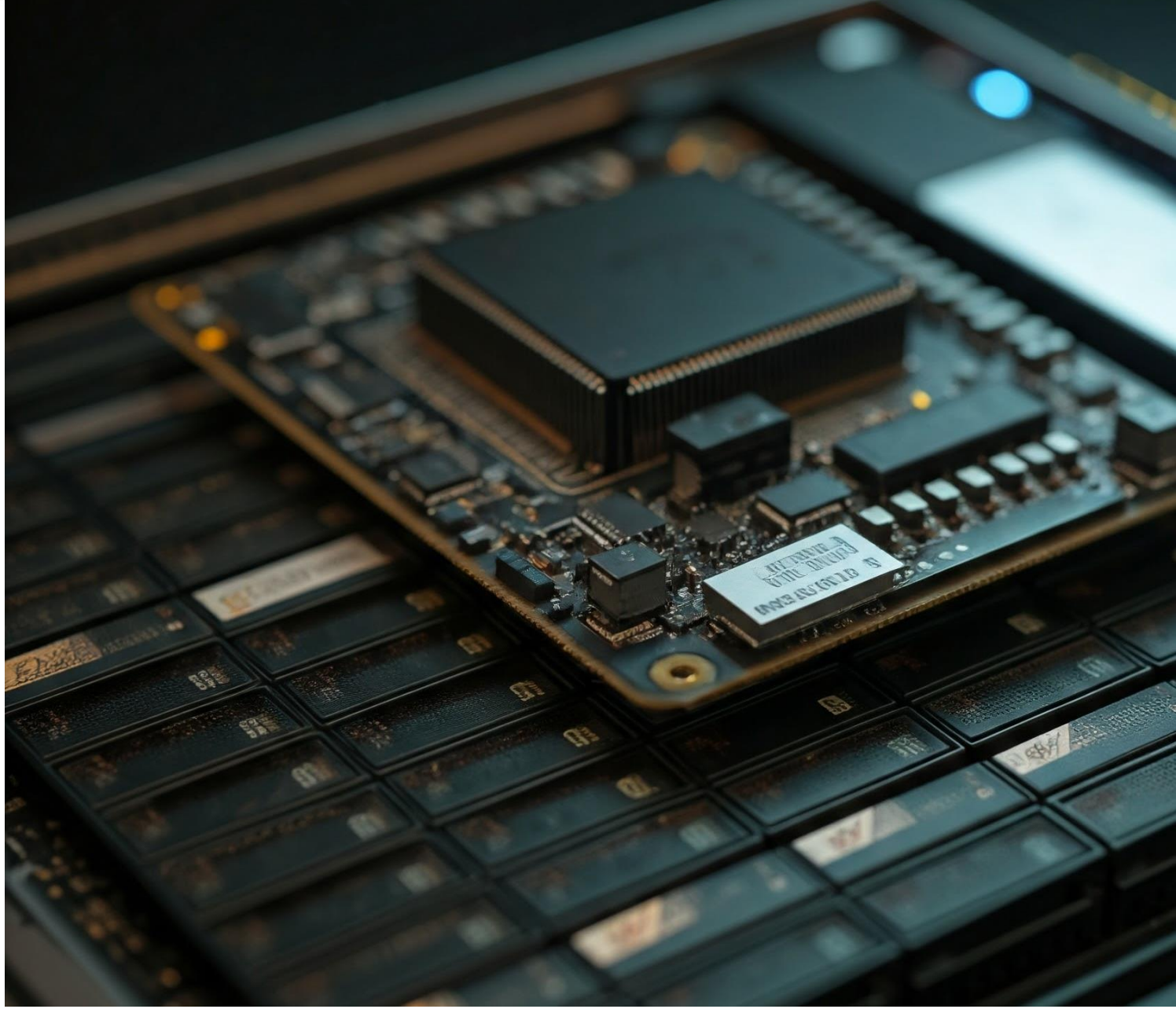
AI and Supercomputing Networks: Detailed Implementation Considerations for the New Challenges of Enhanced Ultra-low Latency Networks

Jonathan Jew
J&M Consultants, Inc
jew@j-and-m.com
Editor ANSI/TIA-942-C Data Center Standard
Chair BICSI Data Center Design Working Group
USTAG ISO/IEC JTC 1 SC25 WG 3 Project Lead for ISO/IEC 11801-5 Data Center Cabling

Artificial Intelligence (AI) Supercomputer servers use GPUs

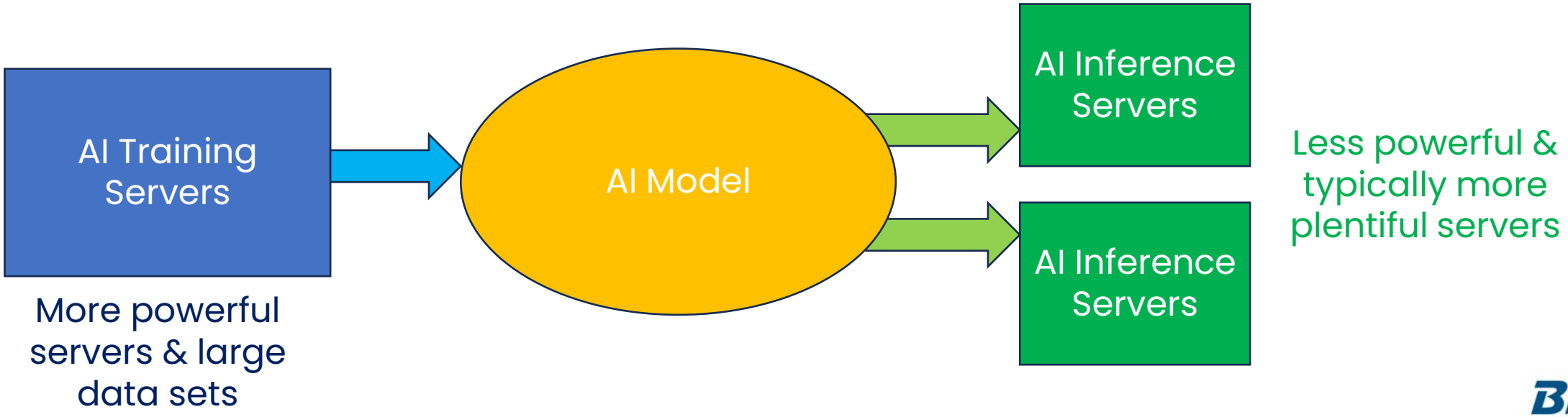
- AI computers are a type of supercomputer optimized for AI workloads
- Both general purpose supercomputers and AI supercomputers use GPUs to perform many calculations in parallel
- GPU processors include Nvidia, AMD, Intel and others
- AI supercomputer servers (called nodes) typically each have 1-2 CPUs and multiple GPUs connected to each other by an internal network



Bicsi CALA

AI Training vs AI Inference Servers

- AI Training supercomputing servers are used to teach an AI model to perform specific task. It requires powerful GPU servers and large amounts of data.
- AI Inference servers use a trained AI model to perform a task. They typically require less powerful hardware that may or may not be supercomputers.



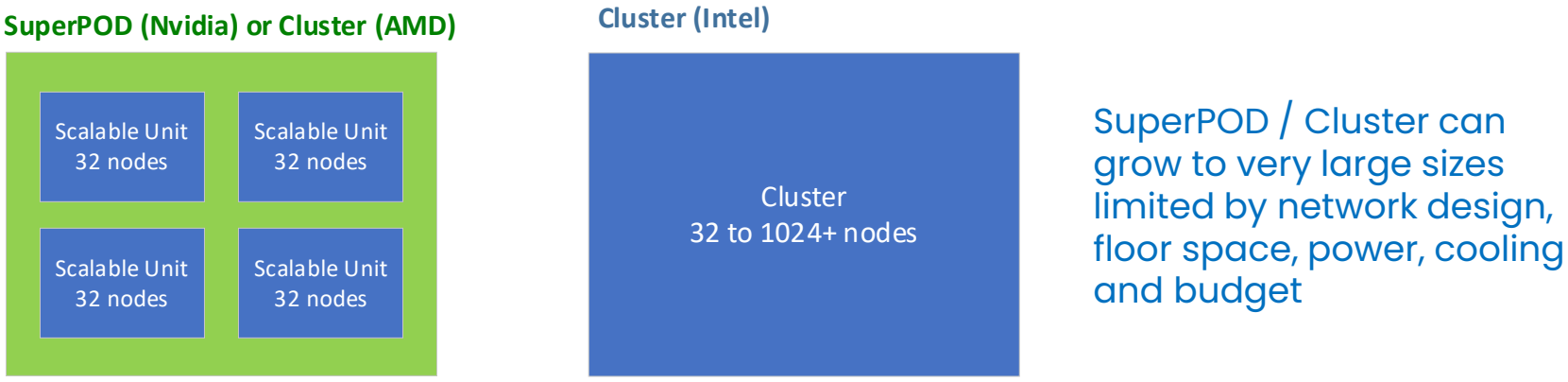
AI Reference Architectures

- Guidelines that define a standardized and optimal configuration for AI supercomputer infrastructure, ensuring performance, scalability, and reliability
- Can be used for general purpose supercomputing
- We will review networking & cabling in current reference architectures for
  - Nvidia Blackwell GB200 & Tensor Core H200
  - AMD Instinct MI300
  - Intel Gaudi3
- These three manufactures provide >95% of GPU chips currently used in AI servers (other than custom chips used internally by companies like Google, Meta, and Amazon)

Bicsi CALA

Groupings of AI Servers

- Nvidia architecture has Scalable Units (SUs) of 32 nodes, multiple SUs form a SuperPOD
- AMD architecture has scalable units of 32 nodes, multiple SUs form a cluster
- Intel reference architecture is a 32-node cluster, but nodes can be added to make larger clusters



SuperPOD / Cluster can grow to very large sizes limited by network design, floor space, power, cooling and budget

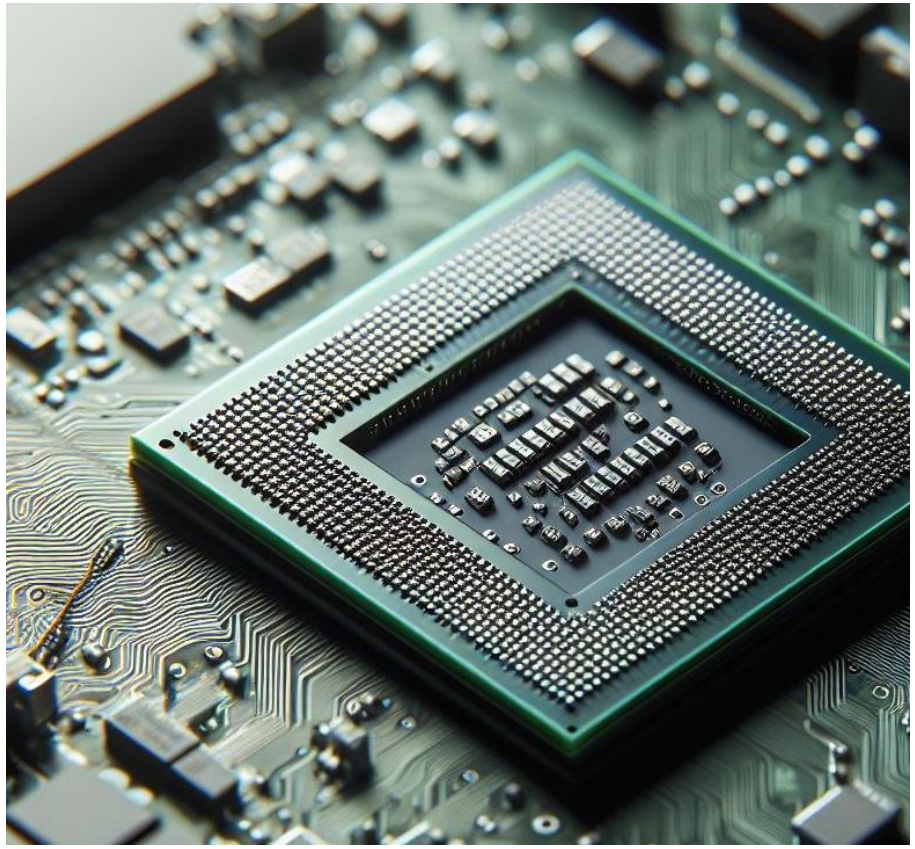
Common Configuration of AI Servers

- Nvidia, AMD, and Intel reference architectures use nodes/servers with 8 GPUs & 2 CPUs
- 6 or 8 ports 400G-800G Ethernet or InfiniBand for GPU network (for training nodes but typically not for inference nodes)
- 2 ports 100G-400G Ethernet or InfiniBand for storage network (connects nodes to storage servers)
- 2 ports 100G Ethernet for in-band management (may be combined with storage network in AMD design)
- 1 port 1 Gbps Ethernet out-of-band (OOB) management (server hardware management - e.g., IPMI, iDRAC, ILO, BMC)
- InfiniBand & Ethernet switches also have a separate switch serial console network using RS-232 (TIA-232-F) protocol over UTP to console servers

Bicsi CALA

GPU Network

- High Speed, low-latency, non-blocking network, for communication between GPUs in different servers/nodes
- Nvidia name – Compute fabric - 8 x 400G InfiniBand (preferred) or 8 x 400G Ethernet - in 2025 these will be 8 x 800G
- AMD name – Backend or Scale-out network - 8 x 400G Ethernet
- Intel name – Accelerator fabric - 6 x 800G Ethernet



Bicsi CALA

InfiniBand vs Ethernet for GPU Networks

- InfiniBand more common with Nvidia deployments due to lower latency provided by RDMA and low overhead protocol
- RDMA (Remote Direct Memory Access) allows data to be transferred by network adapters in different servers directly to memory without involving the CPUs
- Mellanox (now a subsidiary of Nvidia) has a near monopoly (>90%) of InfiniBand switch market
- Ethernet switches use IEEE 802.3 standards and are available from a wide variety of sources at competitive prices
- RoCE (RDMA over Converged Ethernet) protocol along with other networking features other allow Ethernet to have similar low latency
- Colossus (world's largest supercomputer cluster with 100,000 GPUs and being doubled to H200 200,000 GPUs) uses Nvidia Ethernet switches and has latency similar to InfiniBand(~100 ns)
- InfiniBand has a slight edge in latency but Ethernet with RoCE has an advantage in cost, standard IP network layer, and interoperability with existing networks

GPU Network Requirements

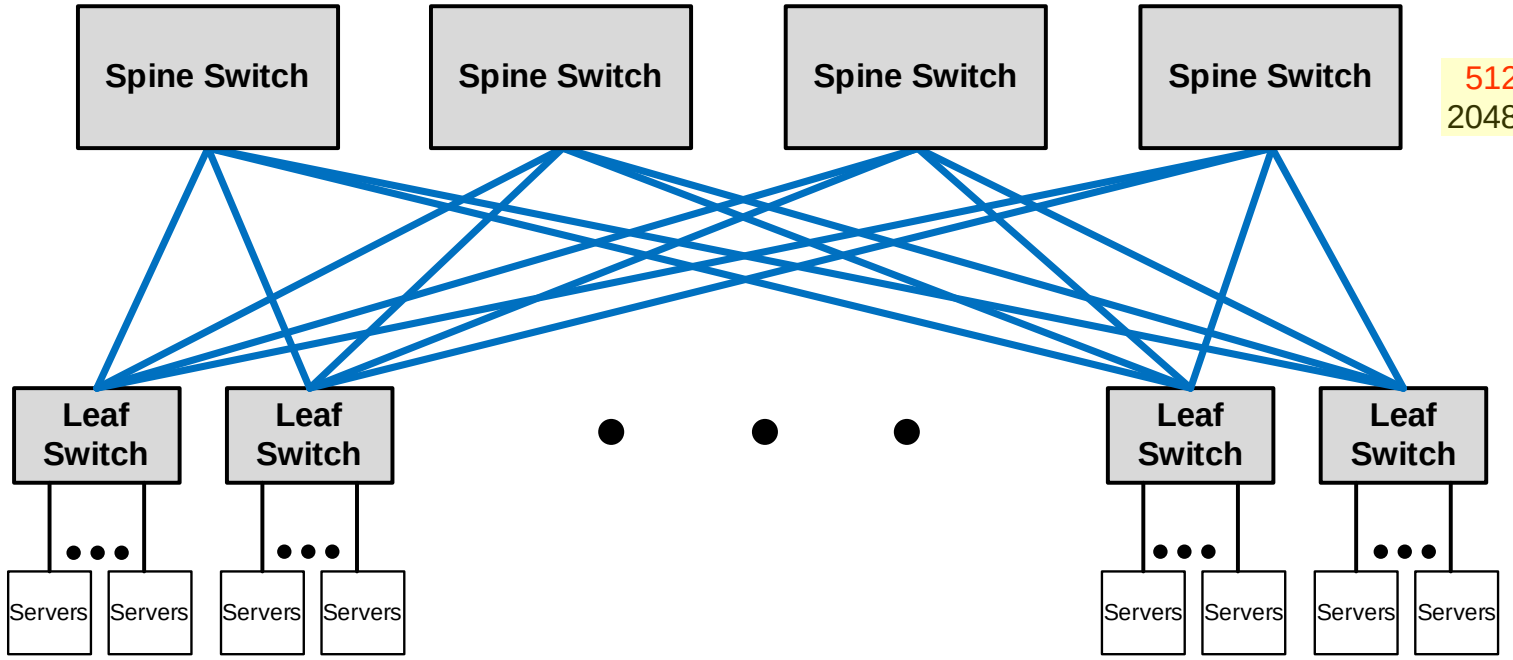
- GPU training workloads require high speeds and low latency
- High Speeds: 400G-800G in 2024-2025
  800G-1.6TB in 2026+
- Use Non-Blocking Clos network (named after Charles Clos [the s is silent]) to minimize intermediate switches and ensure all nodes can communicate all full speed with any other node without congestion
- Fat tree networks are a type of Clos network

Mr. Charles Clos on the right the father of Clos networks

Non-Blocking Leaf-Spine Network
(2 Tier-Fat Tree)

GPU Networks are Non-Blocking Commonly 400G with New Deployments

(all ports may transmit at 400G at the same time)



Max # of nodes depends on radix or port count of switches

512 server ports (64 servers) max with 32-port switches
2048 server ports (256 servers) max with 64-port switches

16 x 400G Connections to Spine from Each Leaf Switch
All Leafs Connected to All Spines

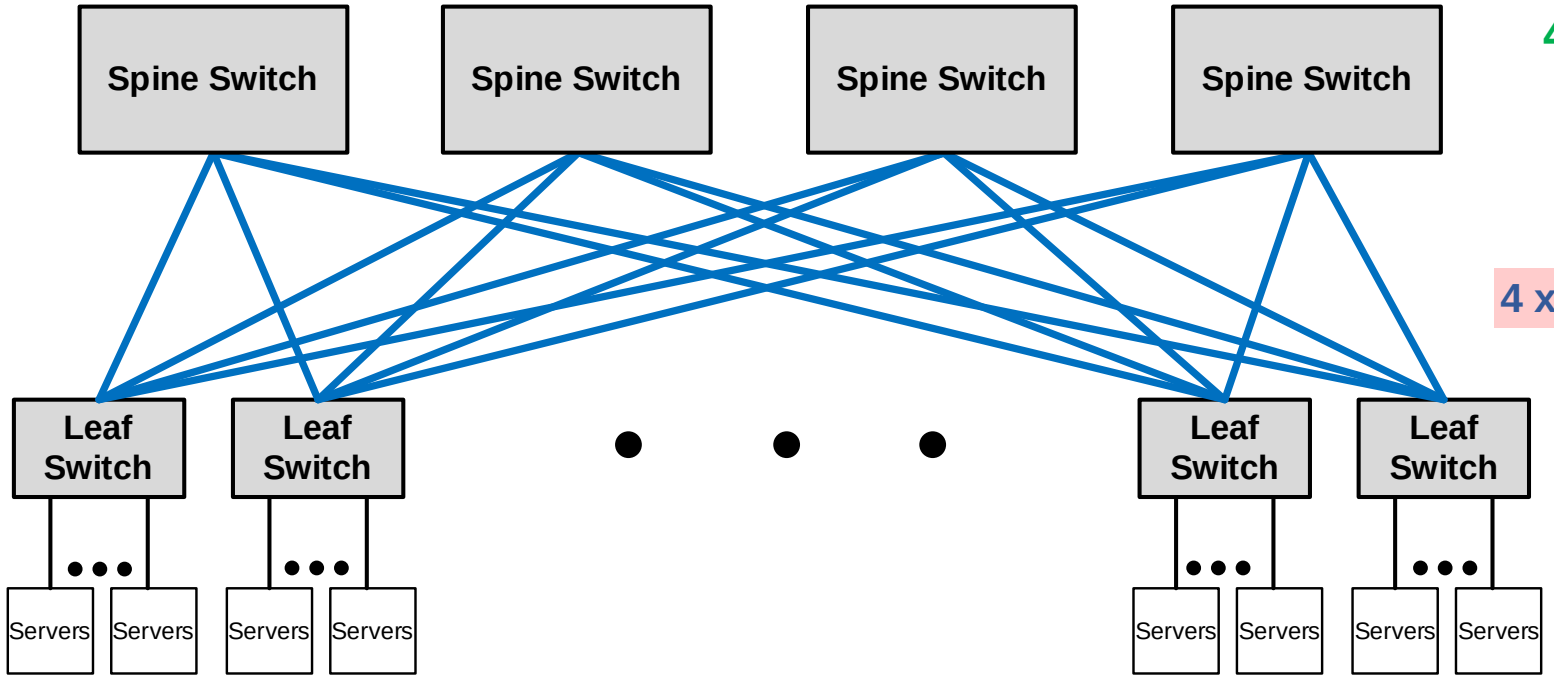
16 x 400G Connections from Servers to Each Leaf Switch

Leaf Switches have ½ of ports for servers and ½ of ports to spine

4:1 Over-Subscribed Leaf-Spine Network

Storage & In-Band Management Networks may be Over-Subscribed and use 100G – 400G ports

(not all ports transmit at full speed concurrently)



4:1 Downlink (toward server) to Uplink (toward higher layer) Over-Subscription Ratio

4 x 400G Connections to Spine from Each Leaf Switch

16 x 400G Connections from Servers to Each Leaf Switch

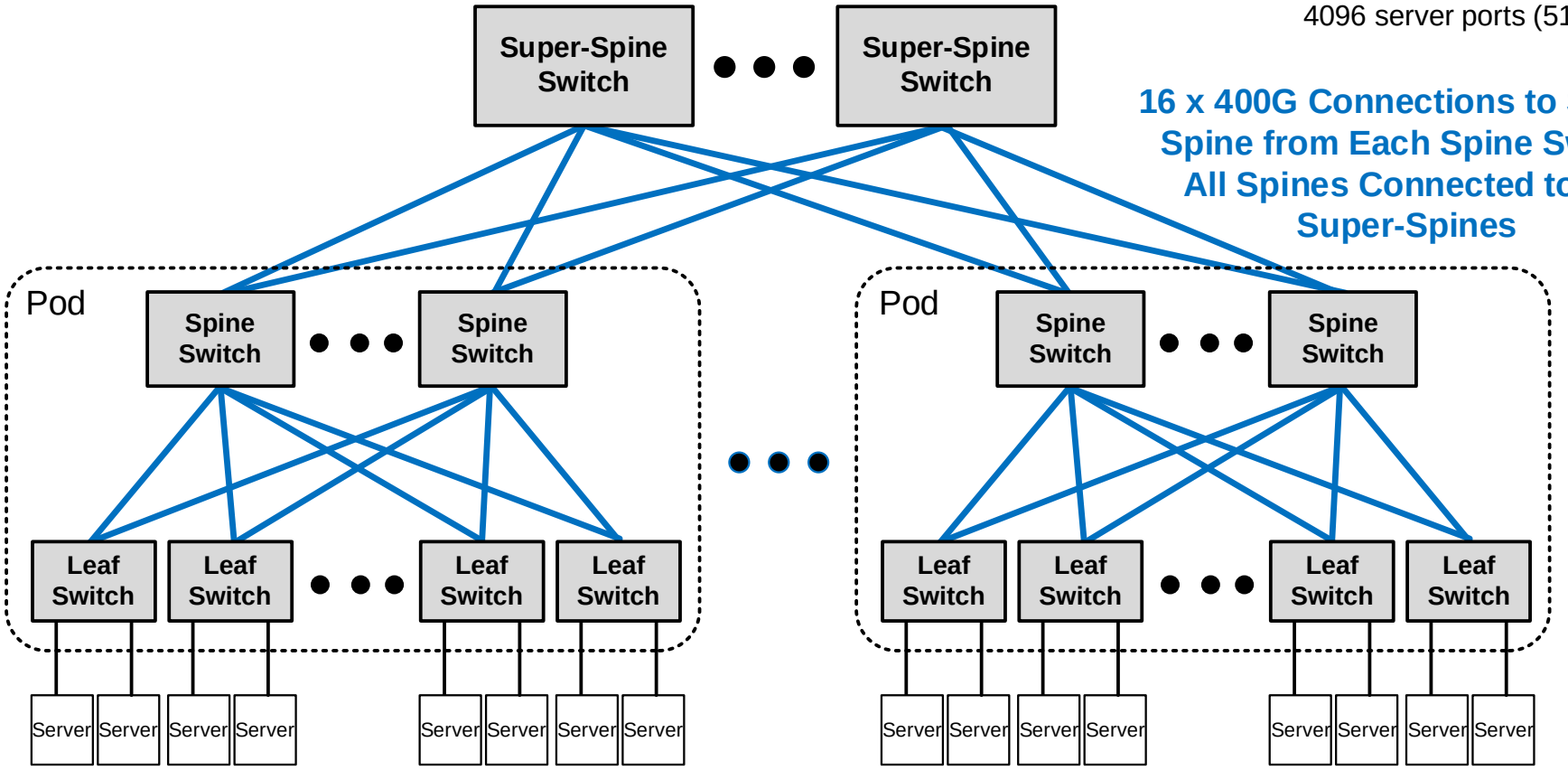
Add Super Spine to Increase Network Size

(3 Tier Fat-Tree)

Additional tier doubles # of nodes

1024 server ports (128 servers) max with 32-port switches

4096 server ports (512 servers) max with 64-port switches



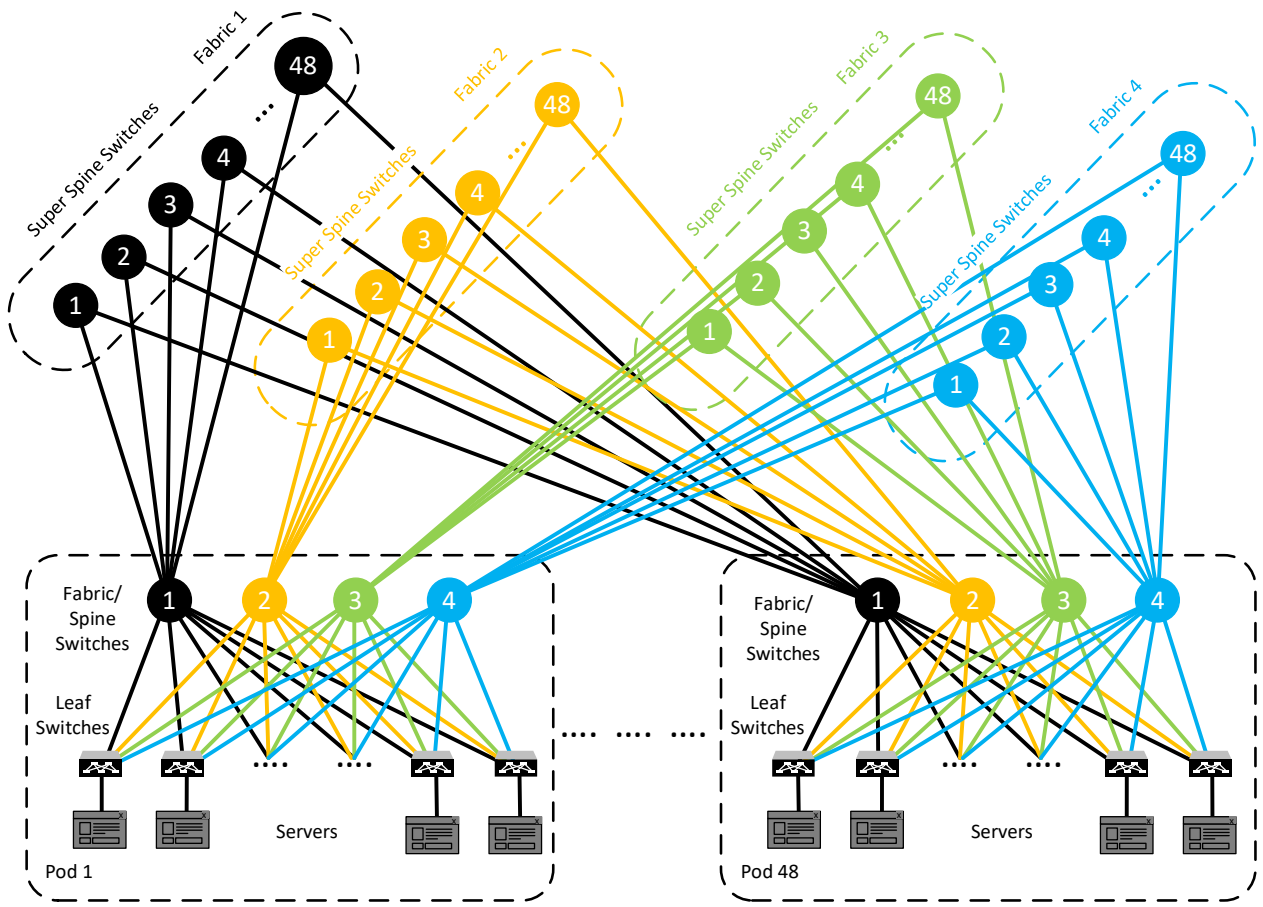
16 x 400G Connections to Super-Spine from Each Spine Switch
All Spines Connected to All Super-Spines

16 x 400G Connections to Spine from Each Leaf Switch
All Leafs Connected to All Spines Within Pod

16 x 400G Connections from Servers to Each Leaf Switch

Leaf & Spine Switches have ½ of ports going down 1-level and ½ of ports going up one level

Hyperscale Fabric for Larger AI Networks

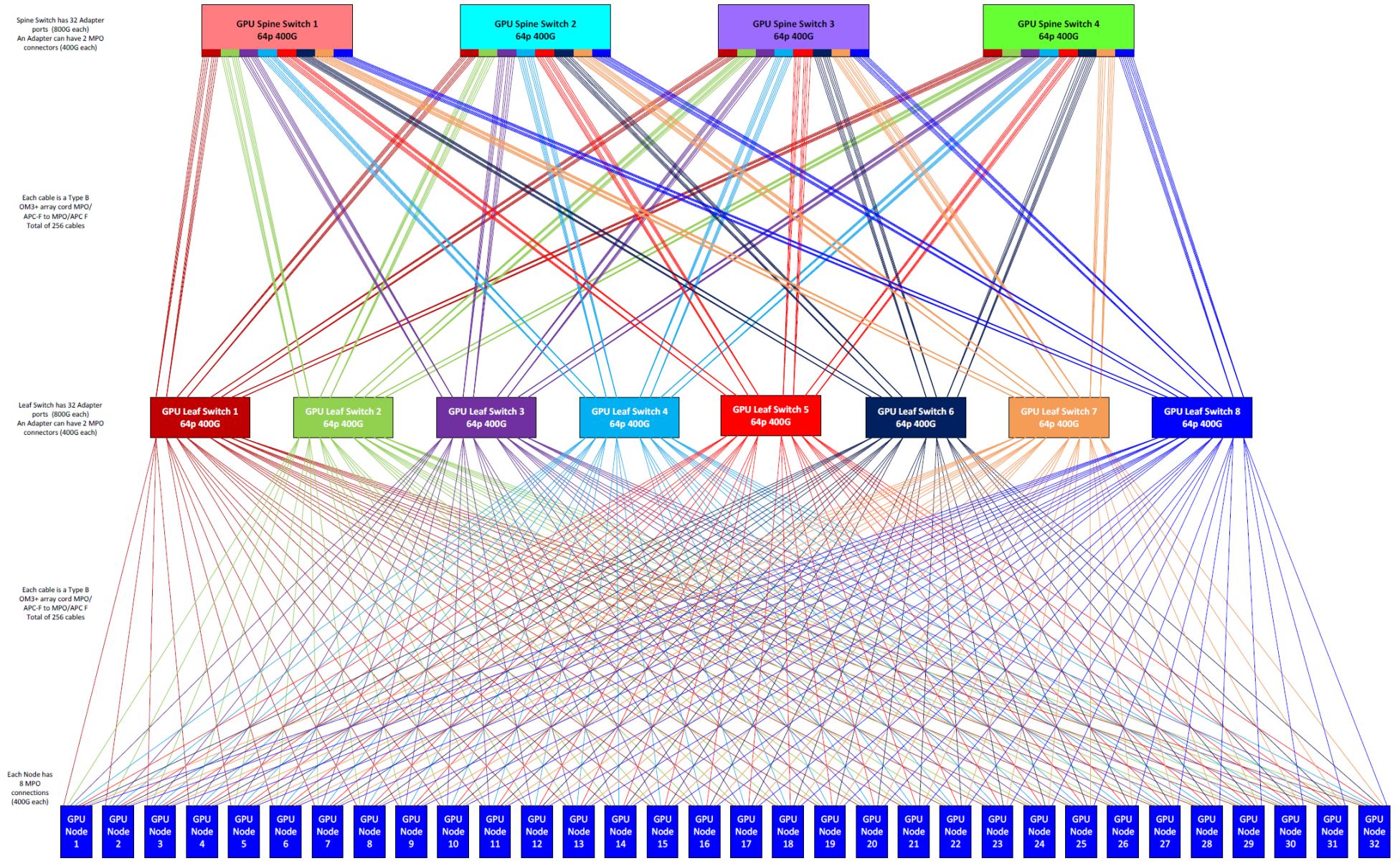


- Another type of Clos network (developed by Meta & named Hyperscale fabric plane Clos design in Cisco MSDC white paper)
- Leaf switches connected to a spine/fabric switch for each of the fabrics
- Each spine/fabric switch connected to all super spine switches in its fabric
- Leaf & spine switches have ½ of ports up 1-level and ½ of ports down 1 level
- The network can grow larger by deploying higher port-count spine & super spine switches or by adding fabrics
- 4 fabric network with 64-port spines & super spines has 8,192 server ports (1024 servers)
- 4 fabric network with 128-port spines & super spines has 32,768 server ports (4096 servers)

Bicsi CALA

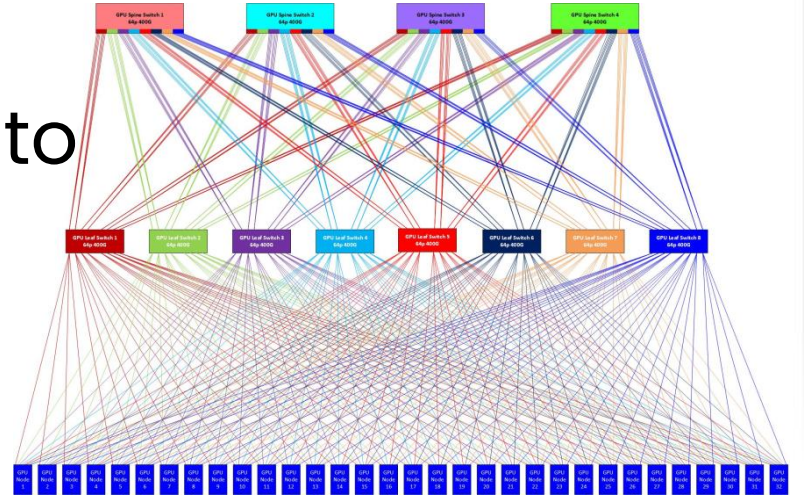
Rail-Optimized Topology

- Ensures no more than one switch between any two nodes in a scalable unit
- Specified for GPU network by Nvidia, but not required in AMD & Intel reference architectures



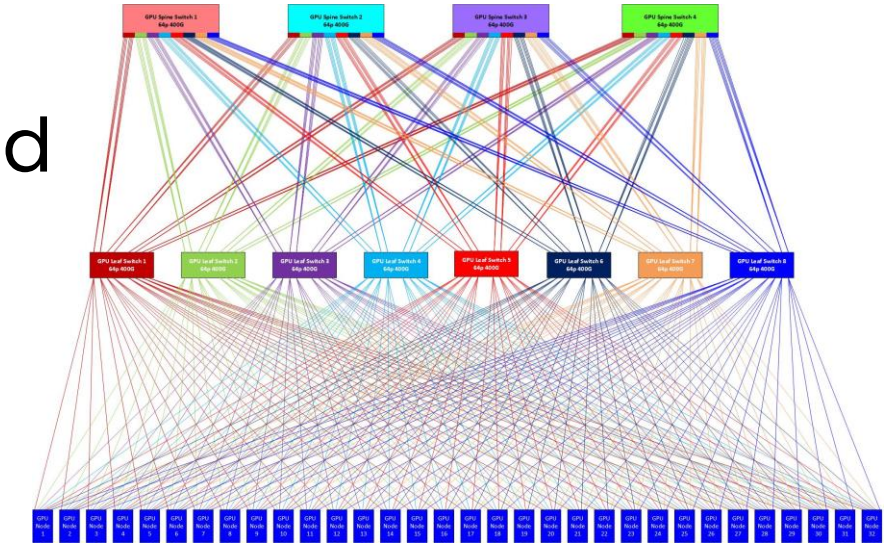
Nvidia GPU Network (Compute Fabric)

- Single non-blocking full fat tree network
- Rail-optimized
- 400G InfiniBand switches (Ethernet with RCoE allowed)
- Each Scalable Unit has 8 x 64-port Leaf Switches with 32 uplinks to spine & 32 ports to nodes (1 port to each node)
- Each node/server has 8 x 400G ports – one port to each Leaf Switch in rail-optimized configuration
- Will be upgraded to 800G in 2025



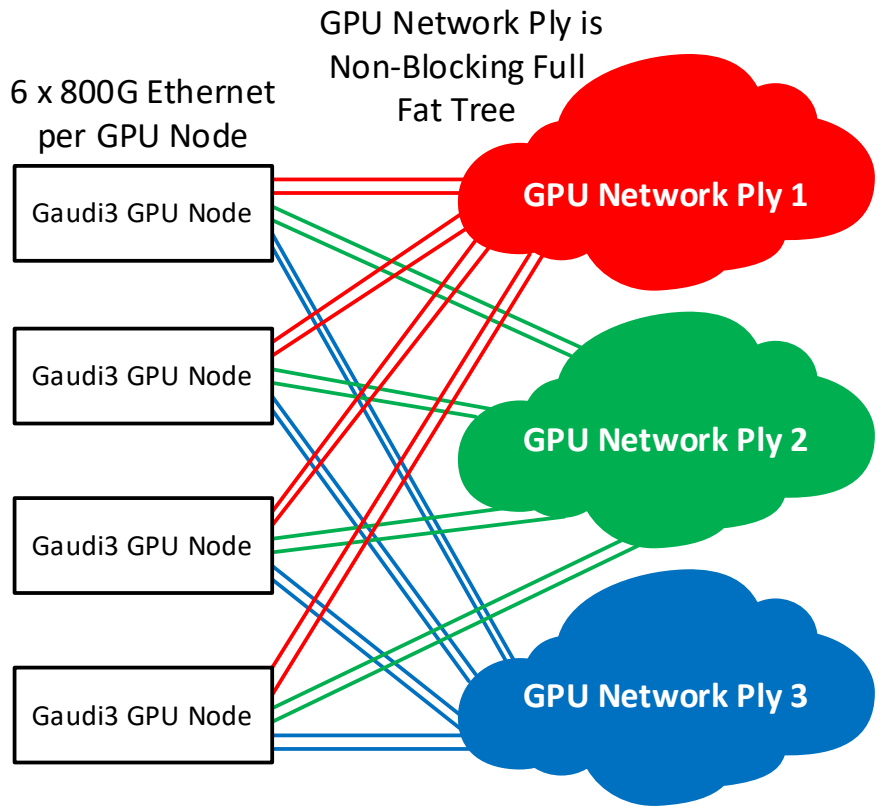
AMD GPU Network (Backend Network)

- Single non-blocking full fat tree network
- May be rail-optimized to minimize latency
- 400G Ethernet network (with RCoE) with a minimum of 64 ports on leaf switches
- Leaf switches have ½ of ports to nodes and ½ of ports to spine
- Each node has 8 x 400G ports to leaf switches
- It is safer, but not necessary for upper layers particularly at higher tiers (e.g., spine to super spine) be under-subscribed to handle uneven traffic



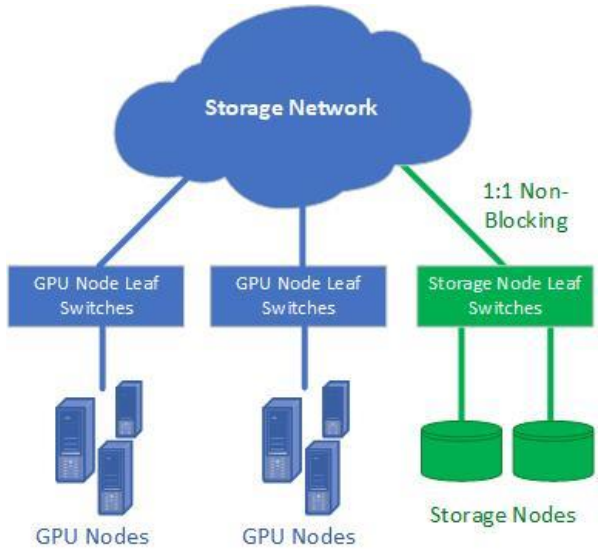
Intel GPU Network (Accelerator Fabric)

- 3 non-blocking full Clos networks (3 accelerator plies) which may be separate or joined at higher tiers
- 800G Ethernet Network (with RCoE)
- Reference architecture uses 12 x 32 port 800G leaf switches – 4 leaf switches/ply
- Each node has 6 x 800G ports – 2 ports per ply



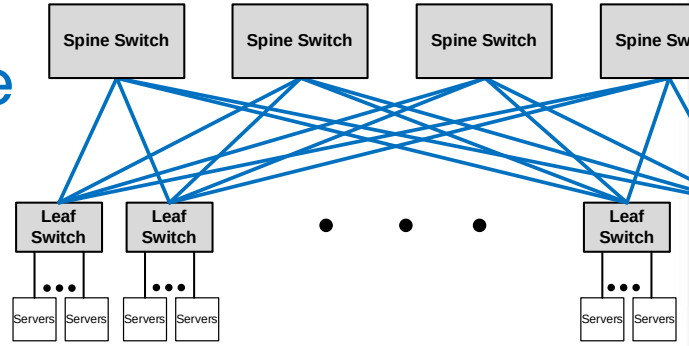
Storage Network

- Separate from GPU Network & connected to Storage nodes. Leaf switches for storage nodes not over-subscribed
- Nvidia (Storage Fabric) - 2 x 400G InfiniBand ports per node to 400G fat tree network with ~4:3 over subscription to compute nodes. 320 Gbps per node
- AMD (Storage Network) – 2 x 400G Ethernet ports per node. May be combined with in-band management to form a Frontend Network –Ethernet with RCoE (recommended) or InfiniBand. May be over-subscribed to compute nodes.
- Intel (Storage Ply) – 2 x 100G Ethernet ports per node to non-blocking full Clos network.



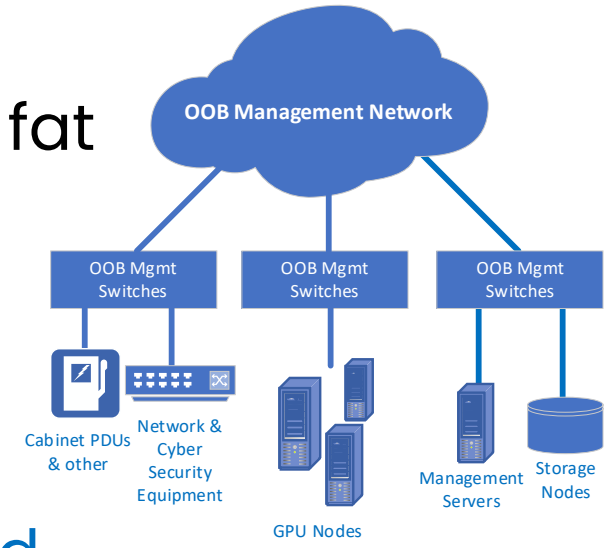
In-Band Management Network

- Uses Ethernet and used for GPU node deployment, management, operation, and GPU node communication with management servers & customer network
- Nvidia (In-Band Network) – 2 x 100G Ethernet to 4:1 over-subscribed leaf-spine network
- AMD (In Band Network typically combined with Storage Network to form a Frontend Network) – if separate it should be 100G Ethernet
- Intel (Control Plane Fabric) – 2 x 100G Ethernet to non-blocking full Clos network



Out-of-Band (OOB) Management Network

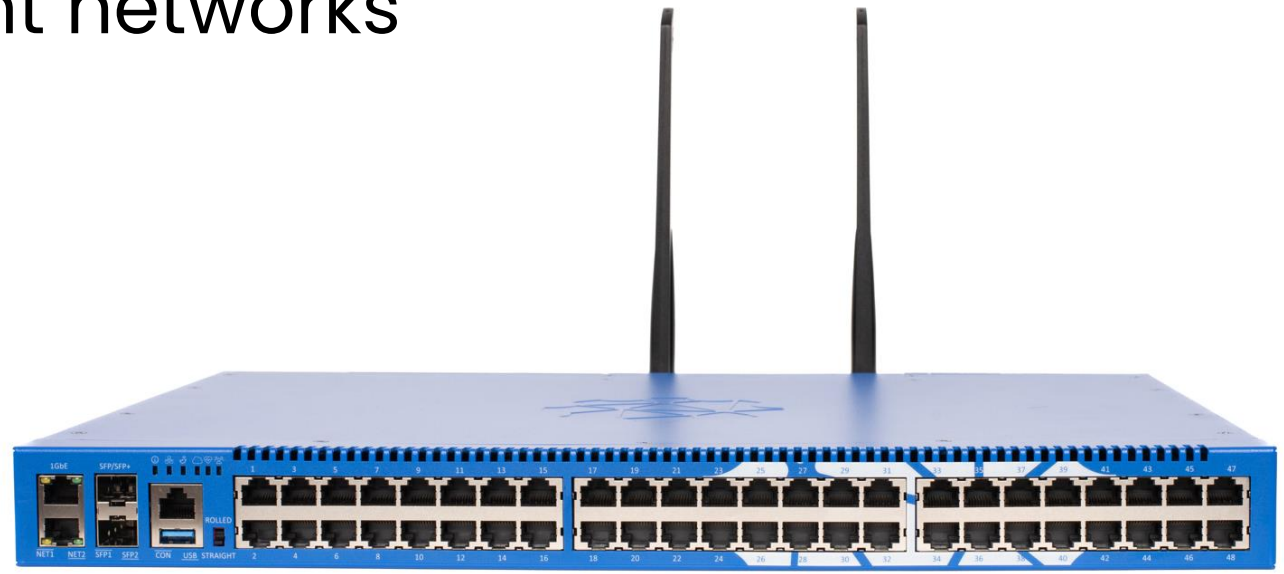
- Connects management ports of all devices including switches, cabinet PDUs, management servers, storage servers, GPU servers
- Typically uses 1 Gbps Ethernet with UTP
- Low bandwidth network with no need for non-blocking fat tree design
- Similar in all three reference designs
- This network may have uplinks to the In-Band Management Network
- All Ethernet-based networks (e.g., GPU, Storage, In-Band, OOB) may be joined to a common backbone (though not necessarily non-blocking to the backbone)



Switch Console Network

- Separate network using console servers to serial ports on switches
- Uses UTP and RS-232 (TIA-232-F) protocol
- Console servers may have uplinks to one of the management networks

Serial console server with cellular modem & Ethernet uplinks

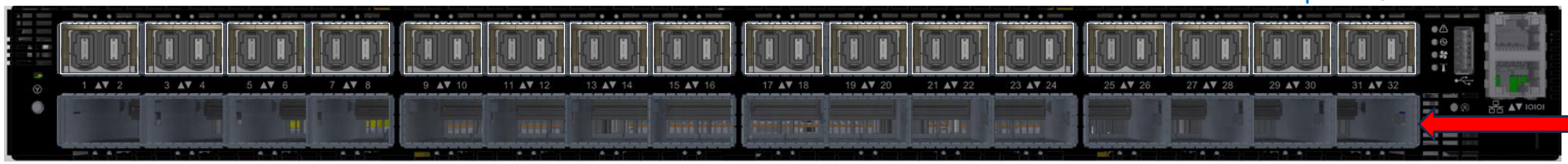


Bicsi CALA

Transceiver Cages on Nodes & Switches

- Ports on Network Interface Cards (NICs) and switches are typically transceiver cages that can accept:
  - Transceivers with optical fiber connector (e.g., LC, MPO) to fiber cabling - distances depend on protocol (100 m to 2-40 km)
  - Direct attach copper cables with transceiver housings affixed on both ends
    - DAC – direct attach cable – passive DAC cables (up to ~3 m)
    - ACC – active copper cable – active DAC cable (up to ~5 m) Amplifies electrical signal
    - AEC – active electrical cables – active DAC cable (up to ~7 m) Amplifies signals and includes clock data recovery to reduce signal jitter
    - Lengths above are for 100G, 400G, 800G cables used for AI networks
  - Active Optical Cables (AOC) fiber cables with transceivers affixed on both ends (up to 100 m)

32-port QSFP-DD switch with transceivers to spine in top ports & openings for DACs on bottom ports



Typical Transceiver Form Factors for AI

- Transceivers and connectors on direct attach cables must be compatible with the transceiver cages on the server NICs and switches
- Typical transceiver form factors used in AI networks:
  - 800G –OSFP or QSFP-DD
  - 400G – OSFP, QSFP-DD, QSFP112
  - 100G – QSFP28

Defined in MSAs (multi-source agreements)

OSFP

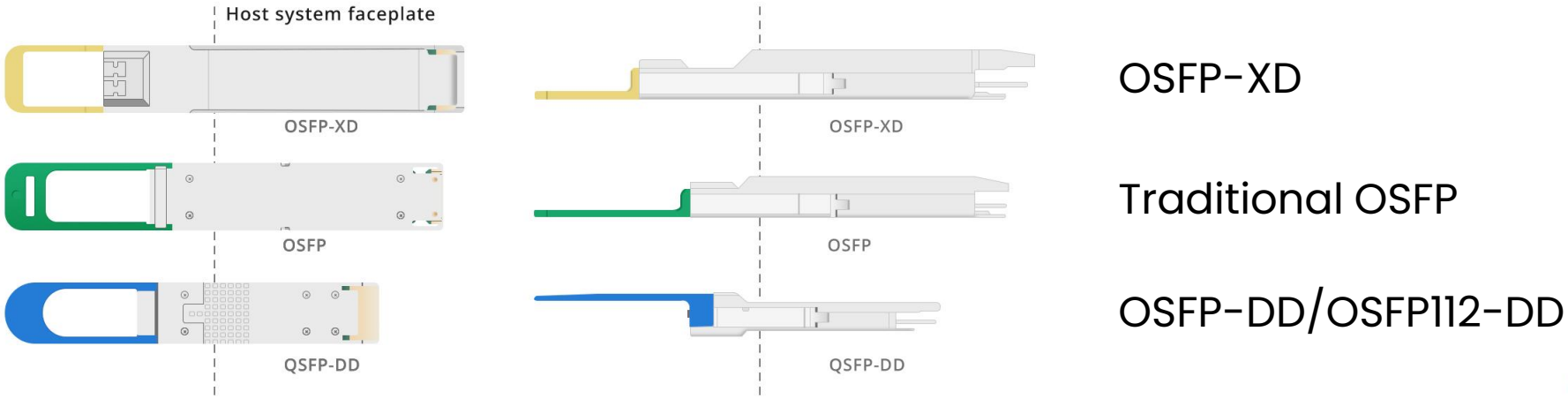
QSFP-DD

QSFP112

QSFP28

Expected Common Transceivers for 1.6T

- QSFP112-DD - A variant of QSFP-DD designed to support higher speeds. QSFP112-DD cage is the same size as QSFP-DD and will typically accept QSFP-DD (800G/400G), QSFP56 (200G), and QSFP28 (100G) transceivers
- OSFP-XD - OSFP update designed for 1.6T and 3.2T. Better power efficiency and higher port density than both OSFP and QSFP112-DD (e.g., 36 vs 32 in 1 RU).
- OSFP – Traditional OSFP design compatible supporting 1.6T, 800G and 400G transceivers, but not compatible with OSFP-XD (or QSFP-DD).



*Source: OSFP1600 and OSFP-XD MSA

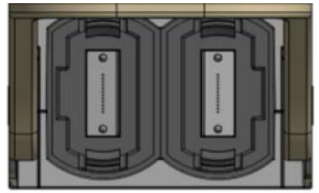
Transceiver Protocols & Coding

- Direct attach cables and transceivers are purchased for a particular network protocol (InfiniBand or Ethernet) and speed. Transceivers are also specific to physical layer protocol (e.g., 400GBASE-DR4: 400G over 8 SMF).
- They typically need to be coded to be recognized by the network equipment (e.g., server NIC or switch) and operate properly - Manufacturers use different codes
- These may use the same or different manufacturer coding at each end (e.g., an AOC between HP and Arista switches)
- Coding information is encoded by the transceiver/cable manufacturer, but can be done in the field with the right equipment
- 3rd party direct attach cables & transceivers should be tested for compatibility with network equipment

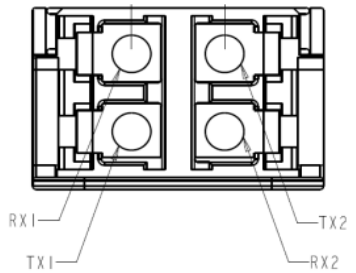
Typical Transceiver Optical Fiber Connectors

- Typical optical fiber connectors used for AI networking
  - 800G – two Base8 angled MPO12 connectors (8 fibers used) for both MMF (100m) and SMF (500m) or two LC/UPC SMF to 2km
  - 400G – one Base8 angled MPO12 connector (8 fibers used) for both MMF (100m) and SMF (500m) or two LC/UPC to 2km
  - 100G – two LC/UPC for both MMF (100m) & SMF (500m-40km) not angled
- IEEE 802.3dj (Mar 2026)
  - 800G over 8 single-mode fibers
  - 1.6 TB over 16 single-mode fibers
  - Possible more widespread adoption of VSFF array connectors like MMC and SN-MT which have ~3x density of MPO

800Gb/s Twin-port 400BASE-DR4 Transceiver



800Gb/s Twin-port 400GBASE-FR4 Transceiver

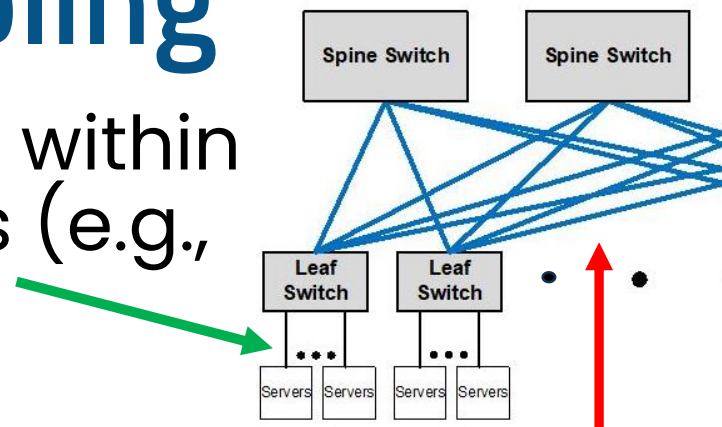


Direct Attach Cables vs Transceivers

- Direct attach copper cables have lowest cost but shortest distances (3-7 m). They have low power consumption.
- AOC cables have longer reach (100 m), but consume more power and are more expensive than DAC
- Transceivers with optical fiber have slightly more latency (but still very low), more flexibility than AOC, easier troubleshooting, and less cost with changes in speeds or protocols

Direct Attach vs Structured Cabling

- Direct attach cables suitable for connections within cabinets and within row to adjacent cabinets (e.g., nodes to leaf switches)
  - DAC, AEC, ACC, Transceivers with cords
- Structured cabling suitable in distributors (e.g., location of spine, fabric, and super spine switches) for cabling from leaf switches to spine/fabric switches and spine/fabric to super spine switches.
  - Uses transceivers with cords to structured cabling
  - Structured cabling infrastructure may be used for changes in speeds & protocols
  - Structured cabling permits use of high-count trunk cables to save space in cable pathways



AI Network Cable Counts & Pathways

- AI networks have lots of cables
- Very important to estimate required pathway sizes (size & # of optical fiber ducts) at locations of spine, fabric, and super spine switches
- # of fiber ducts below is estimate using no structured cabling

Nodes (Servers)	Node-to-Leaf Switches	Leaf Switches-to-Spine Switches	Spine Switches-to-Super Spine Switches	Area of Leaf-Spine Switch AOC (mm^2)	# of 600 mm x 100 mm fiber ducts (GPU Net)	# of 600 mm x 100 mm fiber ducts (All Nets)
128	1024	1024		7235	0.2	0.3
256	2048	2048		14469	0.5	0.6
512	4096	4096	4096	28938	1.0	1.3
1024	8192	8192	8192	57876	1.9	2.5
2048	16384	16384	16384	115753	3.9	5.0

- Structured cabling can use high-count fiber trunks between patch panels to reduce cable volume
  - Avoid mixing heavier cables (power, UTP, large fiber trunks) with small diameter optical fiber cords
  - Use pathways with solid bottoms for fiber cords & small diameter trunks

Minimum # of fiber ducts at main distributor with AOCs only

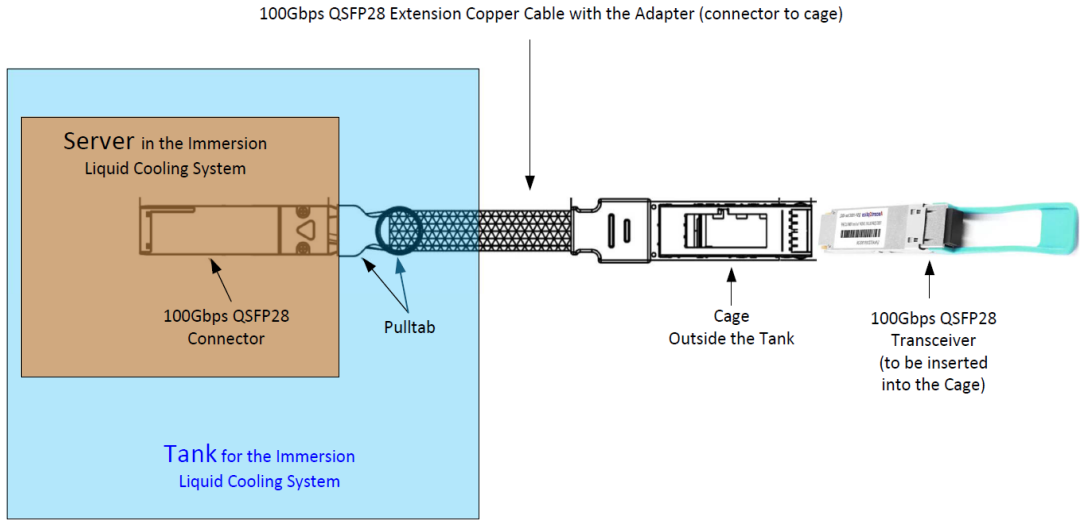
Bicsi CALA

Immersion Cooling Considerations

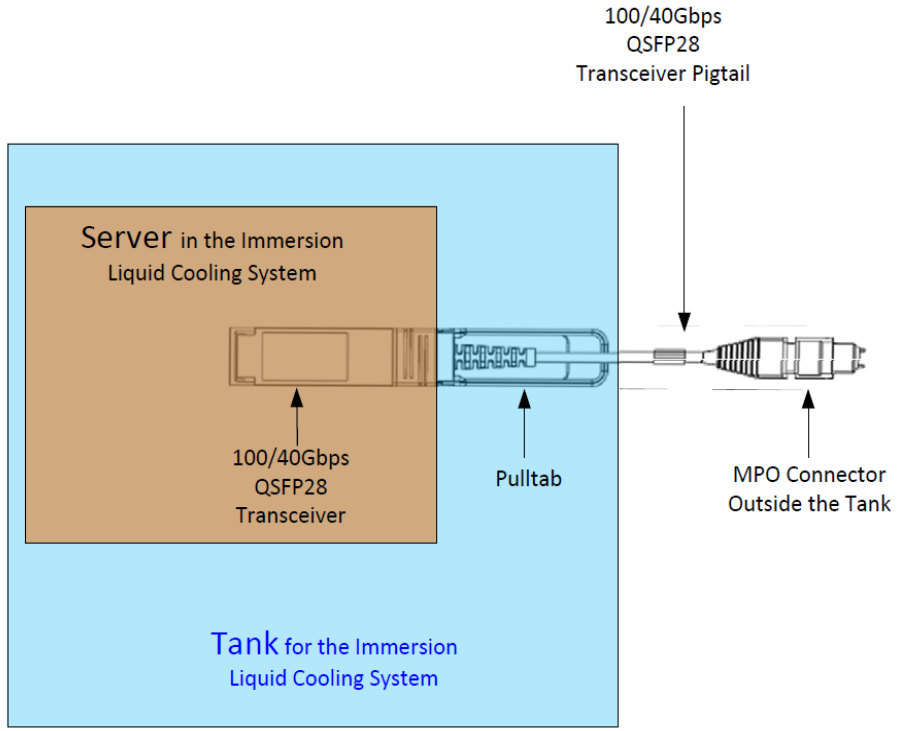
- Cable jackets must be suitable for use in the cooling fluid (e.g., UTP cables with Polyurethane (PUR) jacket)
- Copper direct attach cables (e.g., DAC) work more reliably in fluid than AOC unless AOC is designed for immersion
- Liquid can enter optical fiber connectors degrading performance
- Can use liquid-tight immersion transceivers with built in pigtail (e.g., 400G QSFP-DD transceiver with sealed MMF cable to MPO - MPO end connects to patch panel)
- Can use immersion extender cables (e.g., copper cable with QSFP28 connector to QSFP28 cage - QSFP28 cage holds transceiver out of liquid)
- Any solution should be tested for reliability

Immersion Extender & Immersion Pigtail

100Gbps QSFP28 Extension Copper Cable with the Adapter (connector to cage) with Transceiver for Immersion Liquid Cooling System



100/40Gbps QSFP28 Transceiver Pigtail for Immersion Liquid Cooling System



Questions?

- Jonathan Jew - jew@j-and-m.com
- Editor of ANSI/TIA-942-C
- Chair of working group for ANSI/BICSI 002-2024
- USTAG ISO/IEC JTC 1 SC25 WG 3 Project Lead for ISO/IEC 11801-5 Data Center Cabling