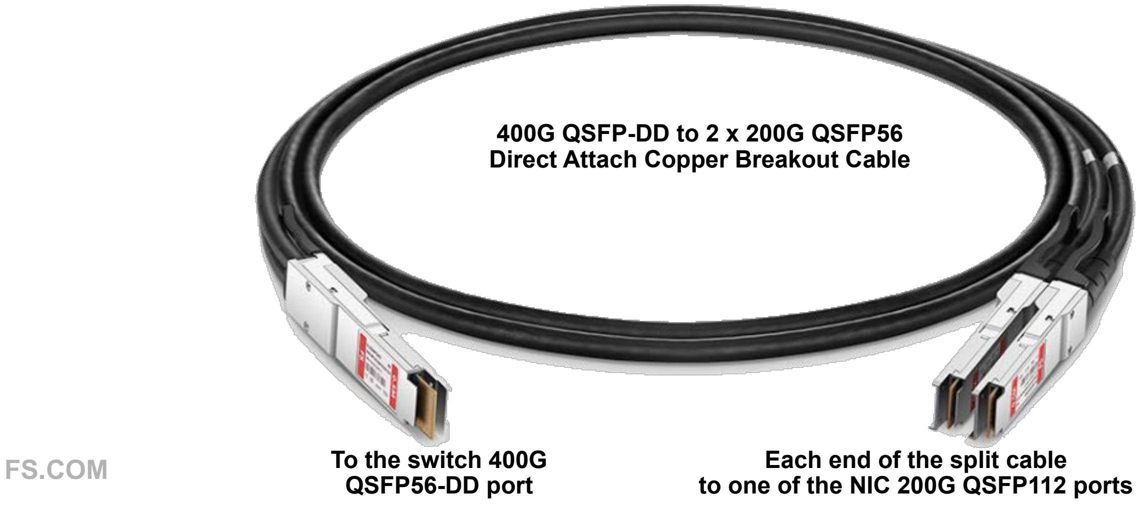
NIC	Cage Connector
P2200G2	QSFP112
N2200G2	QSFP112
P1400G2	QSFP112
N1400G2D	QSFP112-DD

B.3 Determine the Cable Connector Type (for DAC)

For DAC, determine the cable that satisfies the connector type on both ends of the cable (see Table 13).

Table 13: Example Cable Selection

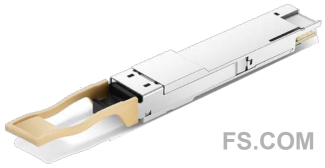
Switch	NIC	Example Required Cable
Dell Z9664F-ON (400G using QSFP56-DD cage)	P2200G2 (2x200G using QSFP112 cage, compatible with QSFP56)	Generic Compatible 400G QSFP-DD to 2x200G QSFP56 Passive Direct Attach Copper Breakout Cable


400G QSFP-DD to 2 x 200G QSFP56
Direct Attach Copper Breakout Cable
FS.COM
To the switch 400G
QSFP56-DD port
Each end of the split cable
to one of the NIC 200G QSFP112 ports

B.4 Transceivers and Optical Cables

1. Choose a transceiver for the switch and NIC ports based on Determine the Switch Port Type and Port Speed and Determine the NIC Port Type for the Specific BCM957608 NIC.
2. Make sure the transceivers use the same center wavelength.
3. Make sure the transceivers satisfy the required range (SR, DR, and so forth).
4. Make sure the transceivers use the same optical cable connector ports.
5. Select optical cable to connect the transceivers:
   a. The cable must support the desired range (SR, DR, and so forth).
   b. The cable must support the connectors to match the transceivers (for example, LC, MTP, and so forth).
   c. The required cable connector will need to support the polish style to match the transceivers (for example, APC or UPC).

Table 14: Example Transceivers and Optical Cables

Switch	**NIC**	**Example required transceivers and optical cable**
Dell Z9664F-ON (400G using QSFP56-DD cage)	P2200G2 (2x200G using QSFP112 cage, compatible with QSFP56)	**Switch Side:** Dell Compatible 400GBASE-SR8 QSFP-DD 850 nm 100 m MTP/MPO-16 APC MMF Optical Transceiver Module  **NIC Side:** *Generic 200GBASE-SR4 QSFP56 850nm 100m MTP/MPO-12 MMF Optical Transceiver Module* **Example Optical Cable:** *MTP-16 to 2 x MTP-8 OM4 Multimode Conversion Harness Cable, 16 Fibers, Magenta*

Appendix C: Interconnect Compatibility List (ICL)

The Interconnect Compatibility List (ICL) is located on ESP and can be downloaded using the following link: BRCM57608 Cable ICL.

Revision History

957608-AN102; June 27, 2024

Updated:

- Removed Broadcom Confidential footer.

957608-AN101; June 18, 2024

Updated:

- Appendix C, Interconnect Compatibility List (ICL) – Updated link to ESP.

95760X-AN100; December 26, 2023

Initial release.

Broadcom reserves the right to make changes without further notice to any products or data herein to improve reliability, function, or design. Information furnished by Broadcom is believed to be accurate and reliable. However, Broadcom does not assume any liability arising out of the application or use of this information, nor the application or use of any product or circuit described herein, neither does it convey any license under its patent rights nor the rights of others.