Spectrum-X 组网设计及软件平台

Feb. 28, 2025

Agenda

- Spectrum-X网络架构设计及BOM
- 应用于Spectrum-X的LinkX 互联组件
- Spectrum-X软件平台 – Cumulus

NVIDIA Spectrum-X网络架构设计及BOM

UPGR_BOM-XL_SPINE

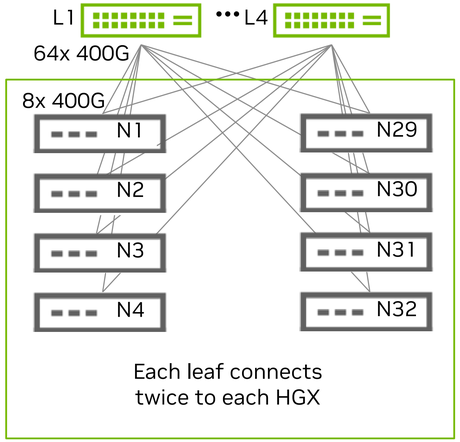
Requirements and Guidelines

- Spectrum-X is optimized for AI cloud use case
- Spectrum-X is focused on GPU-GPU communication – BW and latency
  - Then, we optimize for power, cost and operations
- E/W requires full bisectional BW for GPU-GPU communication
- Different GPUs and servers may have different network topology
- We separate the architecture into 3 different tiers
  - Up to 8K GPUs – two tiers
  - Greater than 8K GPUs – three tiers
  - Exception - 128 GPUs or less – single tier – covered by Spectrum-4
- In order to determine the tier: We ask what is the max scale expected for the cluster over its life-cycle?
  - We would provision the upper tier of the network to accommodate for future growth w/o re-cabling
- N/S with Spectrum-4
  - The majority of the traffic is from storage to compute and vise-versa

E/W – The SU as the basic building block

With Rail-optimized connectivity

- Spectrum-X architecture is modular and is based on Scalable Unit – SU
- A SU is
  - A group of 32 HGX servers connected in a rail-optimized fabric
    - Total of 256 Hopper GPUs
    - Same amount of servers/GPUs as SuperPOD RA
  - Provides a single-hop access across this group of servers
  - Each server has two ports of 400G connected to each leaf switch serving the SU
    - 2 GPU-rails are running through the same switch
- Servers and leaf switches are connected with optics,
  - As a single switch needs to reach many servers – racks
- Allows to localize big amount of the job and reduce load to the rest of the network

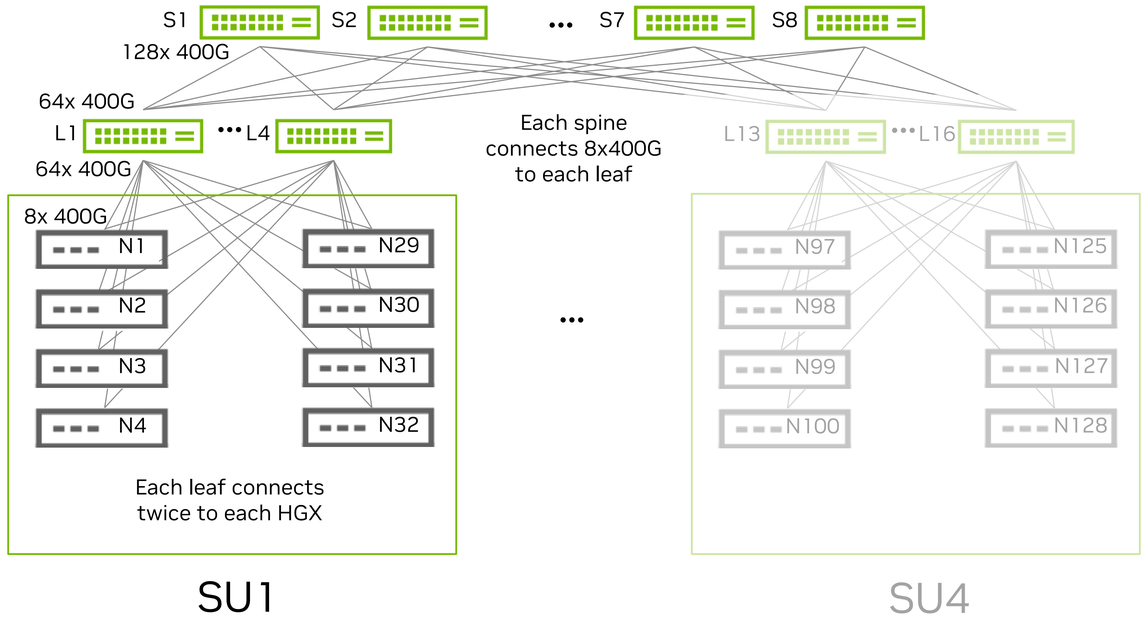


SU

E/W – Two Tiers

Up to 8K Hopper GPUs or 16K L20 GPUs

- Consider the number of GPUs expect to be deployed
- Determine the number of spines to be used:
  - #GPUs <= #total_spine_ports, and
  - #spines should be power-of-2 – 2, 4, 8, etc…
- Example: for 2500 GPUs => 2500/128 ~= 20 => the next power-of-2 is 32
- Then – connect SUs and their respective leaf switches as needed for the capacity requirement
  - And add as-you-grow
  - Leaf switches are deployed per SU count
- While we provision the spines, we don’t need to provision the optics
  - No major additional cost/power is needed

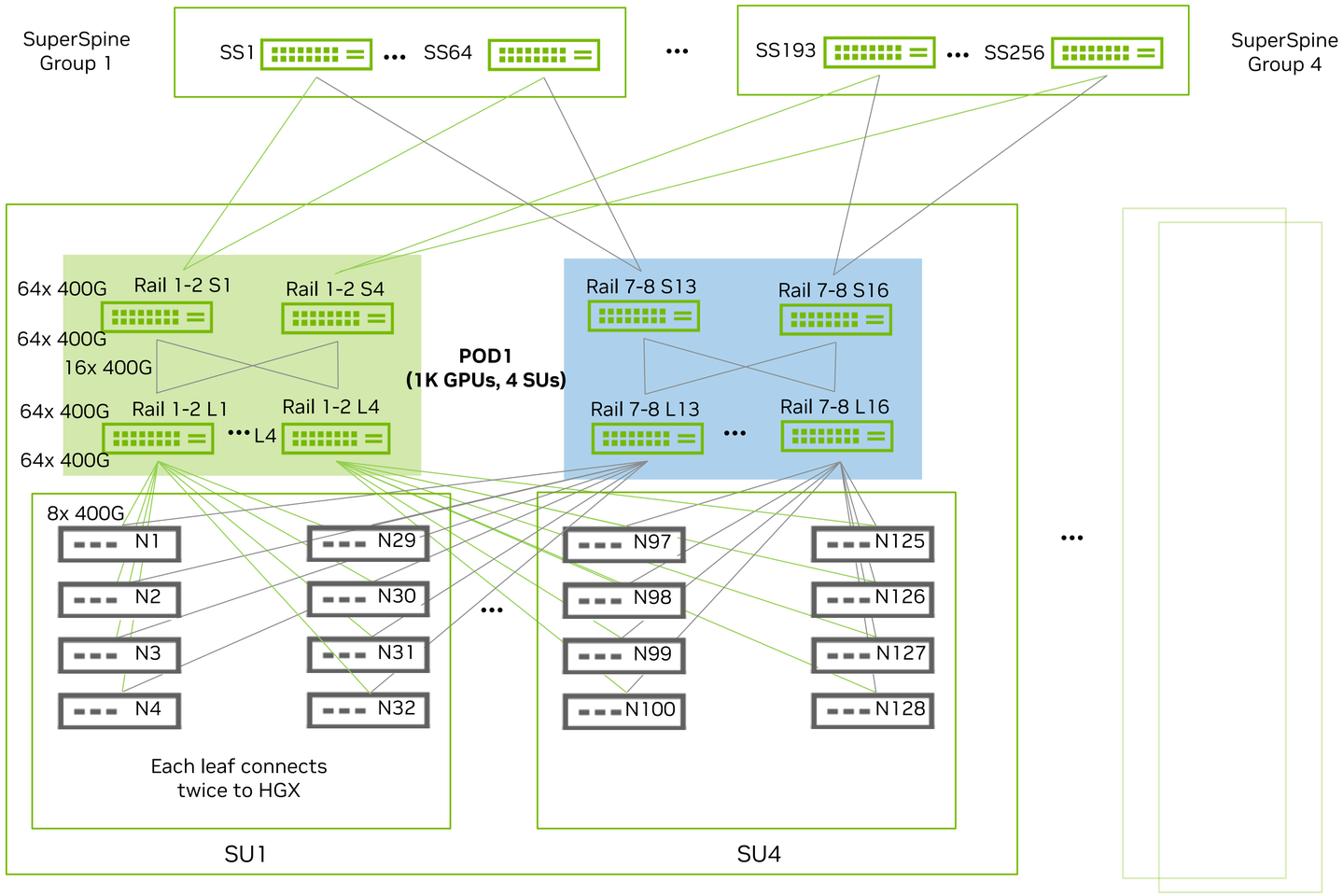


Networking Configuration Tools: https://www.nvidia.com/en-us/networking/configuration-tools/

E/W – Three Tiers

More than 8K Hopper GPUs or 16K L20 GPUs

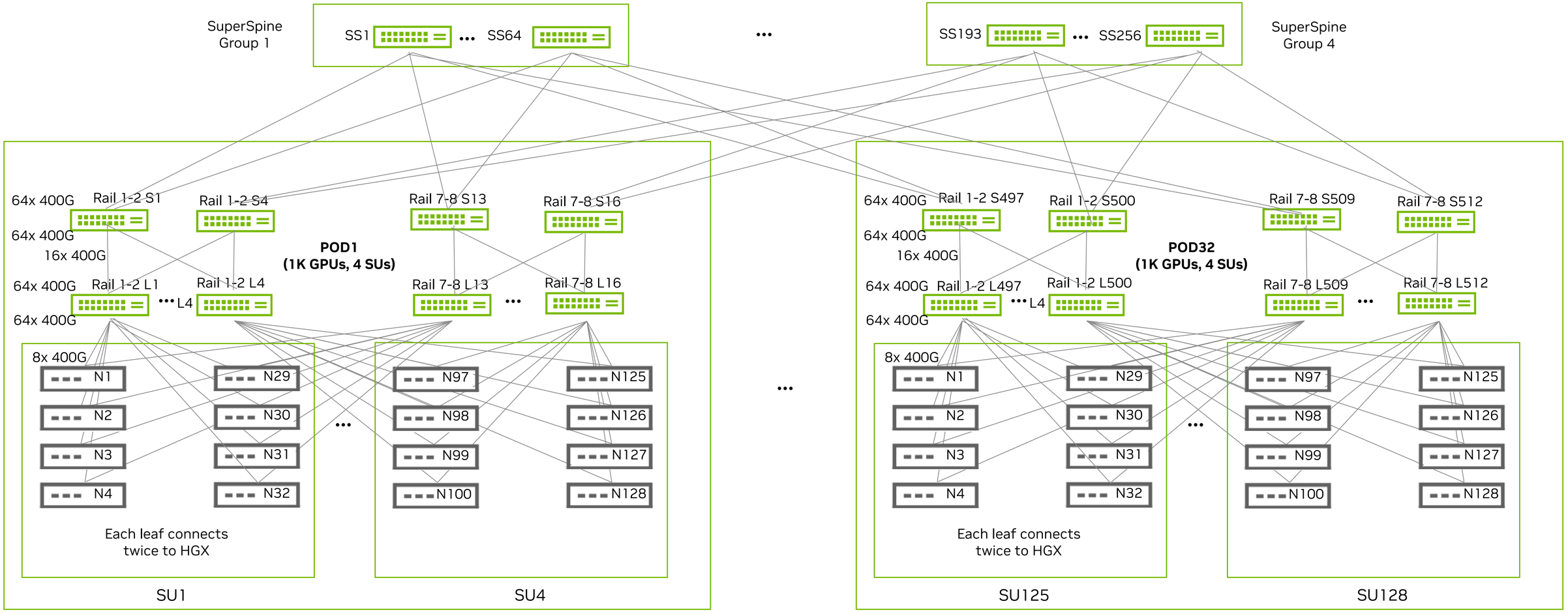
- More flexible and modular design
  - Can adjust to various constraints and scale requirements
  - Often it will adjust to a facility and its space/power constraints
- SUs are first connected using per-rail spine groups
  - Forms rail ”blocks”
- SUs are now aggregated into PODs
  - PODs are non-blocking and data-plane self-contained
  - Add PODs as-you-grow
- PODs are connected using Super Spine groups
  - Rule of thumb - in the amount of spines per rail block
  - All should be deployed at first phase
- Variation: cross-rails within PODs



E/W – Three Tiers

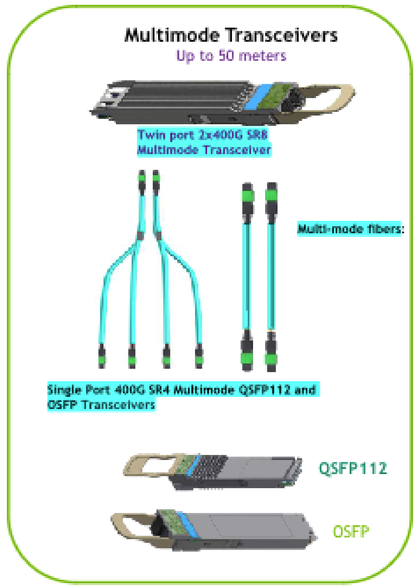
More than 8K Hopper GPUs or 16K L20 GPUs

- Full example of a 32K Hopper GPUs



Interconnect

- Spectrum-X RAs are focused on using optics for operation reasons
- Servers' density and Rail-optimized topology are driving EOR/MOR design
  - Leaf switches are not positioned as Top-of-Rack with their respective servers
- At switch
  - Twin-port OSFP 2xSR4 400G for up to 50m
  - Twin-port OSFP 2xDR4 (or FR4 if needed) 400G for longer reach
- At SuperNIC
  - QSFP112 SR4 400G (mostly)
- Copper for cost and power optimizations can be considered
  - Mostly at the leaf-spine layer
  - A guide is under construction



Connecting Different Speeds to SN5600

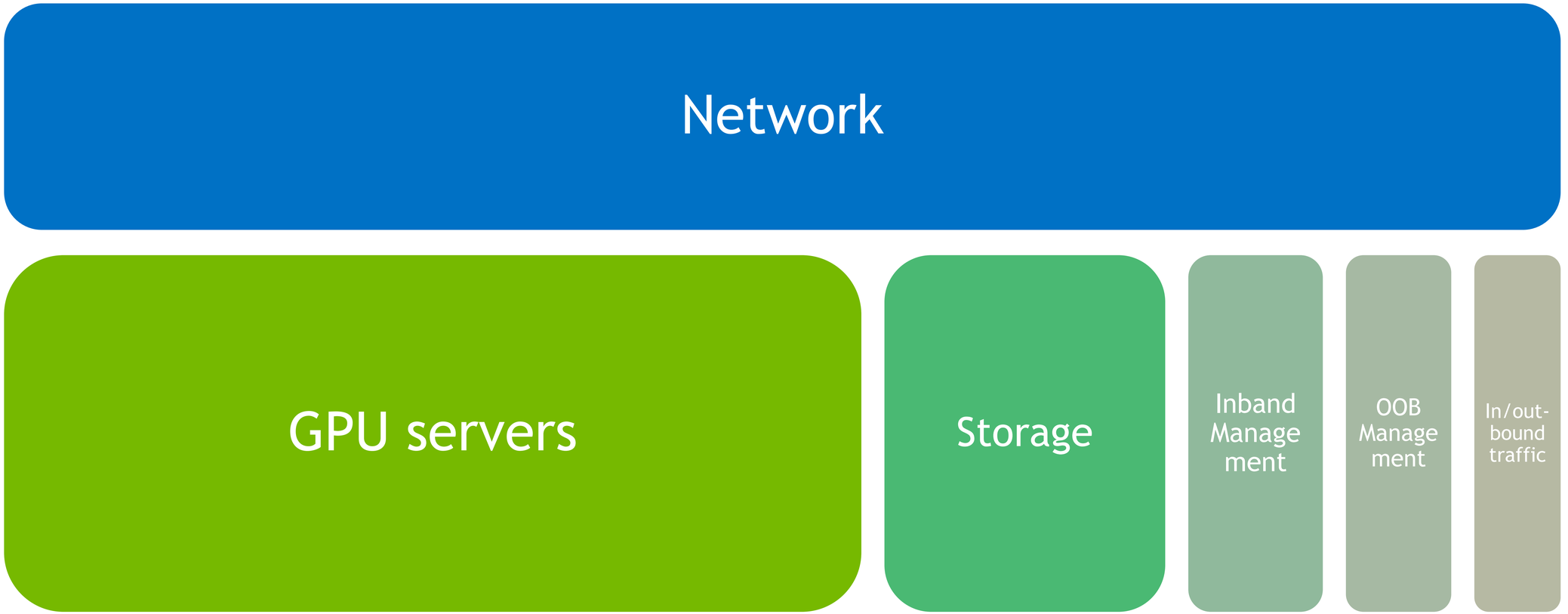
- 400G (4 x 100G)
  - 128 x 400G
  - SN5600 is ideal for 400G connectivity
  - We have the "NDR" LinkX portfolio

- 200G (2 x 100G)
  - 256 x 200G
  - SN5600 is idea for 200G connectivity
  - We leverage the "NDR" LinkX portfolio

- 200G (4 x 50G)
  - 128 x 200G

* as of now we are not using 800G as there is no real need for it

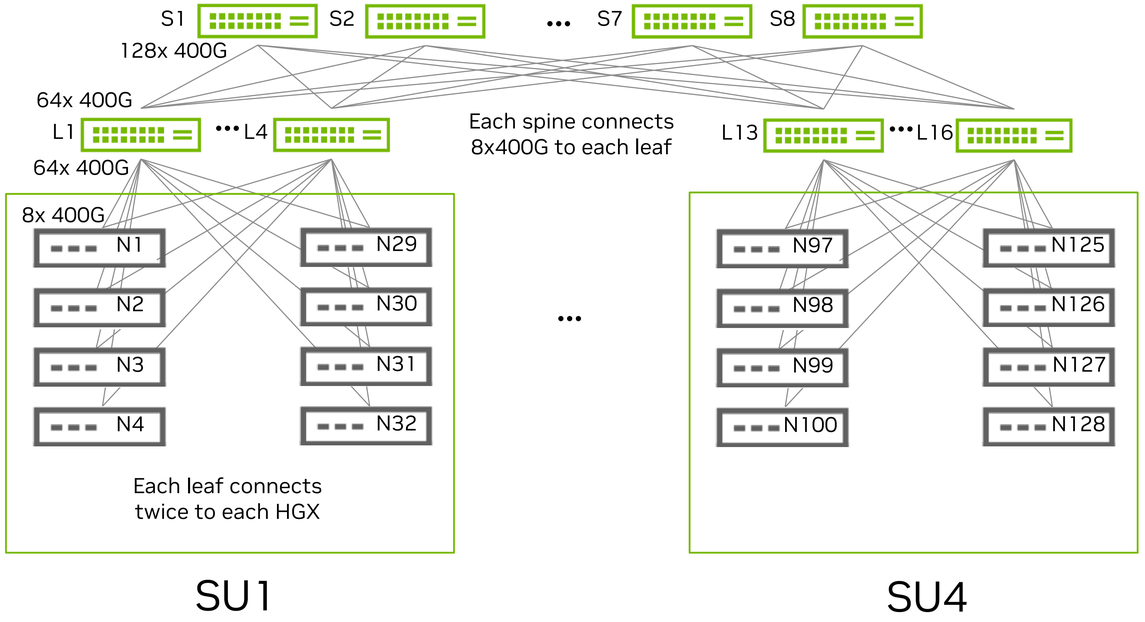
Let’s talk N/S…


Network
GPU servers
Storage
Inband Management
OOB Management
In/out-bound traffic

Spectrum-X BOM example

Spectrum-X Rail Optimized Leaf and Spine

Spectrum-4 + BlueField-3 GPU-to-GPU Fabric for Hopper GPU Cloud (1K GPUs)

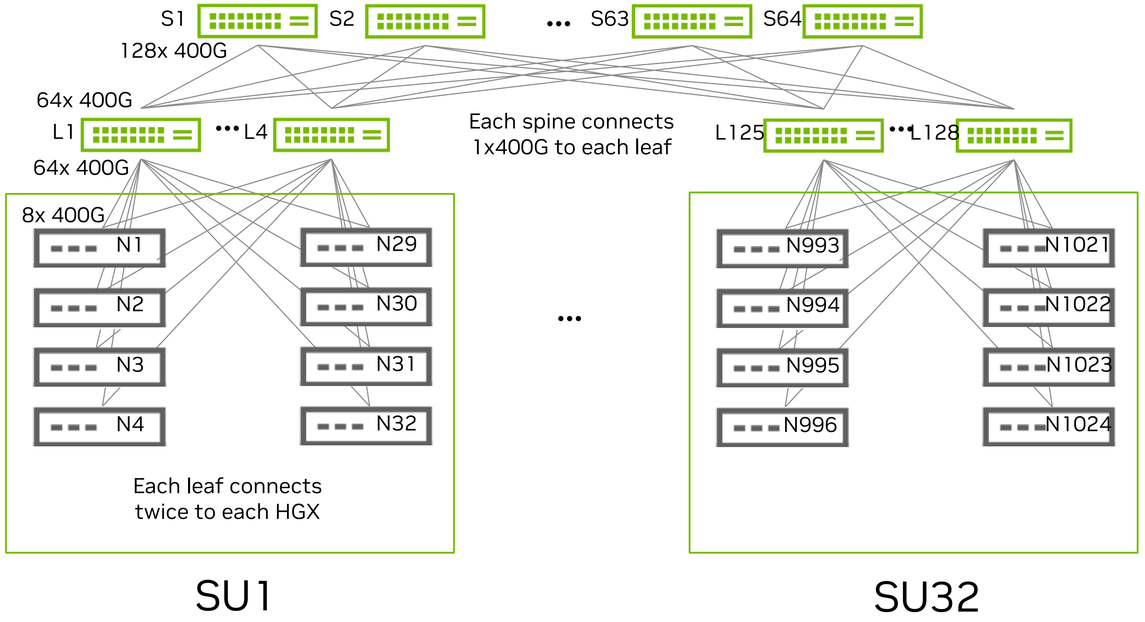


For 1K GPUs

Item	Part Number	Count
SN5600 Ethernet Switches (Leafs)	920-9N42F-00RI-7C0	16
SN5600 Ethernet Switches (Spines)	920-9N42F-00RI-7C0	8
HW & Cumulus Linux Support	781-C5640Z+P3CMI36	24
NetQ (SW & Support)	797-XNQ10Z+P3CMI36	24
OSFP Transceivers on Switches	MMA4Z00-NS	1536
QSFP112 Transceivers on HGX	MMA1Z00-NS400	1024
HGX Node-Leaf Cables	MFP7E10-N030	1024
Leaf-Spine Cables	MFP7E10-N030	1024

Spectrum-X Rail Optimized Leaf and Spine

Spectrum-4 + BlueField-3 GPU-to-GPU Fabric for Hopper GPU Cloud (8K GPUs)

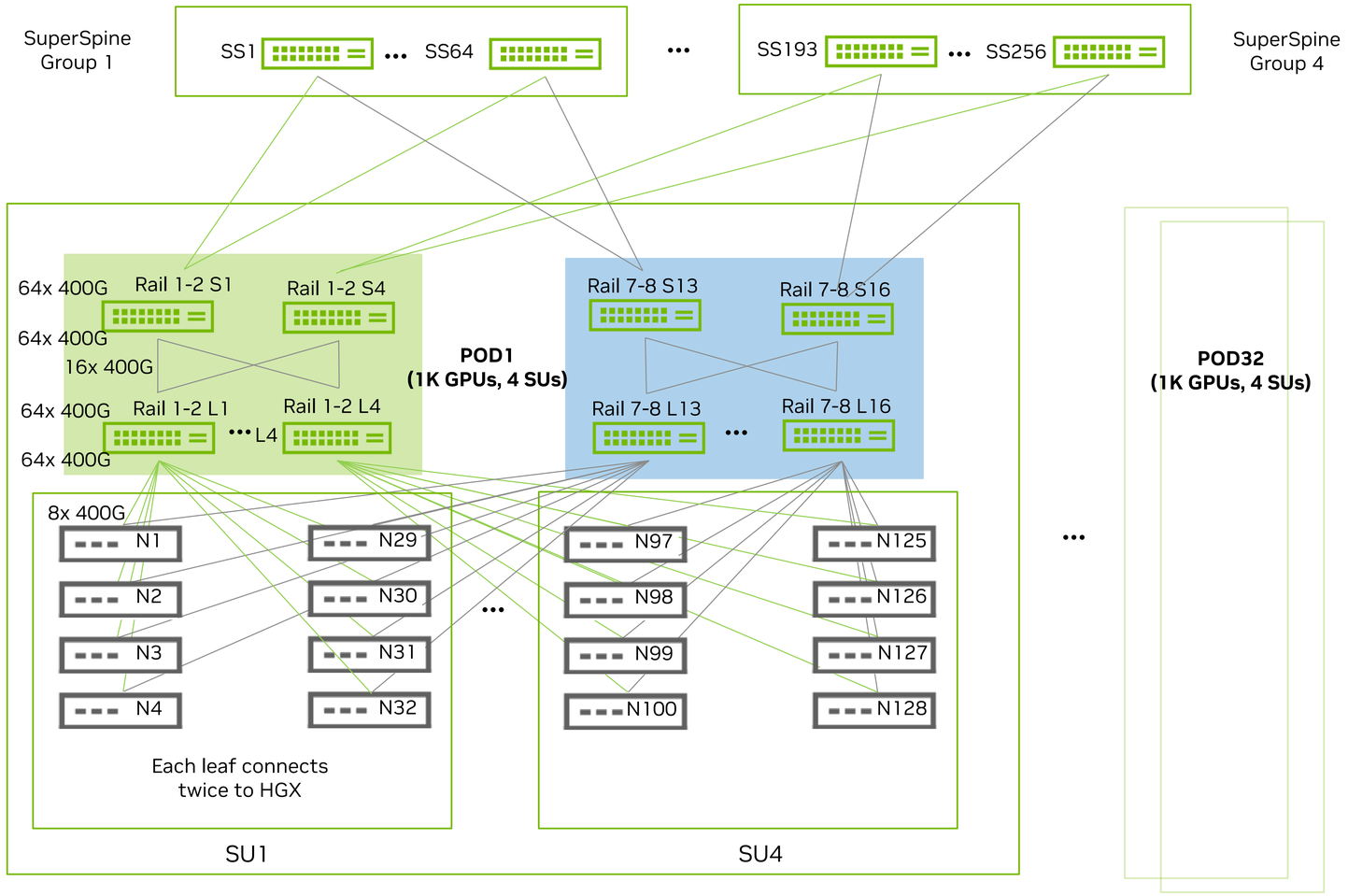


For 8K GPUs

Item	Part Number	Count
SN5600 Ethernet Switches (Leafs)	920-9N42F-00RI-7C0	128
SN5600 Ethernet Switches (Spines)	920-9N42F-00RI-7C0	64
HW & Cumulus Linux Support	781-C5640Z+P3CMI36	192
NetQ (SW & Support)	797-XNQ10Z+P3CMI36	192
OSFP Transceivers on Switches	MMA4Z00-NS	12288
QSFP112 Transceivers on HGX	MMA1Z00-NS400	8192
HGX Node-Leaf Cables	MFP7E10-N030	8192
Leaf-Spine Cables	MFP7E10-N030	8192

Spectrum-X Rail Optimized Leaf and Spine

Spectrum-4 + BlueField-3 GPU-to-GPU Fabric for Hopper GPU Cloud (32K GPUs)

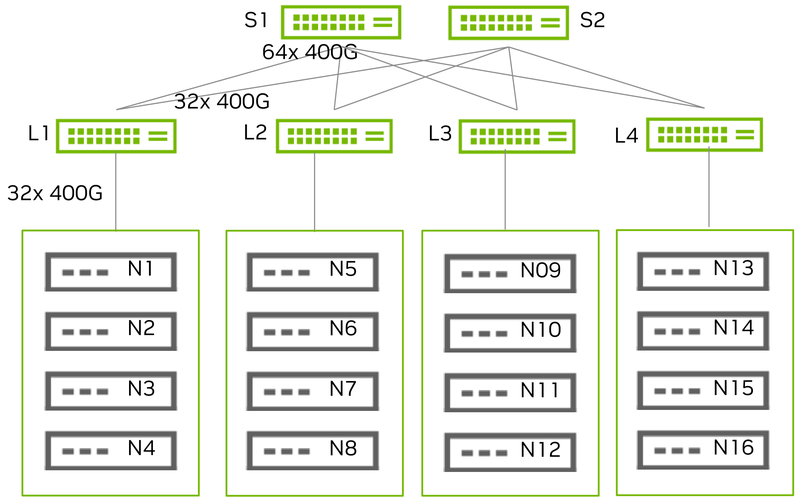


For 32K GPUs (128 SUs)

Item	Part Number	Count
SN5600 Ethernet Switches (Leafs)	920-9N42F-00RI-7C0	512
SN5600 Ethernet Switches (Spines)	920-9N42F-00RI-7C0	512
SN5600 Ethernet Switches (SuperSpines)	920-9N42F-00RI-7C0	256
HW & Cumulus Linux Support	781-C5640Z+P3CMI36	1280
NetQ (SW & Support)	797-XNQ10Z+P3CMI36	1280
OSFP Transceivers on Switches	MMA4Z00-NS	81920
QSFP112 Transceivers on HGX	MMA1Z00-NS400	32768
HGX Node-Leaf Cables	MFP7E10-N030	32768
Leaf-Spine-SuperSpine Cables	MFP7E10-N030	65536

SPC-X POC Topology & BOM

16 Nodes (128 GPUs) - 4 Leafs and 2 Spines



Not Rail Optimized

For 128 GPUs

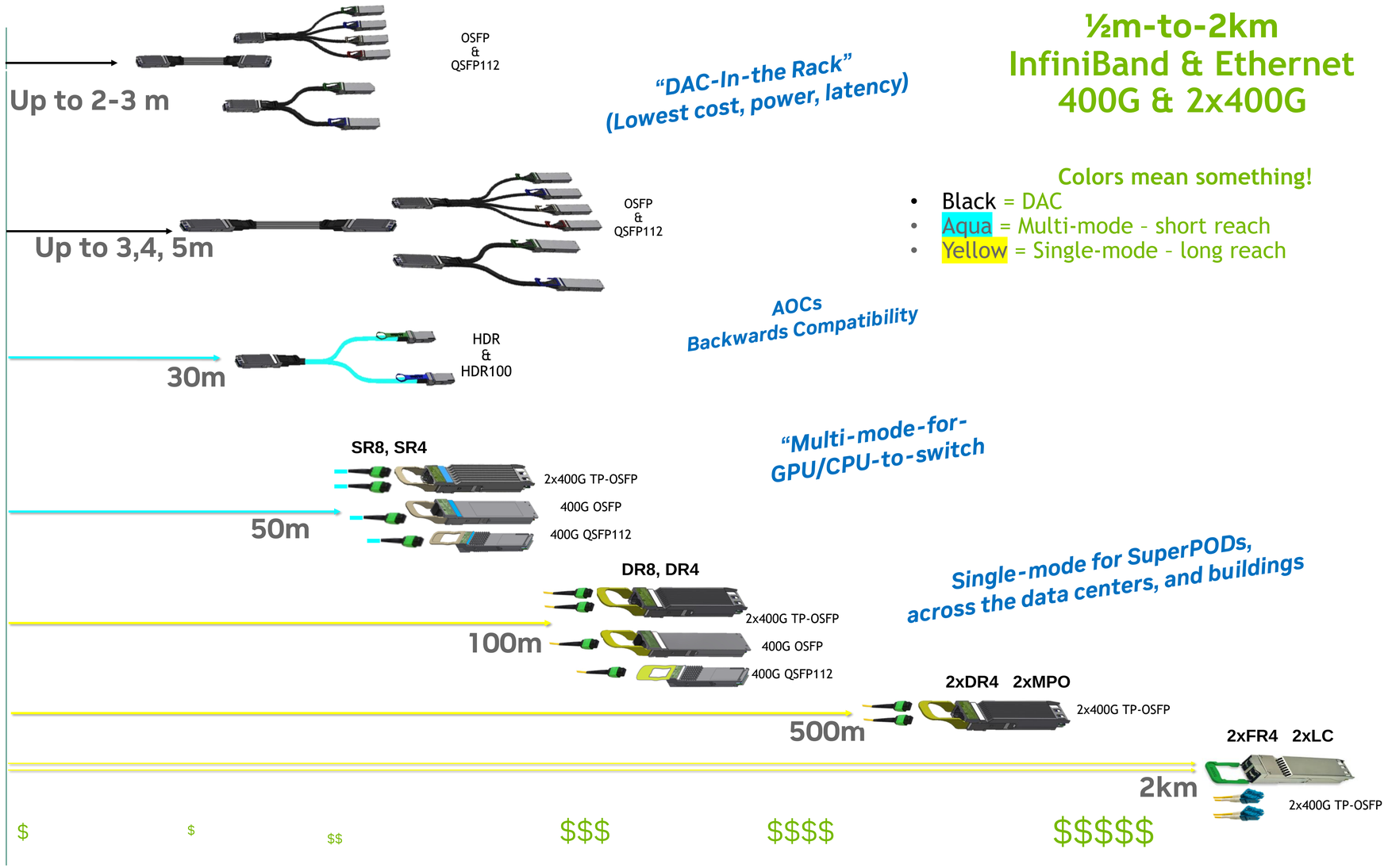
Item	Part Number	Count
SN5600 Ethernet Switches (Leafs)	920-9N42F-00RI-7C0	4
SN5600 Ethernet Switches (Spines)	920-9N42F-00RI-7C0	2
HW & Cumulus Linux Support	781-C5640Z+P3CMI36	6
NetQ (SW & Support)	797-XNQ10Z+P3CMI36	6
OSFP Transceivers on Switches	MMA4Z00-NS	192
QSFP112 Transceivers on HGX	MMA1Z00-NS400	128
HGX Node-Leaf Cables	MFP7E10-N030	128
Leaf-Spine Cables	MFP7E10-N030	128

应用于Spectrum-X的LinkX 互联组件

Lowest Cost Technologies for Each Reach Needed

Based on 100G-PAM4 Signaling

- **DAC** Cables
- ***Linear Active* Copper Cables** (L-ACC)
  (Based on low-power pre-emphasis ICs, not DSPs)
- **Active Optical Cables** (AOC)
- **Multi-Mode** Transceivers
- **Single-Mode** Transceivers



"DAC-In-the Rack"
(Lowest cost, power, latency)

½m-to-2km
InfiniBand & Ethernet
400G & 2x400G

AOCs
Backwards Compatibility

"Multi-mode-for-GPU/CPU-to-switch

Single-mode for SuperPODs, across the data centers, and buildings

Colors mean something!

- Black = DAC
- Aqua = Multi-mode - short reach
- Yellow = Single-mode - long reach

Price Increases as Reach Extends

LinkX 200GbE/HDR Portfolio

DACs, AOCs, Multimode + Single mode Transceivers

Based on 50G-PAM4 Modulation in QSFP56

200G QSFP56 –to- 200G QSFP56

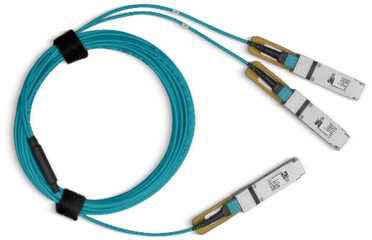
InfiniBand + Ethernet

DACs

AOCs

200G QSFP56 –to- 2x 100G QSFP56

InfiniBand + Ethernet



25G SFP28 Pass thru to QSFP28 Cage

Ethernet only

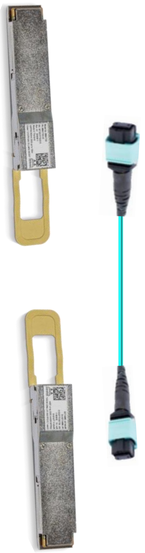
QSA28

SFP28-to-QSFP28 port adapter

200GbE SR4

850nm Multi-mode 100-meters QSFP56

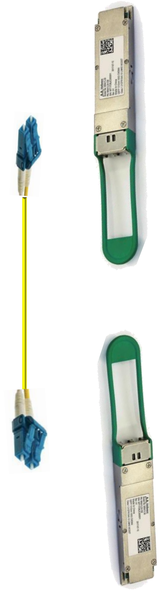
InfiniBand + Ethernet



200GbE FR4

1310nm Single-mode 2km QSFP56

InfiniBand + Ethernet



The Ethernet DACs part numbers were merged with the InfiniBand DACs and are using the InfiniBand part numbers for both IB and EN.
InfiniBand has a lower (better) BER rating than Ethernet (worse), so InfiniBand can be used for both. E
N uses FEC, IB does not or only a lite FEC.
So, InfiniBand parts will work in NVIDIA Ethernet systems, BUT Ethernet parts will not work for InfiniBand systems because of the BER and FEC issues.

Separate 200GbE and HDR portfolio part numbers merged to one set of HDR/200GbE part number.

LinkX 400GbE Portfolio

DACs, ACCs, AOCs, Multimode + Single mode Transceivers

Based on 100G-PAM4 Modulation in OSFP + QSFP112

Same part number for both InfiniBand & Ethernet

New: *Ethernet-only transceivers with new part numbers at reduced prices*

Passive Copper

Up to 3 meters

800G to 2x 400G or 4x 200G
OSFP or QSFP112

OSFP QSFP112

Active Copper

Up to 5 meters

800G to 2x 400G or 4x 200G
OSFP or QSFP112

OSFP QSFP112

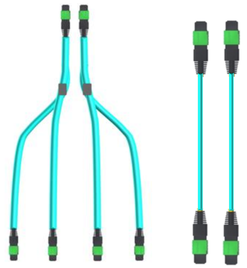
Backwards Compatible OSFP-to-QSFP56

Up to 30-meters Up to 2 meters

Multimode Transceivers

Up to 50 meters

Twin port 2x400G SR8 Multimode Transceiver



Multi-mode fibers:

Single Port 400G SR4 Multimode QSFP112 and OSFP Transceivers

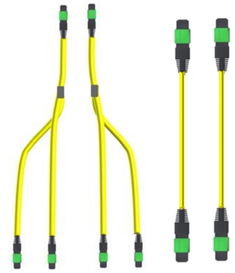
QSFP112

OSFP

Single mode Transceivers

Up to 100, 500 & 2K meters

Twin port 2x400G DR8 Single mode Transceiver



Single-mode fibers:

Single Port 400G DR4 Single mode QSFP112 and OSFP Transceivers

QSFP112

OSFP

LinkX 400GbE/NDR Portfolio

Based on 100G-PAM4 Modulation in OSFP and QSFP112

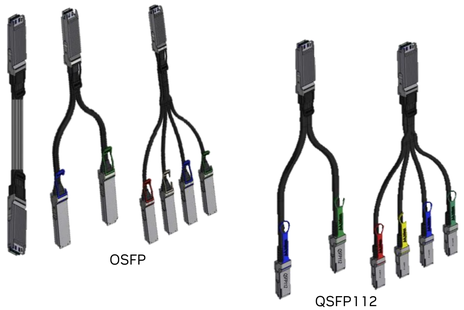
NEW! Additional Ethernet-only transceivers available with lower prices and separate part numbers (-T in OPN)

Ethernet
Spectrum-4 400GbE SN5600
64-cage Twin-port-OSFP

LinkX

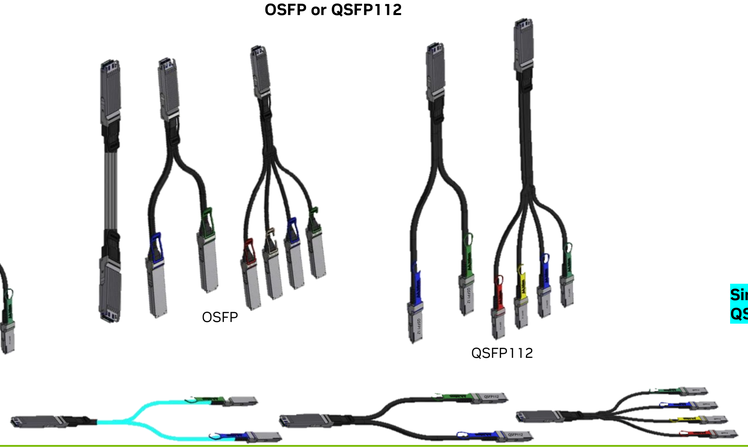
Passive Copper
Up to 3 meters

800G to 2x 400G or 4x 200G OSFP or QSFP112



Active Copper
Up to 5 meters

800G to 2x 400G or 4x 200G OSFP or QSFP112



OPTICS

Up to 50 meters

Twin port 2x400G SR8 Multimode Transceiver

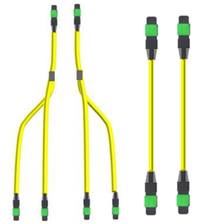
Up to 500 meters

Twin port 2x400G DR8 Single mode Transceiver

Multi-mode fibers:

Single Port 400G SR4 Multimode QSFP112 and OSFP Transceivers

Single-mode fibers:



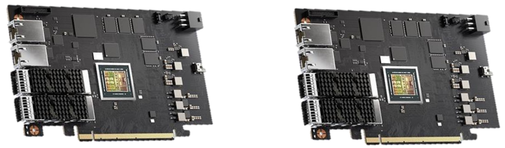
Single Port 400G DR4 Single mode QSFP112 and OSFP Transceivers

ConnectX-7

ConnectX-7/QSFP112 or OSFP 200G & 400G

BlueField-3

BlueField-3/QSFP112 200G & 400G



TWIN-PORT, OPEN-FINNED TOP (IHS) OSFP: OCTAL SMALL FORM-FACTOR PLUG

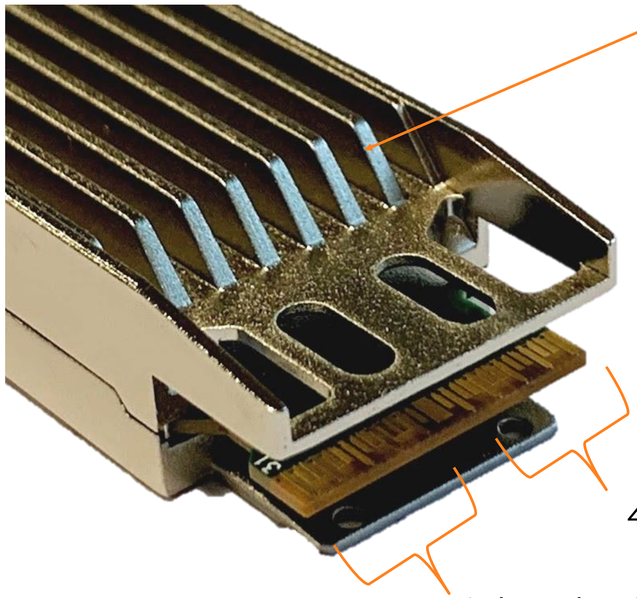
Only used in air-cooled Quantum-2 and SN5600 Ethernet switches

IHS = Integrated Heat Sink =Finned top
RHS =Riding Heat Sink = Flat top

2x 400G (4x100G-PAM4) links

13mm height

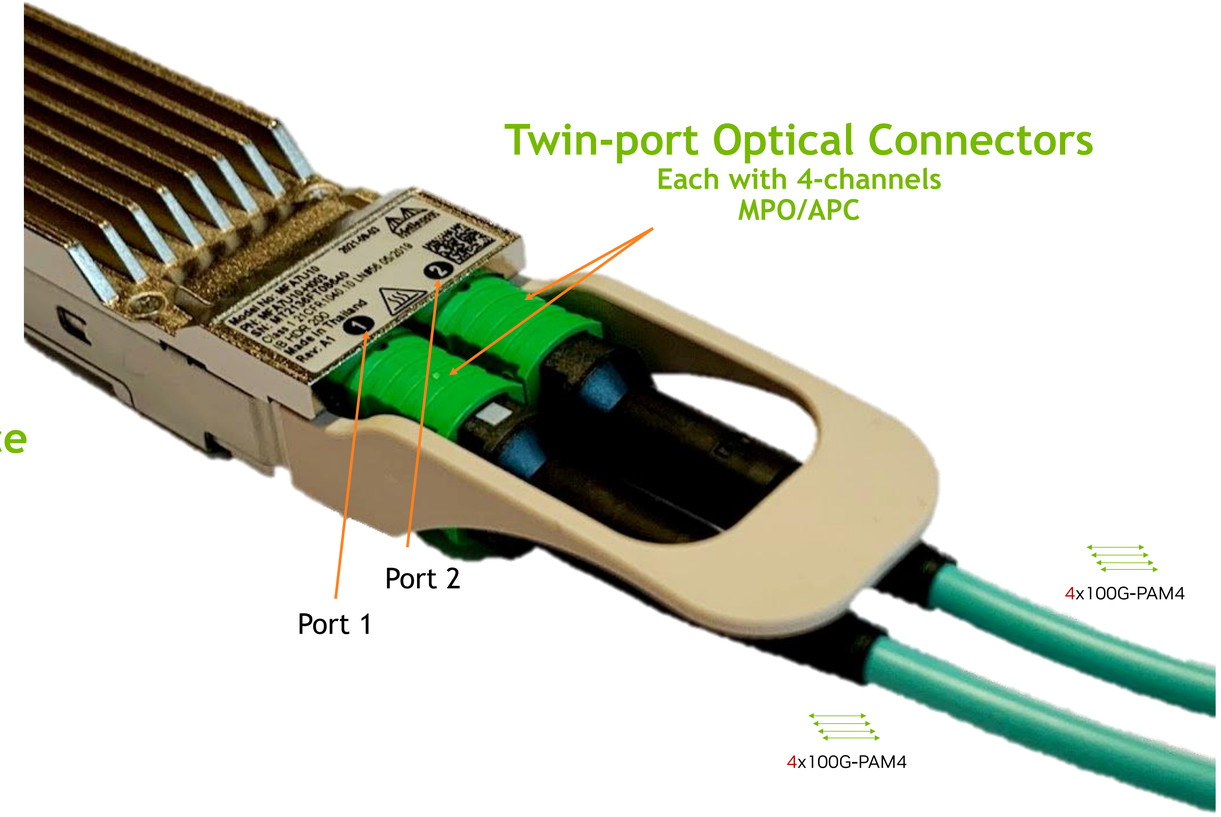
Integrated
Heat
Sink



Twin-port Device
Electrical end

4-channels x 100G-PAM4

4-channels x 100G-PAM4



Twin-port Optical Connectors
Each with 4-channels
MPO/APC

Port 2

Port 1

4x100G-PAM4

4x100G-PAM4

NVIDIA

400G Switches

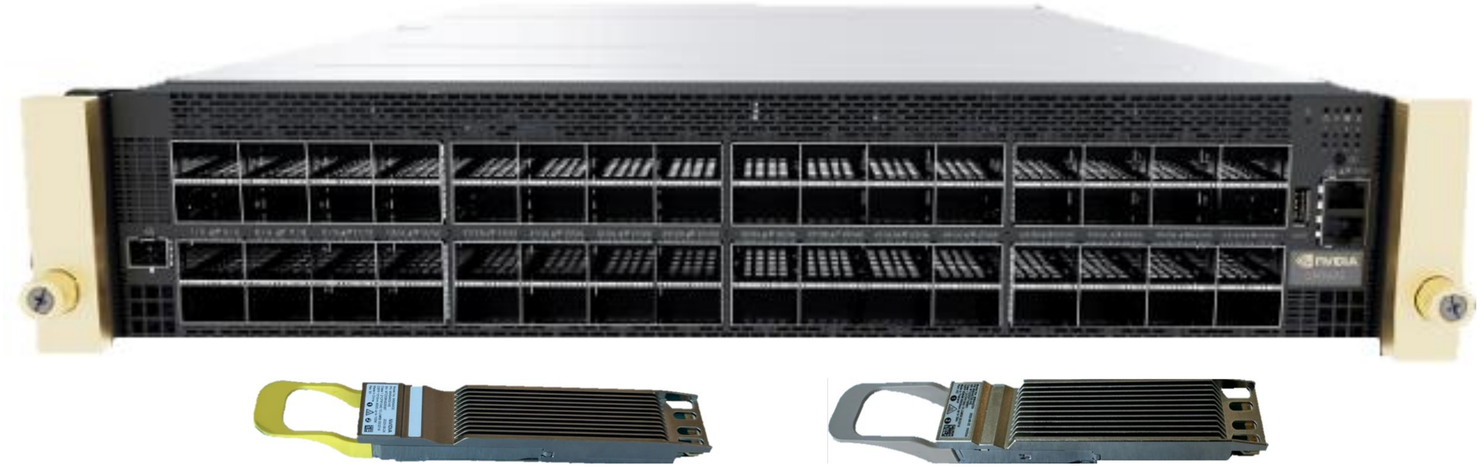
Based on 100G-PAM4 Signaling

SN5600 51.2Tb Ethernet Switch

64-socket, 2RU Chassis

128-ports of **400GbE** (4x100G-PAM4)

256-ports of 200GbE (2x100G-PAM4)



ONLY switches use IHS Finned-Top Transceivers

400GbE/NDR; Each Switch Cage Can Operate AT Different SPEEDS

Each Twin-port can be configured individually

DAC or ACC Cables, Multi-mode or Single-mode Transceivers.

At 400GbE, 200GbE and 100GbE end port speeds

Air-cooled switches accept Twin-port Finned-top devices only

DAC/QSFP56, 28

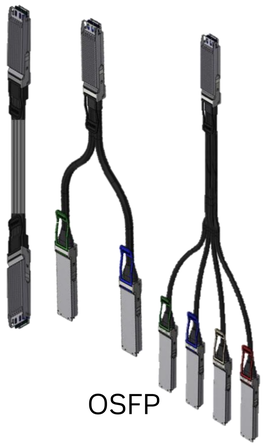
2x200G AOCs
2x100G AOCs

QSFP56
QSFP28

QSFP56

Backward Compatible

DAC/OSFP



OSFP

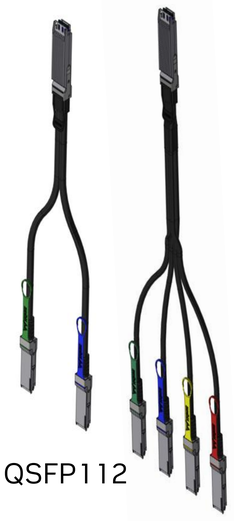
ACC/OSFP

OSFP

DAC/QSFP112

QSFP112

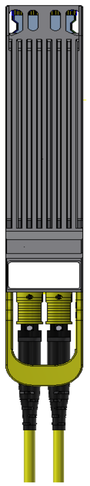
ACC/QSFP112



QSFP112

Multi-mode Transceiver

Single-mode Transceiver

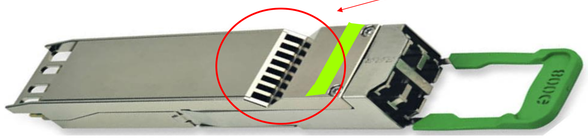


NVIDIA

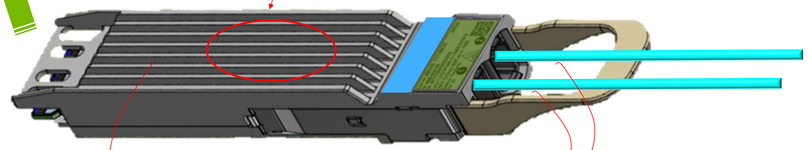
SN5600 ONLY ACCEPTS TWIN-PORT, IHS FINNED-TOP DEVICES

Why? Very few air-cooling holes in QSFP chassis, so fins are added on transceivers

Additional cooling fins integrated on top of transceivers



Closed-Finned top (IHS)
For 2km, 2XFR4 ONLY with dual LC optical connectors.
Used in Quantum2 and SN5600 switches only
with switches using reverse air flow configuration

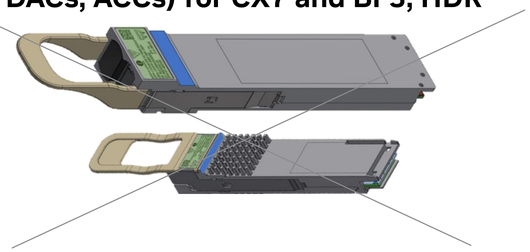


Twin-port (2x400Gb/s)

Open Finned Top (IHS)
(DACs, ACCs, Transceivers)

Finned-top devices are only used in switches
8-channel, 2x400G devices
(Tcvrs, DACs, ACCs)

Everything else uses a flat-top devices (Tcvrs, DACs, ACCs) for CX7 and BF3, HDR



All Twin-port devices for air-cooled switches use **Finned-Tops.**
DGX Hoppers Cedar7 GPU cages only accepts Twin-port devices --- with **flat tops** (-FLT in part number)

Switch cages and DGX cages have different OSFP heights and are not interchangeable (e.g. Flat will not work in Finned cage)
Air-cooled switches can accept open and closed Finned top devices as they are the same height

NEW: Reduced Prices for Ethernet-Only Transceivers

30-40% Lower Prices & New Part Numbers

MMS4X00-NS-T	980-9I30H-F4NM00	BAGHEERA	2x400G Twin-port OSFP Single mode Transceiver	For SN5600 Switch
MMA4Z00-NS-T	980-9I510-F4NS00	LOUIE	2x400G Twin-port OSFP Multimode Transceiver	For SN5600 Switch
MMA4Z00-NS400-T	980-9I51S-F4NS00	LOUIE400	400GbE OSFP Multimode Transceiver	For CX7/OSFP
MMA1Z00-NS400-T	980-9I693-F4NS00	QLOUIE400	400GbE QSFP112 Multimode Transceiver	For CX7/BF3 / QSFP112

800G-to-800G Switch-to-Switch & Adapter, DPU

Direct Attach Copper Cables (DAC)

"Direct Attach Copper" DAC cables have only a tiny configuration EEPROM IC and only copper wires.
Offers:

- Lowest latency 5-15ns
- Lowest power <0.1W
- Lowest cost
- Highest reliability

400Gb/s Spectrum-4 Ethernet
800G Twin-port-OSFP Switches

Ethernet 400GbE SN5600
64-cage Twin-port-OSFP

OSFP End Points
1,2 meters

Direct Attach Copper Cable
800G to 800G
OSFP to OSFP

MCP4Y10-N00A	(0.5m)	30 AWG
MCP4Y10-N001	(1m)	30 AWG
MCP4Y10-N002	(2m)	26 AWG

OSFP End Points
4-to-5 meters

Direct Attach Copper Cable
800G to 2x 400G
OSFP to 2xOSFP

MCP7Y00-N001	(1m)	30 AWG
MCP7Y00-N01A	(1.5m)	30 AWG
MCP7Y00-N002	(2m)	26 AWG
MCP7Y00-N02A	(2.5m)	26 AWG
MCP7Y00-N003	(3m)	26 AWG

Direct Attach Copper Cable
800G to 4x 200G
OSFP to 4xOSFP

MCP7Y50-N001	(1m)	30 AWG
MCP7Y50-N01A	(1.5m)	30 AWG
MCP7Y50-N002	(2m)	26 AWG
MCP7Y50-N02A	(2.5m)	26 AWG
MCP7Y50-N003	(3m)	26 AWG

400Gb/s Spectrum-4 Ethernet
800G Twin-port-OSFP Switches

Ethernet 400GbE SN5600
64-cage Twin-port-OSFP

400G/200G ConnectX-7
OSFP

QSFP112 End Points
4-to-5 meters

Direct Attach Copper Cable
800G to 2x 400G
OSFP to 2x QSFP112

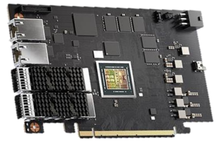
MCP7Y10-N001	(1m)	30AWG
MCP7Y10-N01A	(1.5m)	30AWG
MCP7Y10-N002	(2m)	26 AWG
MCP7Y10-N02A	(2.5m)	26 AWG
MCP7Y10-N003	(3m)	26 AWG

Direct Attach Copper Cable
800G to 4x 200G
OSFP to 4x QSFP112

MCP7Y40-N001	(1m)	30AWG
MCP7Y40-N01A	(1.5m)	30AWG
MCP7Y40-N002	(2m)	26 AWG
MCP7Y40-N02A	(2.5m)	26 AWG
MCP7Y40-N003	(3m)	26 AWG

400G/200G ConnectX-7
QSFP112

400G/200G BlueField-3
QSFP112



800G-to-800G Switch-to-Switch & Adapter, DPU

Linear - Active Copper Cables (LACC)

"Linear" Active DAC cables have a small pre-emphasis IC instead of a large, high power & latency DSP IC.

Offers:

- Low latency 10-20ns
- Low power <0.8-1.5Watts/end
- Lower cost (than optics)

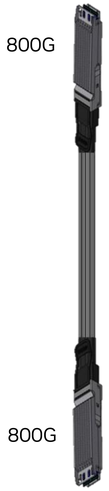
These are NOT Active Electric Cable that use a DSP IC in each end

400Gb/s Spectrum-4 Ethernet
800G Twin-port-OSFP Switches

Ethernet 400GbE SN5600
64-cage Twin-port-OSFP

OSFP End Points

3,4,5 meters



800G to 800G
Twin-port OSFP to Twin-port OSFP
MCA4J80-N003 (3m) 30 AWG
MCA4J80-N004 (4m) 30 AWG
MCA4J80-N005 (5m) 26 AWG
1.5 Watts/end

OSFP End Points

4-to-5 meters

800G to 2x 400G
OSFP to 2x OSFP
MCA7J60-N004 (4m) 30 AWG
MCA7J60-N005 (5m) 26 AWG
1.5 Watts and 0.8Watts

800G to 4x 200G
OSFP to 4x OSFP
MCA7J70-N004 (4m) 30 AWG
MCA7J70-N005 (5m) 26 AWG
1.5 and 0.4 Watts

QSFP112 End Points

4-to-5 meters

800G to 2x 400G
OSFP to 2x QSFP112
MCA7J65-N004 (4m) 30 AWG
MCA7J65-N005 (5m) 26 AWG
1.5 Watts and 0.8Watts

800G to 4x 200G
OSFP to 4x QSFP112
MCA7J75-N004 (4m) 30 AWG
MCA7J75-N005 (5m) 26 AWG
1.5 Watts and 0.4Watts

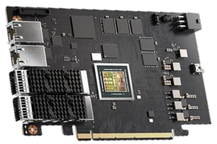
400Gb/s Spectrum-4 Ethernet
800G Twin-port-OSFP Switches

Ethernet 400GbE SN5600
64-cage Twin-port-OSFP

400G/200G ConnectX-7
OSFP

400G/200G ConnectX-7
QSFP112

400G/200G BlueField-3
QSFP112



800G-to-800G Switch-to-Switch & Adapter, DPU

Multimode Optics Up to 50-meters

SR4=Short Reach 4-channels (50m)

Multimode optics enables fiber reaches up to 50-meters and lower cost than Single mode optics

800G Electrical end
+
2x400G Optical end

400Gb/s Spectrum-4 Ethernet
800G Twin-port-OSFP Switches

Ethernet 400GbE SN5600
64-cage Twin-port-OSFP

Twin port OSFP 2x400G 2xSR4
Multimode Transceiver
MMA4Z00-NS-T (50m)
OSFP Finned-top
Dual MPO-12/APC
15 Watts
Louie

800G

400G

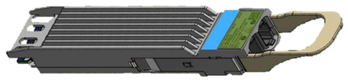
200G

400G

200G

800G-to-800G

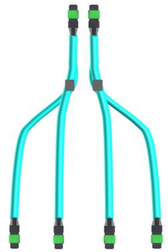
Multi-mode fibers:
MFP7E10-N0XX (XX = 03, 05, 07,10, 15, 20, 25, 30, 35, 40, 50) meters



800G-to-2x 400G

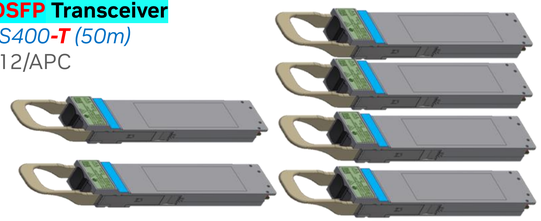
Multi-mode fibers:
MFP7E10-N0XX (XX = 03, 05, 07,10, 15, 20, 25, 30, 35, 40, 50) meters

800G-to-4x 200G



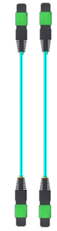
Multimode fibers Splitters
MFP7E20-N0XX (XX = 03, 05, 07, 10, 15, 20, 30, 40, 50) meters

Single port 400G SR4
Multimode OSFP Transceiver
MMA4Z00-NS400-T (50m)
Single MPO-12/APC
8 Watts

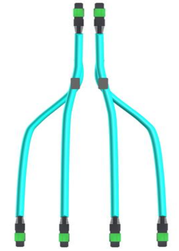


400G/200G ConnectX-7 OSFP

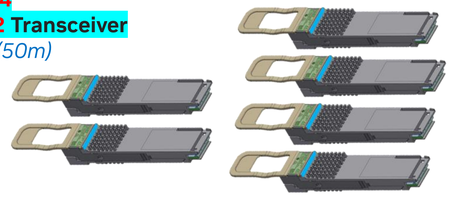
800G-to-2x 400G



800G-to-4x 200G



Single Port 400G SR4
Multimode QSFP112 Transceiver
MMA1Z00-NS400-T (50m)
Single MPO-12/APC
8 Watts

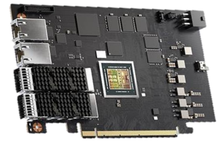


Both OSFP and QSFP112 transceivers can be mixed on split ends;

Use 400G transceivers for 200G links.
2 fiber on split ends creates 200G and reduces power

400G/200G ConnectX-7 QSFP112

400G/200G BlueField-3 QSFP112



400Gb/s Spectrum-4 Ethernet
800G Twin-port-OSFP Switches

Ethernet 400GbE SN5600
64-cage Twin-port-OSFP

Spectrum-X软件平台-Cumulus

NVIDIA SPECTRUM PLATFORM

Connecting NVIDIA Solutions With Accelerated Ethernet Switch Technologies

ACCELERATED
Best-in-class hardware performance with cloud-scale software efficiency

INNOVATIVE
5th generation in-house ASIC design optimizes Cloud, AI, & storage workloads

NVIDIA NETQ | NVIDIA AIR
NETWORK SOLUTION TOOLS

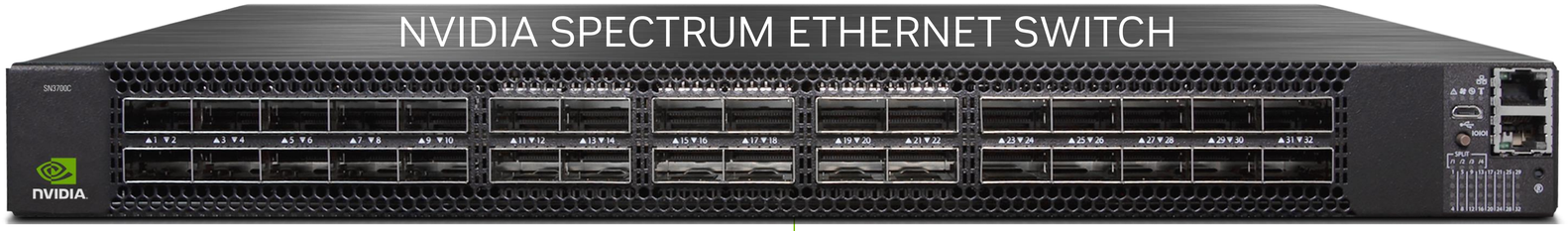
NVIDIA CUMULUS | PURE SONIC
NETWORK OPERATING SYSTEM

NVIDIA SPECTRUM ASIC
SWITCH ASIC

OPTIMIZED
Faster network deployments with lowest TCO and highest ROI

RELIABLE
Exclusive features enabling fairness, predictability and actionable visibility

NVIDIA SPECTRUM ETHERNET SWITCH

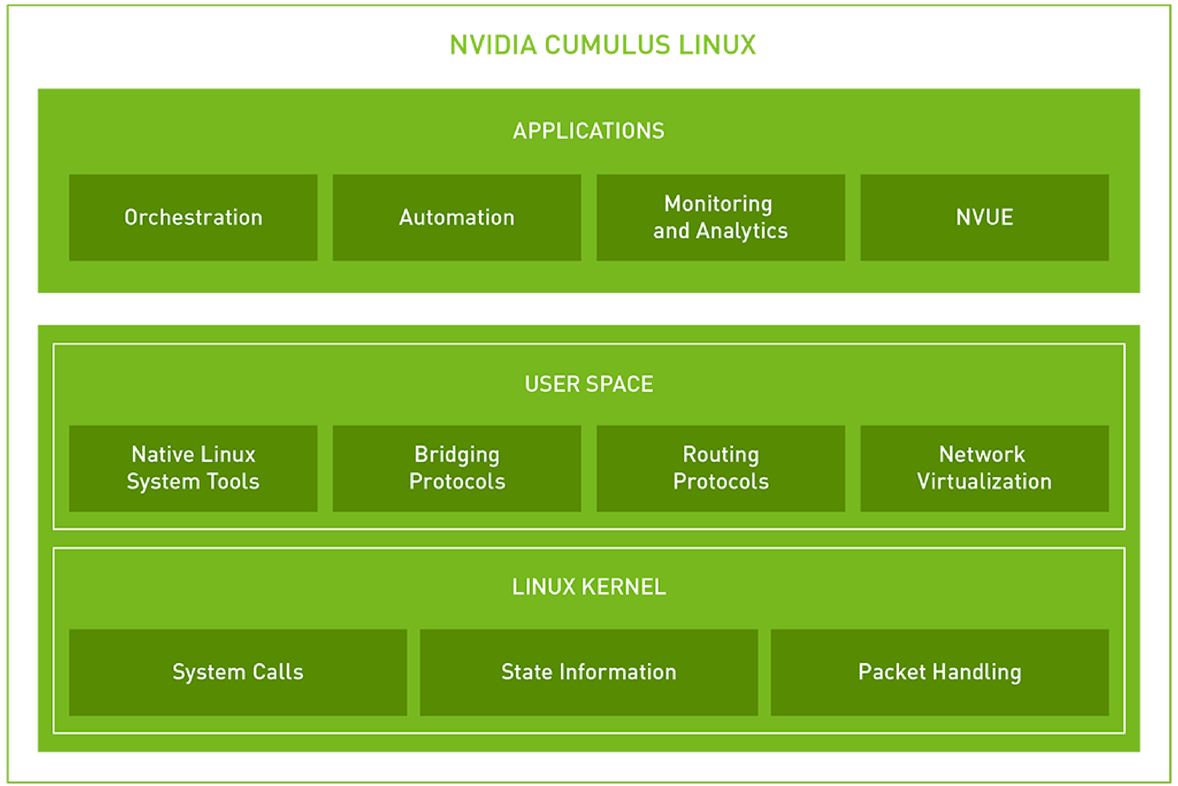


DGX | EGX | HGX | OVX

NVIDIA SPECTRUM ETHERNET NIC/DPU

NVIDIA Cumulus Linux

Flagship NOS for Spectrum Ethernet switches



Innovative open network operating system to automate, customize, and scale data center networks. The Cumulus operational model enables faster deployments and simpler management.

HIGHEST PERFORMANCE

Accelerated RoCE, Adaptive Routing, and more coupled with highest scale

SIMPLIFIED OPERATIONS

Using digital twin, built-in automation, and open and programmable APIs (NVUE)

POWERED BY NVIDIA

Optimized NOS for E2E NVIDIA data center solutions (tested & validated)

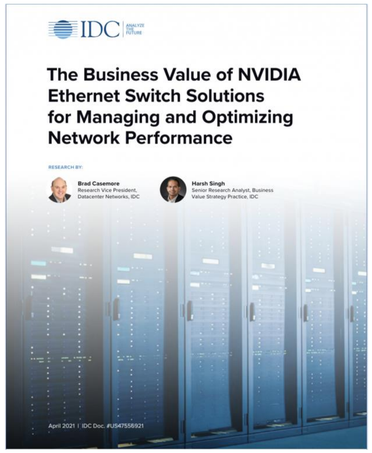
Cumulus integrates into full stack NVIDIA solutions (Aerial, DGX, EGX, OVX, GeForce NOW)

CUMULUS LINUX BUSINESS VALUE

PROGRAMMABLE INFRASTRUCTURE

FULLY FEATURED WITH NO LICENSE COST

CENTERED AROUND AUTOMATION



Staff Efficiencies Benefits

42% more efficient
Network management staff

36% less staff time
Spent “keeping the lights on”
120% more time innovating

Business and Operations Benefits

64% reduction
In unplanned downtime

$7.6 million additional
Revenue gained/protected

**IDC Business Value Whitepaper, Sponsored by NVIDIA, The Business Value of NVIDIA Ethernet Switch Solutions for Managing and Optimizing Network Performance, IDC Doc. #US47556921, April 2021*

NVIDIA SPECTRUM ETHERNET SWITCH PORTFOLIO

BUILT FOR SPEED AND SCALE FROM 1G TO 800G

NVIDIA OPTIMIZED

E2E solution validation & support
OVX / DGX / EGX
GeForce / NVIDIA AI Enterprise

SWITCH INNOVATION

5th Gen In-House ASIC Design
Intelligence in the hardware
Actionable Visibility
Unique Half-Width Form Factor

ACCELERATED DC SWITCHES

High Performance Ethernet
AI & Storage Optimized
At Cloud Scale

*"[NVIDIA Spectrum] switches allow our customers to support more apps, more users and more locations with **higher performance and lower latency while reducing cost, space and power**."*

-GM, Hewlett Packard Enterprise

Hewlett Packard Enterprise · Lenovo · NetApp · Microsoft Azure · VAST · Rakuten

Broad Ethernet Switch Portfolio

Breakout of Spectrum Ethernet Switches by Family

Spectrum ASIC	Spectrum-2 ASIC	Spectrum-3 ASIC	Spectrum-4 ASIC
1/10/25/40/50/100GbE 16-52 ports	1/10/25/40/50/100/200GbE 16-128 ports	1/10/25/40/50/100/200/400GbE 32-128 ports	10/25/40/50/100/200/400/800GbE 32-256 ports
SN2000	**SN3000**	**SN4000**	**SN5000**
 SN2100: 16x100G	 SN3420: 48x25G + 12x100G	 SN4410: 48x100G + 8x400G	 SN5400: 64x400G Q4'23
 SN2010: 18x25G + 4x100G	 SN3700-C: 32x100G	 SN4600-C: 64x100G	 SN5600: 64x800G
 SN2201: 48x1G + 4x100G	 SN3700-V: 32x200G	 SN4600-V: 64x200G	
	 SN3750-SX: 32x200G	 SN4700: 32x400G	

Cumulus Linux

Operational Innovation for Accelerated Data Centers

Simplicity	Performance	Flexibility
		
Built for Automation Data Center Digital Twins	Acceleration through AR, ROCE Highest Scale	NVUE Object Model Open-Source Innovation

Cumulus Linux

Built for Automation, Built for the Cloud

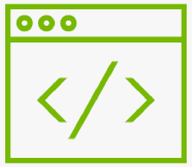
Based on decades of work at Google and Cisco

World's first native Linux Network Operating System

Leaders in Linux networking

- FRR, VRF, EVPN, ifupdown2, ONIE

Native automation, streaming telemetry

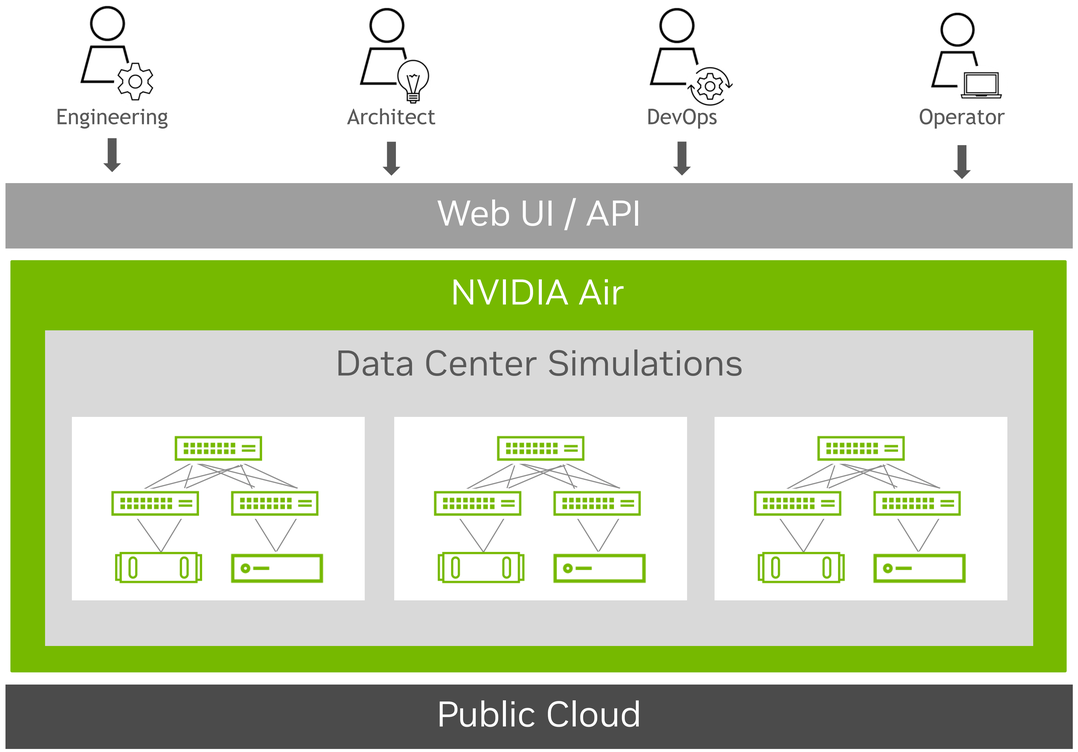


Simplified Configurations

Advanced data center simulation

NVIDIA Air

Hosted data center digital twins with end-to-end platform simulation (Network + GPU)



Create a full-size virtual data center at no cost with network & compute featuring pre-built and bring-your-own images. Validate solutions with security, automation, upgrades, failover, monitoring & interoperability.

FAST SPIN-UP

Create a digital twin in seconds, no HW required

HIGHEST ACCURACY

The most true-to-life emulation for ASICs & GPUs

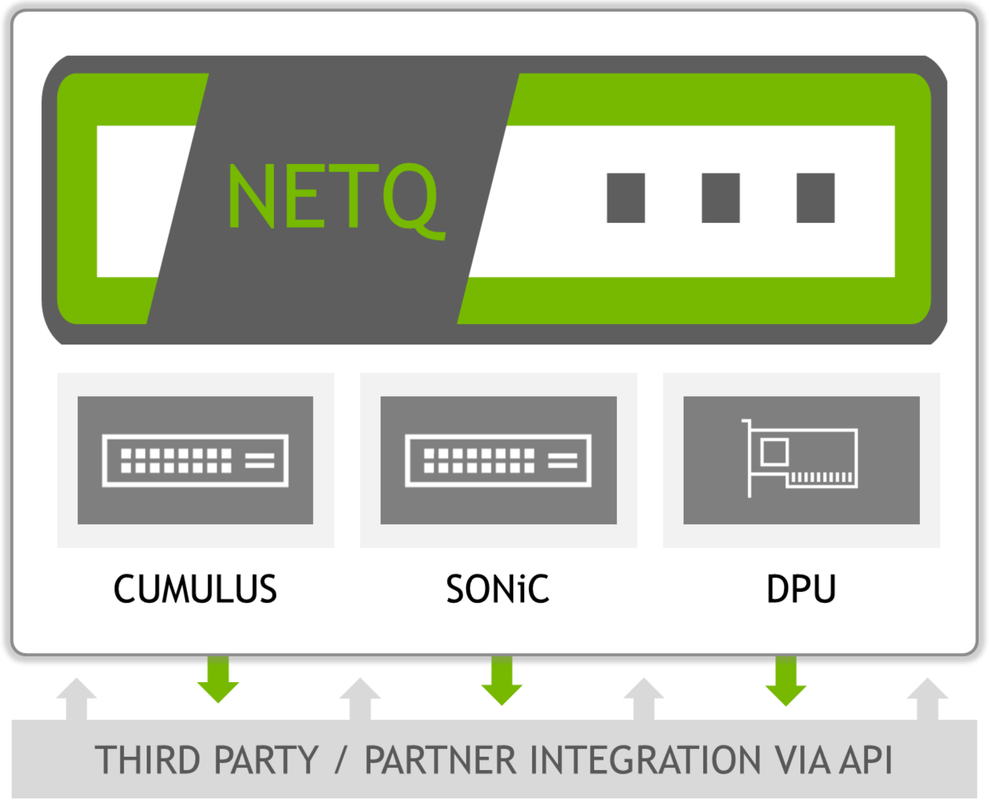
VALIDATE THEN DEPLOY

Deploy solutions with confidence to the physical data center

NVIDIA Air simulates full stack NVIDIA solutions (DGX, EGX, OVX)

Visibility and Validation for Streamlined Operations

NVIDIA NetQ – Switch & DPU



Using advanced telemetry, NetQ makes it easier to troubleshoot and automate network workflows in real time while reducing maintenance and downtime.

CENTRALIZED VISIBILITY

Monitor & trouble-shoot NVIDIA network: DPUs and switches (Cumulus and SONiC)

NETWORK HEALTH

Validations of network to identify and fix issues quickly

DATA AGGREGATION

Centrally gather & analyze data across entire network

NetQ integrates into full stack NVIDIA solutions (Aerial, DGX, EGX, OVX, Morpheus, HBN/DPU)

Features needed to accelerate AI

Spectrum Switches with Cumulus Ensure Superior Performance for AI Workloads

ADAPTIVE ROUTING

30% performance gains using E2E NVIDIA Adaptive Routing (Switch to NIC)

ACCELERATED RoCE

4x faster provisioning of AI clusters using Spectrum innovations

BEST-IN-CLASS LATENCY

Spectrum provides **10X** lower latency

CONGESTION CONTROL

VIP Congestion Control (VIP CC) through in-band telemetry

FAIRNESS & BUFFERS

Shared buffer delivers **fairness & 100% reduction** in buffer usage

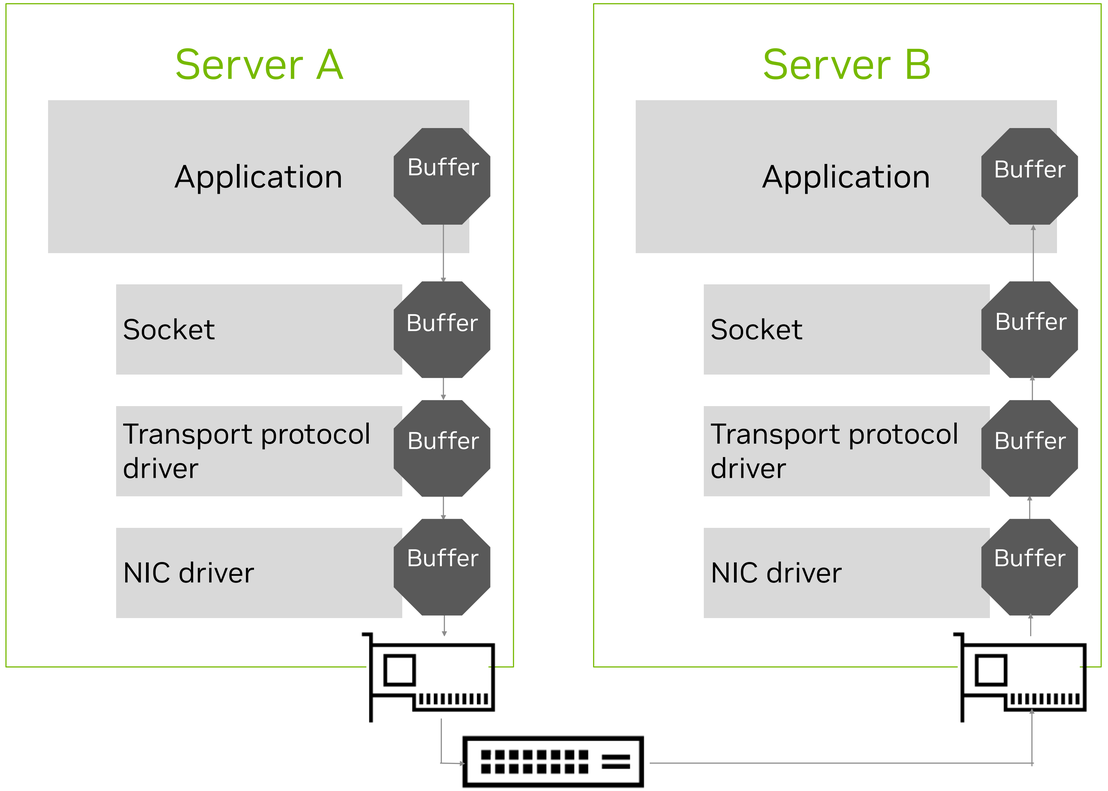
HIGH BANDWIDTH

Dense 200G for AI/ML training workloads

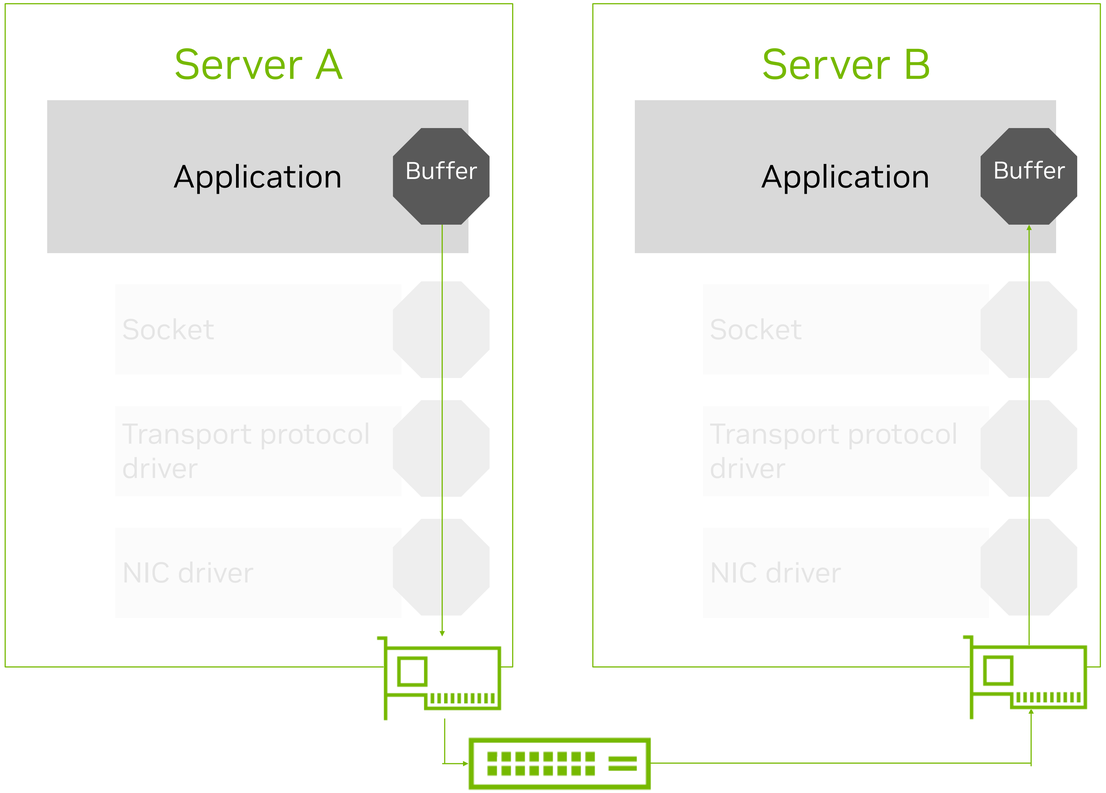
Speed up Application Performance with RoCE

NVIDIA Cumulus Linux Switches and NICs offer Accelerated RoCE Support

Problem: TCP requires multiple hops for application communication

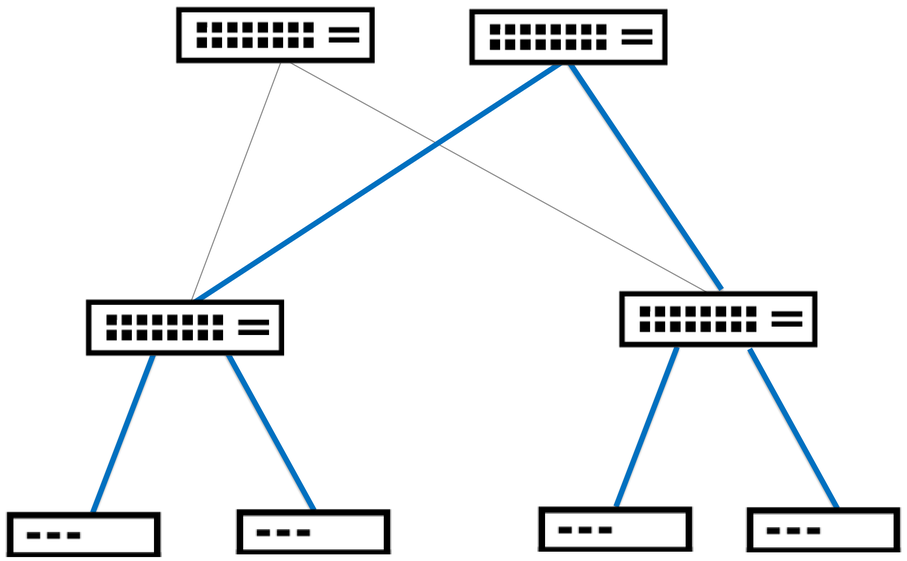


Solution: RoCE speeds up applications through remote direct memory access (RDMA)

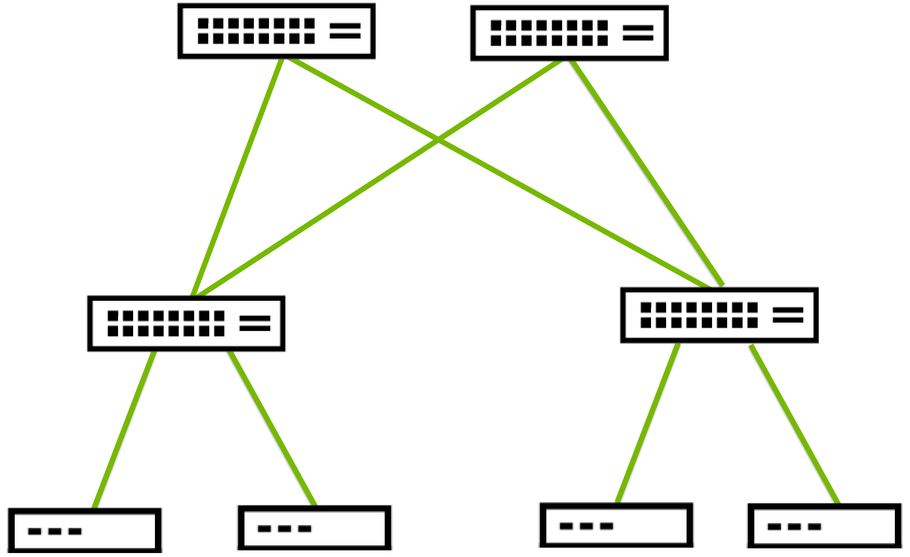


Adaptive Routing With Cumulus Linux

Supercomputing Network Innovation



50% bandwidth utilization on ISLs with non-blocking topology



Full bandwidth utilization throughout fabric

Traditional ECMP

- Static hashing
- Independent of traffic conditions
- Bigger flows = higher chance for congestion
- High tail latency

Adaptive Routing

- Congestion based port selection
  - Flowlet-aware: eliminates out-of-order Packets
  - Reduce tail latency
  - RoCE OOO placement for highest efficiency

Accelerated one touch roce

Complexity-free App Speedup for High Performance Workloads

do roce

Set up lossy or lossless RoCE with two lines of code:

```
cumulus@switch:mgmt:~$ nv set qos roce mode lossy
cumulus@switch:mgmt:~$ nv config apply
```

Enable RoCE, EVPN and other protocols via the NVIDIA User Experience (NVUE) object model

show roce

Verify RoCE buffers, utilization, bytes, packet counts, and more:

```
cumulus@switch:mgmt:~$ nv show qos roce
```

Spend less time tweaking RoCE parameters with NVIDIA Ansible NVUE Modules

check roce

Verify RoCE consistency and congestion settings across switches:

```
cumulus@switch:mgmt:~$ netq check roce
```

Validate your RoCE settings and gain fabric-level performance insights with NVIDIA NetQ

NVIDIA Cumulus Linux is the Flagship NOS for Accelerated Ethernet

Simplicity · Performance · Flexibility

HIGHEST PERFORMANCE

Optimized NOS for End-to-End NVIDIA Data Center solutions (tested & validated)

PRE-VALIDATED DEPLOYMENT

Digital twin, built-in automation, and open APIs for 95% faster deployments

FASTEST ROI

Lowest TCO with no-license NOS for 41% lower Opex and 360% 5-year ROI