Reference Cluster Design Architecture MI300X / MI325X

128 GPU to 1024 GPU

Version 4.0 – August 2025

AMD
together we advance_

Cluster - 128 GPU to 1024 GPU (16 to 128 Nodes) *

Cluster Size	128 to 1024 GPU
Platforms	Dell XE9680 Lenovo SR685a V3 SMCI AS-8125GS
OS	Ubuntu 22.04 (or above)
Linux Kernel	5.15 – 6.80
ROCm	6.3.3 (or above)

*Genericized BoM – removing device specifics including dependencies

Storage Type	
Local Storage	1.6 TB (or greater)
Utility Storage	Pure, Vast, RYO
Bulk Storage	Pure, Vast, WekaIO
Scratch Storage	Vast, DDN, WekaIO, Hammerspace
Archive/Object Storage	S3 Compatible

Network - 128 GPU to 1024 GPU (128 to 1024 Nodes)*

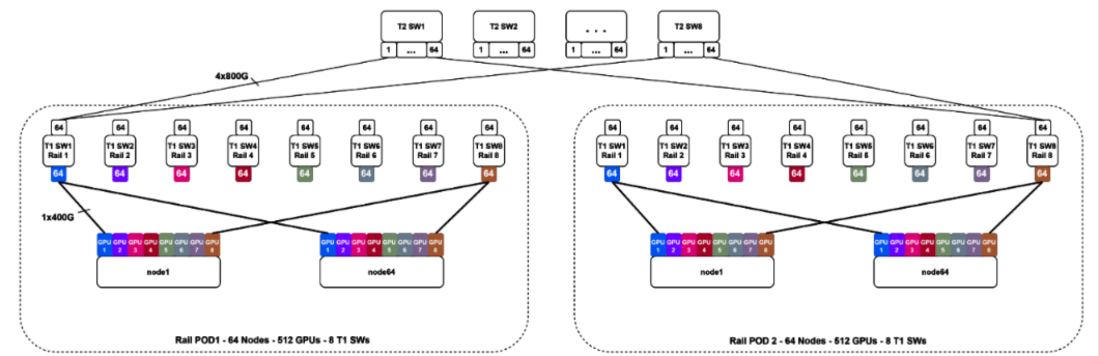
Backside Network Topology	2 Tier Rail Optimized / Fat Tree
NIC	Pollara 400 BCM957608 (Thor2)
Switch	Arista, Dell, Juniper, Cisco (TH 4/5, Jericho/Ramon)
Network OS	SONiC, Junos, EOS, IOS
Subscription Ratio	1:1.16=16% Undersubscribed (AMD recommendation)
Optics	Vendor ACL/HCL transceivers or Direct Attach Copper
Fabric	RoCEv2 Ethernet

Frontside Network Segment	Adapter Recommended
All-in One Network	Ethernet 100GbE 2-port QSFP28 Adapter
Storage Network (Optional)	Ethernet 100GbE 2-port QSFP28 Adapter
Virtualization Network (Optional)	Ethernet 100GbE 2-port QSFP28 Adapter
Host In-Band	Ethernet 10/25GbE 4-Port SFP28 Adapter
BMC OOB Mgt	1G Copper

* Genericized BoM – removing device specifics including dependencies

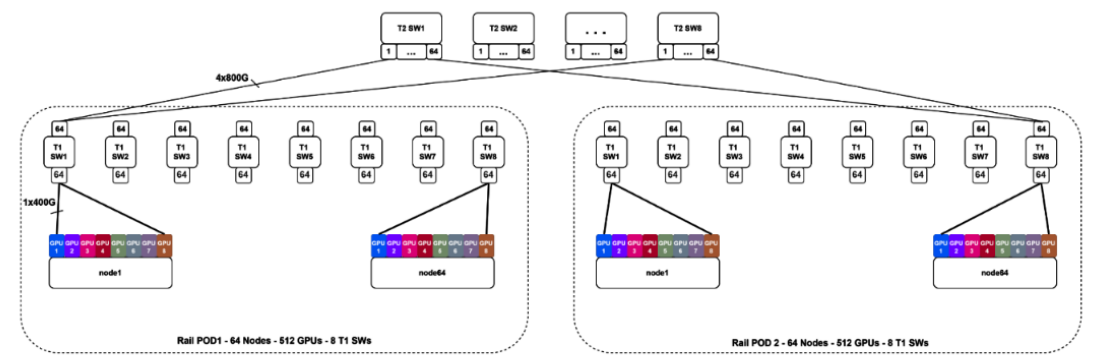
128 GPU to 1024 GPU – General Network Layouts

Rail Optimized (1:1 example)



- Rail design - 2 pods
- 64 Nodes per pod – 512 GPUs
- Sizing of network can be done through lower/increase spine layer and adjusting pod size.
- Lower latency for AI workloads – optimized CPU communication
- Improved performance for large scale models

Tree Topology (1:1 example)



- Tree design – 2 pods
- 64 Nodes per pod – 512 GPUs
- Sizing of network can be done through lower/increase spine layer and adjusting pod size.
- Non blocking bandwidth and scalability
- Load balancing and fault tolerance
- Diverse workloads

Sample Cluster BOMs
16 – 128 Nodes

Sample 128GPU (16N) Single Switch eBOM

Rack #1, #2, #4 - GPU Racks

MFG	MFG Part No / SKU	Material Description	Total Qty 3 Racks	Notes
HPE	S4Q28A	ProLiant Compute XD685 Air Cooling Server	12	GPU Nodes
AMD		Universal baseboard (UBB) module with with eight AMD Instinct™ MI325X Accelerators	12	GPU board
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	3	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	6	Data Center Rack PDUs

Rack #3 - GPU + Network Rack

MFG	MFG Part No / SKU	Material Description	Total Qty 1 Rack	Notes
Arista	DCS-7060X6-64PE-F	Arista 7060X6, 64 x 800GbE OSFP switch, front-to-rear air, 2xAC ; L3 license ; cloudvision 3yr ; NBD HW support	1	L1 BE Net Switch
Arista	OSFP-800G-2XDR4	Arista 800GBASE-2XDR4 OSFP Transceivers ; 2x 400GBASE-XDR4 Transceiver, Dual MPO-12 connector, 2km over parallel SMF	64	L1 BE Net Switch Transceivers
FS	69008	2m MTP® Jumper, MTP®-12 APC (Female) to MTP®-12 APC (Female), 12 Fibers, Single Mode (OS2)	128	L1 BE Net Switch to NIC Fiber Cables
Arista	DCS-7050CX3-32S-F	Arista DCS-7050CX3-32S-F	1	FE Net Switch
Arista	MMA1B00-C100D	Arista Networks QSFP-100G-SR4 Compatible QSFP28 MPO-12/UPC MMF Optical Transceiver Module, Support 4 x 25G-SR	16	FE Switch Transceivers
Broadcom	BCM957608-P2200GQF00	Broadcom P2100G - 2 x 100GbE PCIe NIC	16	FE Net NIC Adapter
FS.com	QSFP-100G-SR4	Broadcom compatible 100GBASE-SR4 QSFP100 Transceiver	16	FE Net NIC Adapter Transceivers
FS.com	12FMTPOM4	MTP® Jumper, MTP®-12 UPC (Female) to MTP®-12 UPC (Female), 12 Fibers, Multimode (OM4)	4	FE Net NIC Adapter to Switch Cables
Arista	DCS-7010TX-48-F	Arista 7010TX, 48x 10/100/1000 RJ45 & 4 x 25G SFP (1/10/25GbE) switch, front to rear air, 2xAC ; 3yr cloudvision ; 3yr NBD HW replacement	1	OOB MGMT Switch
FS.com	C6UTPSGSPVC	Cat6 28AWG Snagless Unshielded (UTP) PVC CM Slim Ethernet Network Patch Cable	16	OOB MGMT Cat 6 Cables
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	1	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	2	Data Center Rack PDUs

Sample 512GPU (64N) POD-Rail-Optimized eBOM

Rack #1 - #4, #6-#13, #15-#18 - GPU Racks

MFG	MFG Part No / SKU	Material Description	Total Qty 16 Racks	Notes
HPE	S4Q28A	ProLiant Compute XD685 Air Cooling Server	64	GPU Servers
AMD		Universal baseboard (UBB) module with with eight AMD Instinct™ MI325X Accelerators	64	GPU boards
AMD	Pollara-400-1Q400P	AMD Pollara 400	512	NIC Adapters
FS	184798	400GBASE-XDR4 QSFP112 Single Mode Transceiver, MPO-12 APC, 2km reach over parallel SMF	512	NIC Transceivers
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	16	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	32	Data Center Rack PDUs

Rack #5 - L1 BE Network Rack

MFG	MFG Part No / SKU	Material Description	Total Qty 1 Rack	Notes
Arista	DCS-7060X6-64PE-F	Arista 7060X6, 64 x 800GbE OSFP switch, front-to-rear air, 2xAC ; L3 license ; cloudvision 3yr ; NBD HW support	8	L1 BE Net Switch
Arista	OSFP-800G-2XDR4	Arista 800GBASE-2XDR4 OSFP Transceivers ; 2x 400GBASE-XDR4 Transceiver, Dual MPO-12 connector, 2km over parallel SMF	64	L1 BE Net Transceivers
FS	69008	2m MTP® Jumper, MTP®-12 APC (Female) to MTP®-12 APC (Female), 12 Fibers, Single Mode (OS2)	128	L1 Be Net Fiber Cables
Arista	DCS-7260CX3-64E-F	Arista 7260X3, 64x100GbE QSFP & 2xSFP+ Enhanced switch, front-to-rear air, 2xAC	2	FE Net Switch
Arista	MMA1B00-C100D	Arista Networks QSFP-100G-SR4 Compatible QSFP28 MPO-12/UPC MMF Optical Transceiver Module, Support 4 x 25G-SR	32	FE Net Switch Transceivers
Broadcom	BCM957608-P2200GQF00	Broadcom P2100G - 2 x 100GbE PCIe NIC	32	FE Net NICs
FS.com	QSFP-100G-SR4	Broadcom compatible 100GBASE-SR4 QSFP100 Transceiver	32	FE Net NIC Transceivers
FS.com	12FMTPOM4	MTP® Jumper, MTP®-12 UPC (Female) to MTP®-12 UPC (Female), 12 Fibers, Multimode (OM4)	32	FE Net fiber Cables
Arista	DCS-7010TX-48-F	Arista 7010TX, 48x 10/100/1000 RJ45 & 4 x 25G SFP (1/10/25GbE) switch, front to rear air, 2xAC ; 3yr cloudvision ; 3yr NBD HW replacement	1	OOB MGMT Switch
FS.com	C6UTPSGSPVC	Cat6 28AWG Snagless Unshielded (UTP) PVC CM Slim Ethernet Network Patch Cable	64	OOB MGMT Cat 6 Cables
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	1	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	2	Data Center Rack PDUs

Rack #14 - L1 FE Network Rack

MFG	MFG Part No / SKU	Material Description	Total Qty 1 Rack	Notes
Arista	DCS-7260CX3-64E-F	Arista 7260X3, 64x100GbE QSFP & 2xSFP+ Enhanced switch, front-to-rear air, 2xAC	2	FE Net Switch
Arista	MMA1B00-C100D	Arista Networks QSFP-100G-SR4 Compatible QSFP28 MPO-12/UPC MMF Optical Transceiver Module, Support 4 x 25G-SR	32	FE Net Switch Transceivers
Broadcom	BCM957608-P2200GQF00	Broadcom P2100G - 2 x 100GbE PCIe NIC	32	FE Net NICs
FS.com	QSFP-100G-SR4	Broadcom compatible 100GBASE-SR4 QSFP100 Transceiver	32	FE Net NIC Transceivers
FS.com	12FMTPOM4	MTP® Jumper, MTP®-12 UPC (Female) to MTP®-12 UPC (Female), 12 Fibers, Multimode (OM4)	32	FE Net fiber Cables
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	1	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	2	Data Center Rack PDUs

Rack #XX - L2 Spine Network Rack

MFG	MFG Part No / SKU	Material Description	Total Qty 1 Rack	Notes
Arista	DCS-7060X6-64PE-F	Arista 7060X6, 64 x 800GbE OSFP switch, front-to-rear air, 2xAC ; L3 license ; cloudvision 3yr ; NBD HW support	8	L2 BE Net Switch
Arista	OSFP-800G-2XDR4	Arista 800GBASE-2XDR4 OSFP Transceivers ; 2x 400GBASE-XDR4 Transceiver, Dual MPO-12 connector, 2km over parallel SMF	64	L2 BE Net Transceivers
FS	69008	2m MTP® Jumper, MTP®-12 APC (Female) to MTP®-12 APC (Female), 12 Fibers, Single Mode (OS2)	128	L2 Be Net Fiber Cables
Arista	DCS-7260CX3-64E-F	Arista 7260X3, 64x100GbE QSFP & 2xSFP+ Enhanced switch, front-to-rear air, 2xAC	2	FE Net Spine Switch
Arista	MMA1B00-C100D	Arista Networks QSFP-100G-SR4 Compatible QSFP28 MPO-12/UPC MMF Optical Transceiver Module, Support 4 x 25G-SR	64	FE Net Spine Switch Transceivers
FS.com	12FMTPOM4	MTP® Jumper, MTP®-12 UPC (Female) to MTP®-12 UPC (Female), 12 Fibers, Multimode (OM4)	64	FE Net fiber Cables
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	1	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	2	Data Center Rack PDUs

Sample 1024GPU (128N) POD-Rail-Optimized eBOM

Rack #1 - #4, #6-#13, #15-#22, #24-#31, #33-#36 - GPU Racks

MFG	MFG Part No / SKU	Material Description	Total Qty 32 Racks	Notes
HPE	S4Q28A	ProLiant Compute XD685 Air Cooling Server	128	GPU Servers
AMD		Universal baseboard (UBB) module with with eight AMD Instinct™ MI325X Accelerators	128	GPU boards
AMD	Pollara-400-1Q400P	AMD Pollara 400	1024	NIC Adapters
FS	184798	400GBASE-XDR4 QSFP112 Single Mode Transceiver, MPO-12 APC, 2km reach over parallel SMF	1024	NIC Transceivers
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	32	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	64	Data Center Rack PDUs

Rack #5,23 - L1 BE Network Racks

MFG	MFG Part No / SKU	Material Description	Total Qty 2 Racks	Notes
Arista	DCS-7060X6-64PE-F	Arista 7060X6, 64 x 800GbE OSFP switch, front-to-rear air, 2xAC ; L3 license ; cloudvision 3yr ; NBD HW support	16	L1 BE Net Switch
Arista	OSFP-800G-2XDR4	Arista 800GBASE-2XDR4 OSFP Transceivers ; 2x 400GBASE-XDR4 Transceiver, Dual MPO-12 connector, 2km over parallel SMF	128	L1 BE Net Transceivers
FS	69008	2m MTP® Jumper, MTP®-12 APC (Female) to MTP®-12 APC (Female), 12 Fibers, Single Mode (OS2)	256	L1 Be Net Fiber Cables
Arista	DCS-7260CX3-64E-F	Arista 7260X3, 64x100GbE QSFP & 2xSFP+ Enhanced switch, front-to-rear air, 2xAC	4	FE Net Switch
Arista	MMA1B00-C100D	Arista Networks QSFP-100G-SR4 Compatible QSFP28 MPO-12/UPC MMF Optical Transceiver Module, Support 4 x 25G-SR	64	FE Net Switch Transceivers
Broadcom	BCM957608-P2200GQF00	Broadcom P2100G - 2 x 100GbE PCIe NIC	64	FE Net NICs
FS.com	QSFP-100G-SR4	Broadcom compatible 100GBASE-SR4 QSFP100 Transceiver	64	FE Net NIC Transceivers
FS.com	12FMTPOM4	MTP® Jumper, MTP®-12 UPC (Female) to MTP®-12 UPC (Female), 12 Fibers, Multimode (OM4)	64	FE Net fiber Cables
Arista	DCS-7010TX-48-F	Arista 7010TX, 48x 10/100/1000 RJ45 & 4 x 25G SFP (1/10/25GbE) switch, front to rear air, 2xAC ; 3yr cloudvision ; 3yr NBD HW replacement	2	OOB MGMT Switch
FS.com	C6UTPSGSPVC	Cat6 28AWG Snagless Unshielded (UTP) PVC CM Slim Ethernet Network Patch Cable	128	OOB MGMT Cat 6 Cables
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	2	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	4	Data Center Rack PDUs

Rack #14, #32 - L1 FE Network Racks

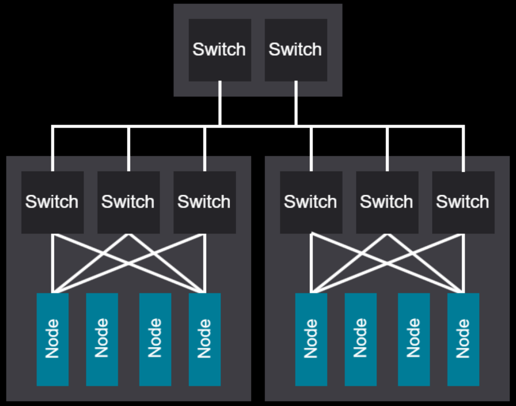
MFG	MFG Part No / SKU	Material Description	Total Qty 2 Racks	Notes
Arista	DCS-7260CX3-64E-F	Arista 7260X3, 64x100GbE QSFP & 2xSFP+ Enhanced switch, front-to-rear air, 2xAC	4	FE Net Switch
Arista	MMA1B00-C100D	Arista Networks QSFP-100G-SR4 Compatible QSFP28 MPO-12/UPC MMF Optical Transceiver Module, Support 4 x 25G-SR	64	FE Net Switch Transceivers
Broadcom	BCM957608-P2200GQF00	Broadcom P2100G - 2 x 100GbE PCIe NIC	64	FE Net NICs
FS.com	QSFP-100G-SR4	Broadcom compatible 100GBASE-SR4 QSFP100 Transceiver	64	FE Net NIC Transceivers
FS.com	12FMTPOM4	MTP® Jumper, MTP®-12 UPC (Female) to MTP®-12 UPC (Female), 12 Fibers, Multimode (OM4)	64	FE Net fiber Cables
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	2	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	4	Data Center Rack PDUs

Rack #XX - L2 Spine Network Rack

MFG	MFG Part No / SKU	Material Description	Total Qty 1 Rack	Notes
Arista	DCS-7060X6-64PE-F	Arista 7060X6, 64 x 800GbE OSFP switch, front-to-rear air, 2xAC ; L3 license ; cloudvision 3yr ; NBD HW support	8	L2 BE Net Switch
Arista	OSFP-800G-2XDR4	Arista 800GBASE-2XDR4 OSFP Transceivers ; 2x 400GBASE-XDR4 Transceiver, Dual MPO-12 connector, 2km over parallel SMF	64	L2 BE Net Transceivers
FS	69008	2m MTP® Jumper, MTP®-12 APC (Female) to MTP®-12 APC (Female), 12 Fibers, Single Mode (OS2)	128	L2 Be Net Fiber Cables
Arista	DCS-7260CX3-64E-F	Arista 7260X3, 64x100GbE QSFP & 2xSFP+ Enhanced switch, front-to-rear air, 2xAC	2	FE Net Spine Switch
Arista	MMA1B00-C100D	Arista Networks QSFP-100G-SR4 Compatible QSFP28 MPO-12/UPC MMF Optical Transceiver Module, Support 4 x 25G-SR	64	FE Net Spine Switch Transceivers
FS.com	12FMTPOM4	MTP® Jumper, MTP®-12 UPC (Female) to MTP®-12 UPC (Female), 12 Fibers, Multimode (OM4)	64	FE Net fiber Cables
APC	AR3357SP	APC NetShelter SX, Server Rack Enclosure, 48U, Shock Packaging, 2000 lbs, Black, 2258H x 750W x 1200D mm	1	Data Center Racks
APC	APDU10452SM	APC by Schneider Electric NetShelter 42-Outlets PDU	2	Data Center Rack PDUs

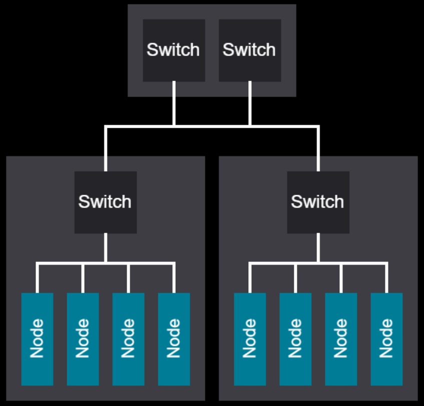
Cluster Network

Basic Network Topologies



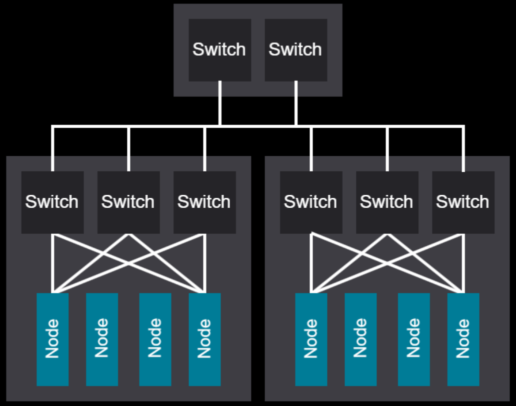
2-Tier Rail Network

- Enables large scalable unit sizes for large jobs or replica sizes
- Efficient for workloads favoring ring-based collectives
- Higher infrastructure costs



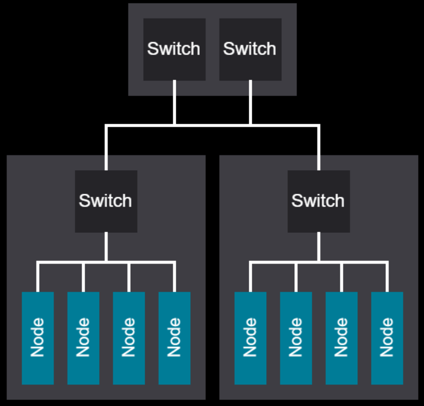
2-Tier Tree Network

- Efficient for small workloads or replicas
- Easy to add capacity with proper planning
- Potentially allows for lower infrastructure costs
- Limited blast radius compared to rail networks



3-Tier Rail TH5/J3 Network

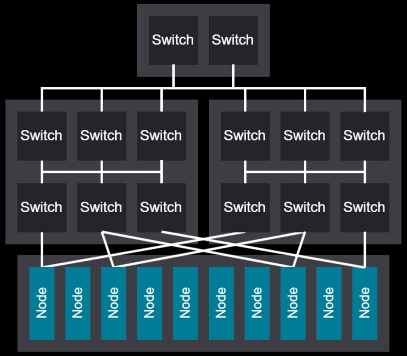
- Spine switches replaced with a 2-tier Jericho3-AI/Ramon3 fabric for increased maximum cluster size
- Deep buffers and scheduled fabric aid greatly with congestion issues in large clusters at a small latency penalty



3-Tier Tree TH5/J3 Network

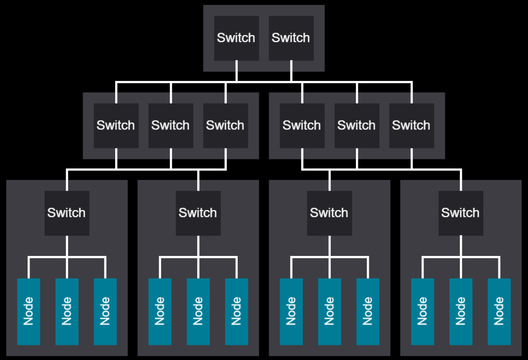
- Same benefits from switch to scheduled spine fabric as with rail, but retains the primary characteristics of tree networks

Basic Network Topologies Continued


Switch Switch
Switch Switch Switch Switch Switch Switch
Switch Switch Switch Switch Switch Switch
Node Node Node Node Node Node Node Node Node Node

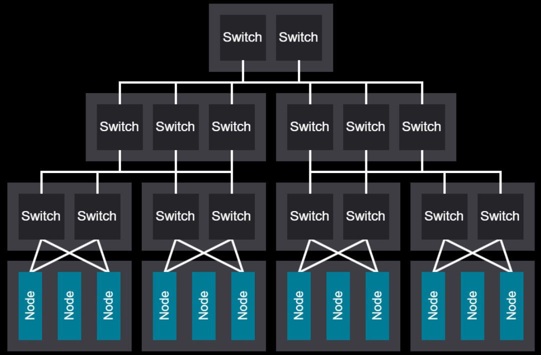
3-Tier Rail Optimized Network

- Allows for massive scalable unit sizes
- Best ring-based collective performance at scale; at the cost of poor any-any performance


Switch Switch
Switch Switch Switch Switch Switch Switch
Switch Switch Switch Switch
Node Node Node Node Node Node Node Node Node Node Node Node

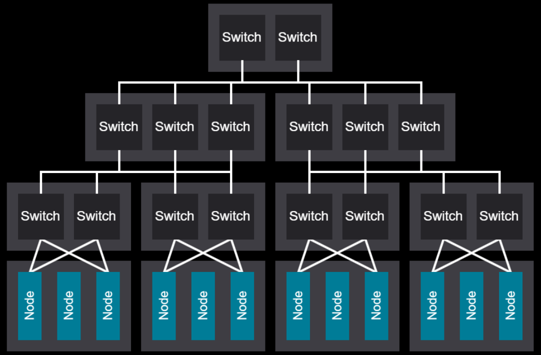
3-Tier Tree Network

- Allows for massive cluster sizes
- Best any-any performance at scale
- Suitable for a “campus style” deployment


Switch Switch
Switch Switch Switch Switch Switch Switch
Switch Switch Switch Switch Switch Switch Switch Switch
Node Node Node Node Node Node Node Node Node Node Node Node

3-Tier Hybrid Rail Network

- Allows for massive cluster sizes with large scalable units
- Favors ring-based collectives, but does not sacrifice significant any-any performance on large jobs
- Suitable for a “campus-style” deployment

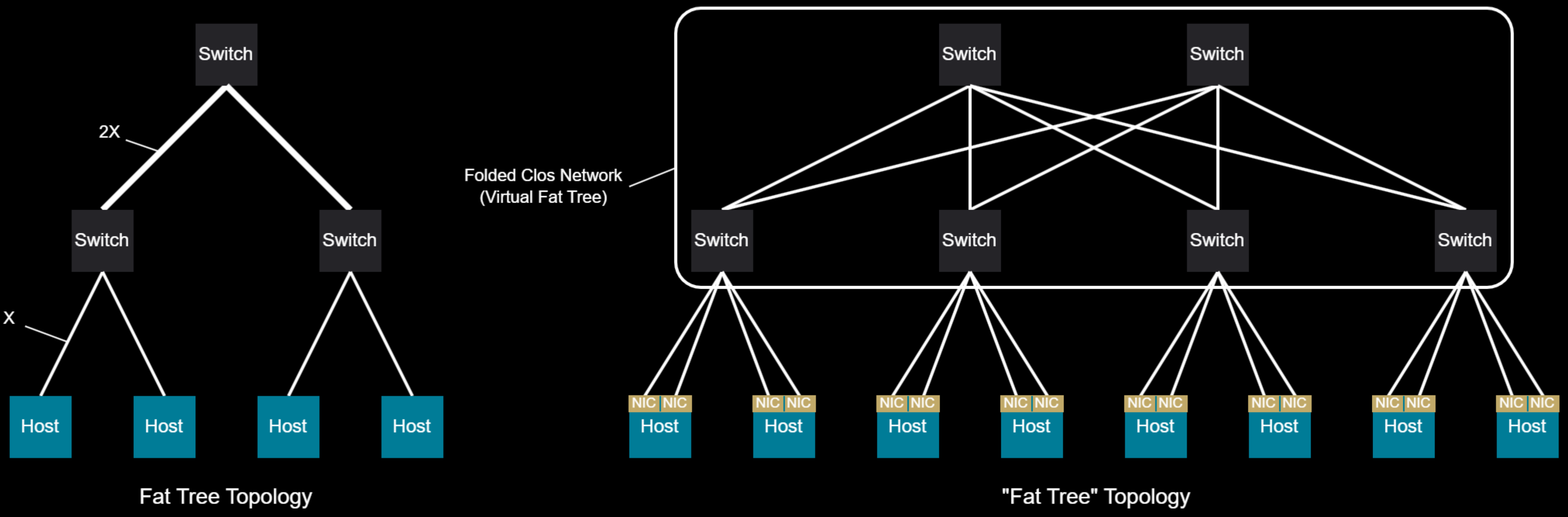

Switch Switch
Switch Switch Switch Switch Switch Switch
Switch Switch Switch Switch Switch Switch Switch Switch
Node Node Node Node Node Node Node Node Node Node Node Node

3-Tier Fully Scheduled Rail Network

- Medium sized scalable units
- Excellent congestion performance due to deep buffers and scheduled fabric
- Technical limitations limit cluster size to ~32K GPUs in recommended configuration

Tree and Rail

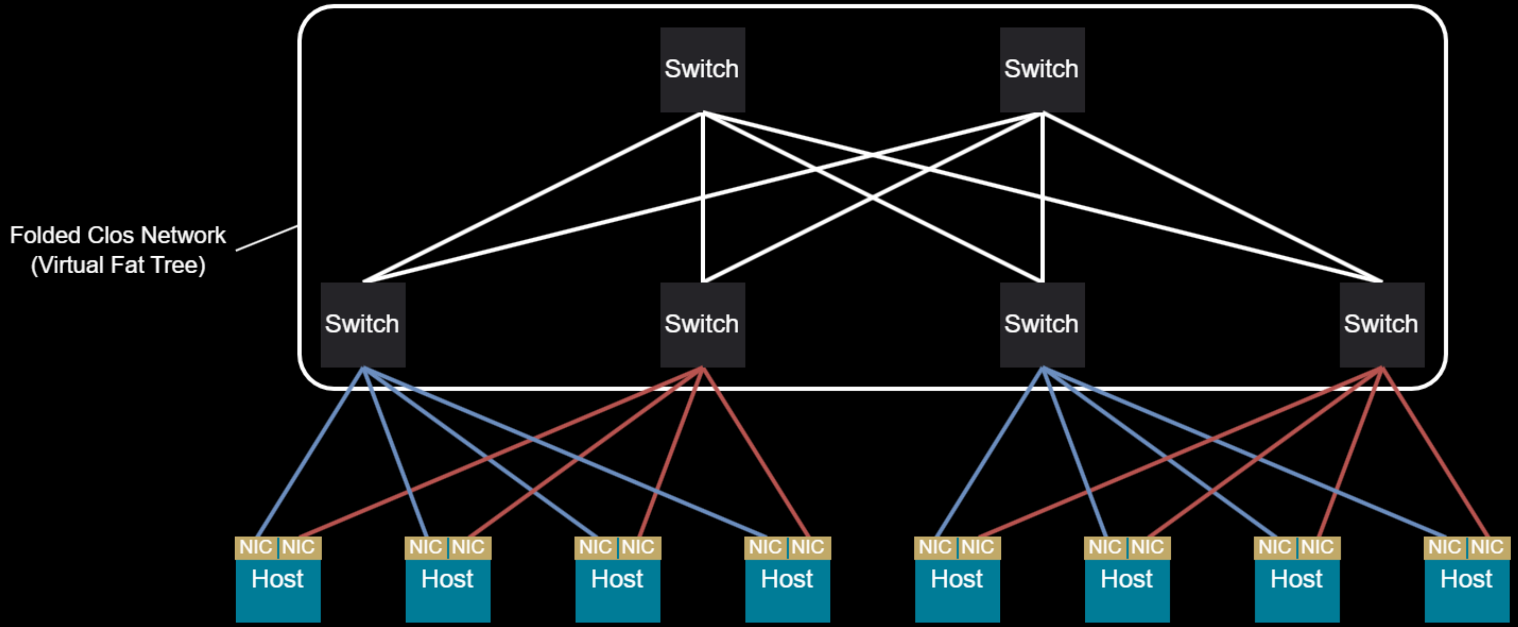
Fat Tree Terminology In Cluster Design



The canonical fat tree topology is a network concept where a switch's connection to upstream peers has at least parity bandwidth with the total aggregate bandwidth of its downstream connections. This causes links between switches to become "fatter" as they get closer to the core.

The "fat tree" topology for AI/ML clusters instead refers to how a host is connected to its upstream switches; in this case all host NICs terminate on the same switch. It can also be considered 1-rail network. The network itself is generally a 3-stage or 5-stage folded Clos network due the fixed radix of network switches.

Rail Networks In Cluster Design

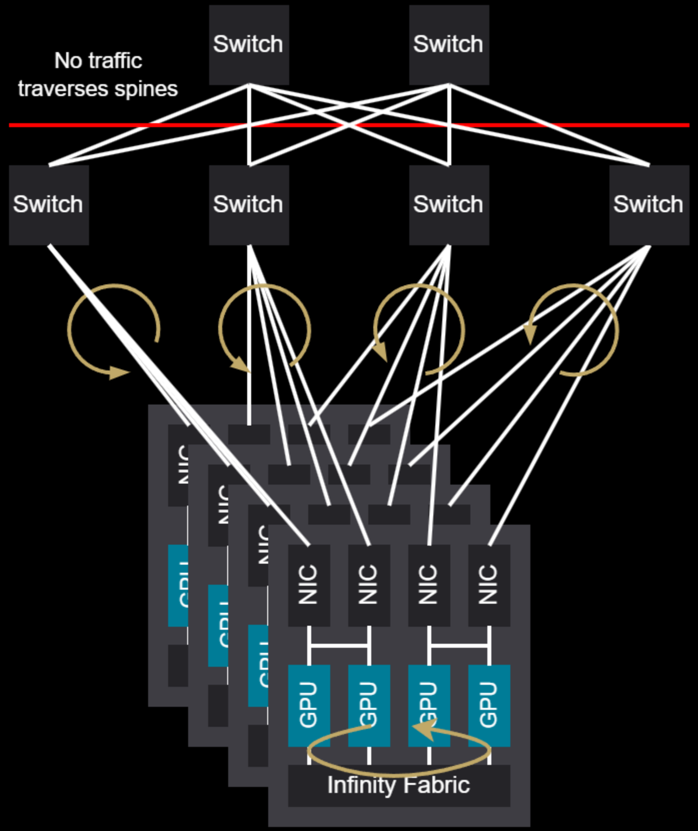


Rail Topology

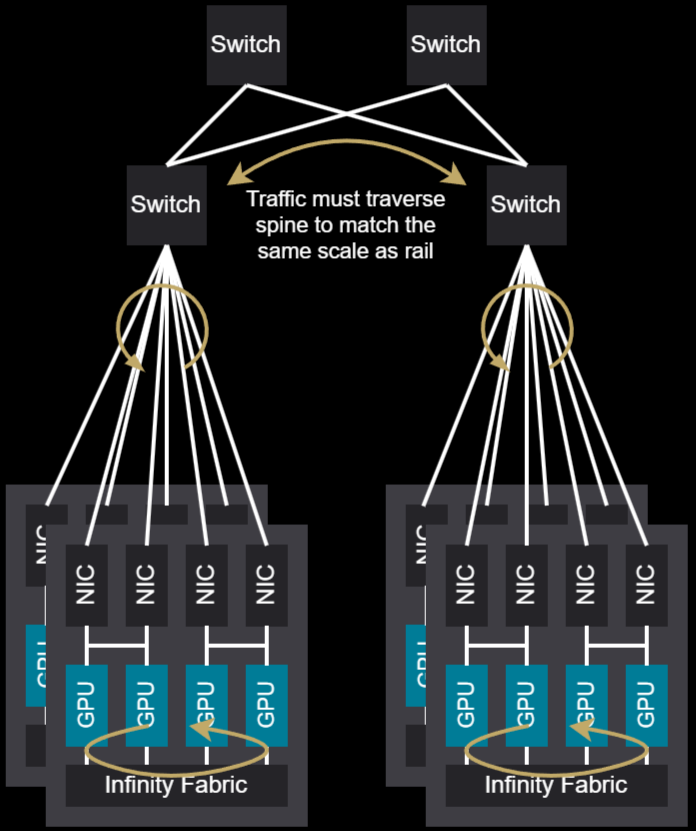
Rail networks leverage the same folded Clos network as tree networks, but host connections are instead aggregated onto switches based on NIC rank. These shared ranks are referred to as rails and allow the network to provide preferential latency for connections which share the same rail. The downside to this design is any traffic which needs to cross rails/ranks must traverse either the network spine layer, or Infinity Fabric (PXN).

The above example shows an example 2-rail network, with the rails colored blue and red to differentiate them.

Rail Enables Larger Single Hop Ring Domains

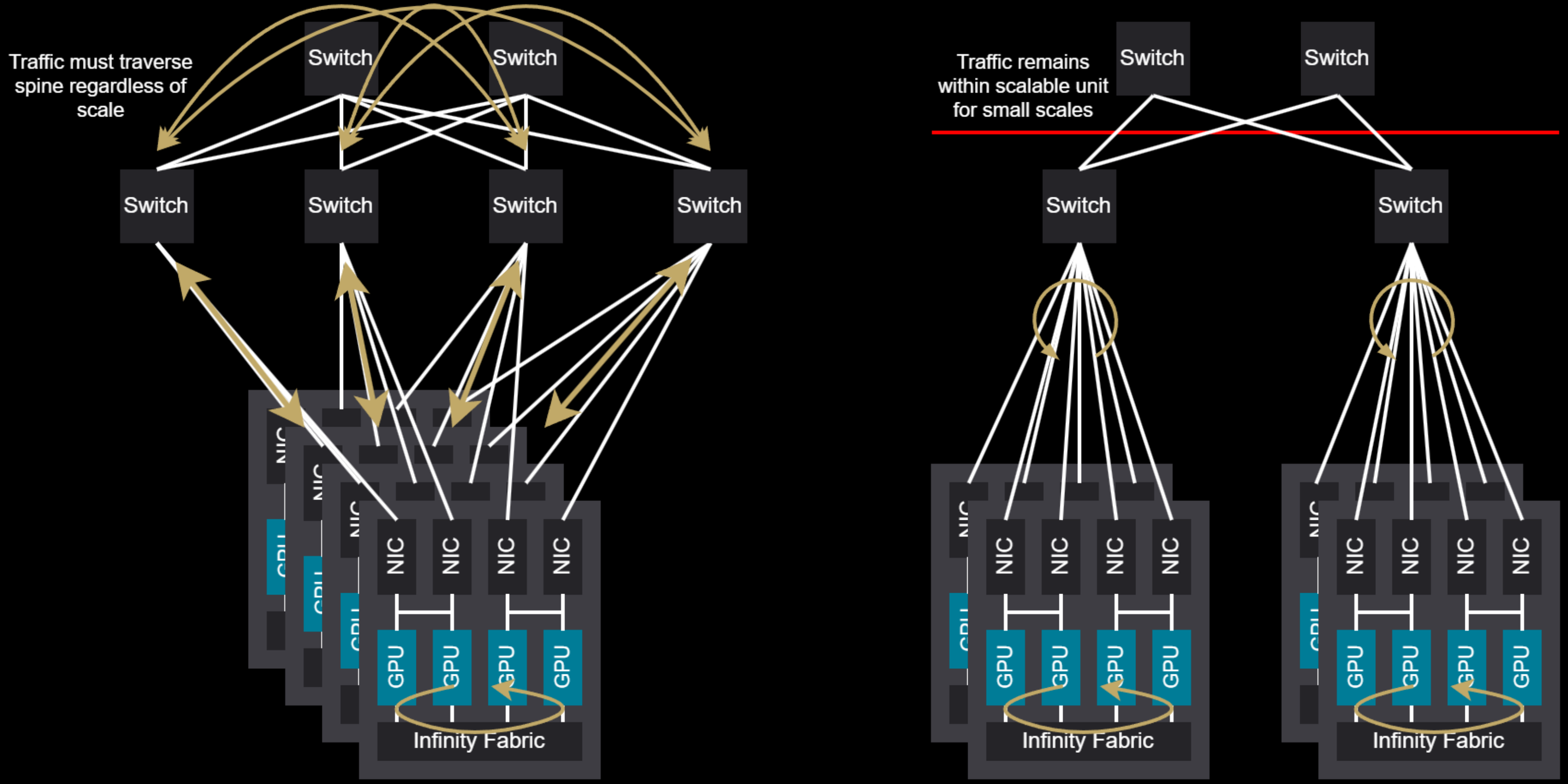


No traffic traverses spines



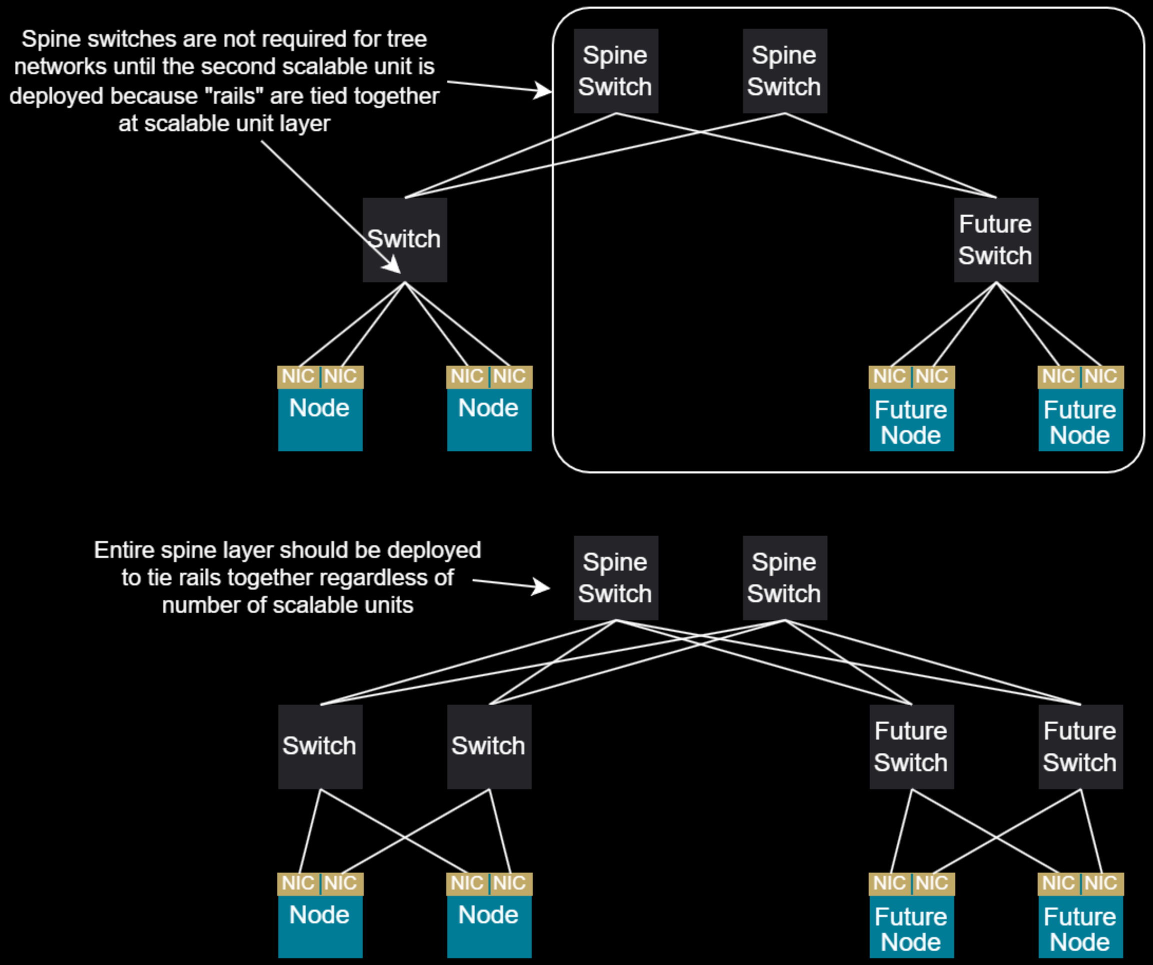
Traffic must traverse spine to match the same scale as rail

Tree Handles Cross-Rank Traffic Better



Scaling Networks

Cluster Backend Deployment Strategy for 2-tier Networks



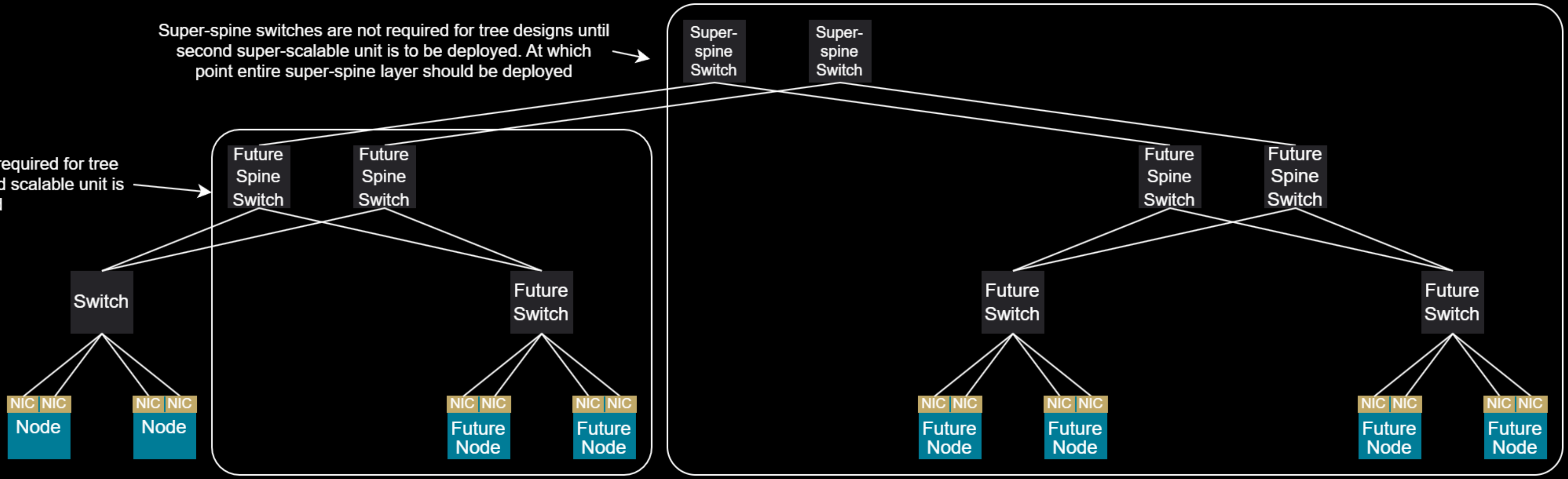
Spine switches are not required for tree networks until the second scalable unit is deployed because "rails" are tied together at scalable unit layer

Entire spine layer should be deployed to tie rails together regardless of number of scalable units

Cluster Backend Deployment Strategy for 3-tier Tree Networks

Super-spine switches are not required for tree designs until second super-scalable unit is to be deployed. At which point entire super-spine layer should be deployed

Spine switches are not required for tree networks until the second scalable unit is deployed

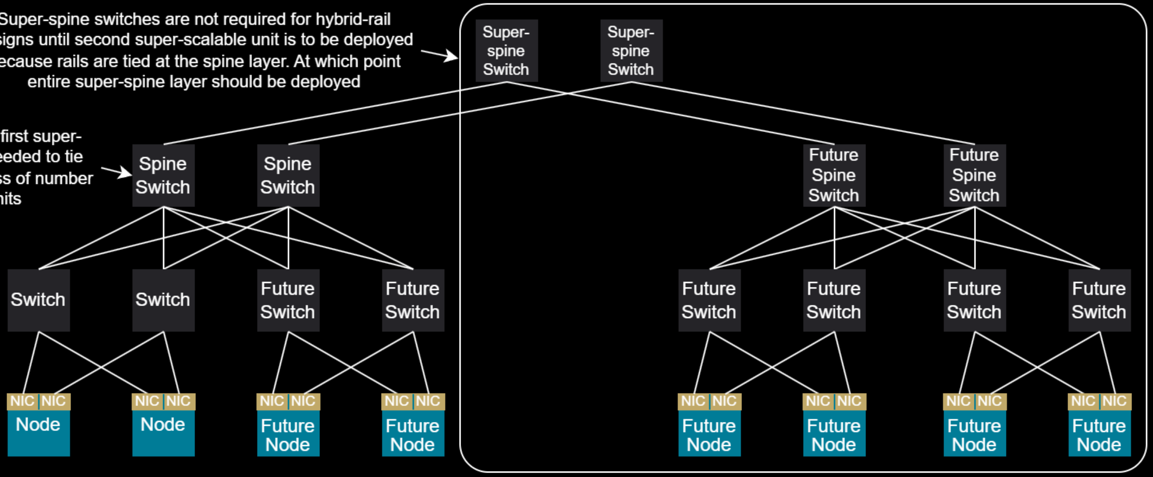


AMD
together we advance_

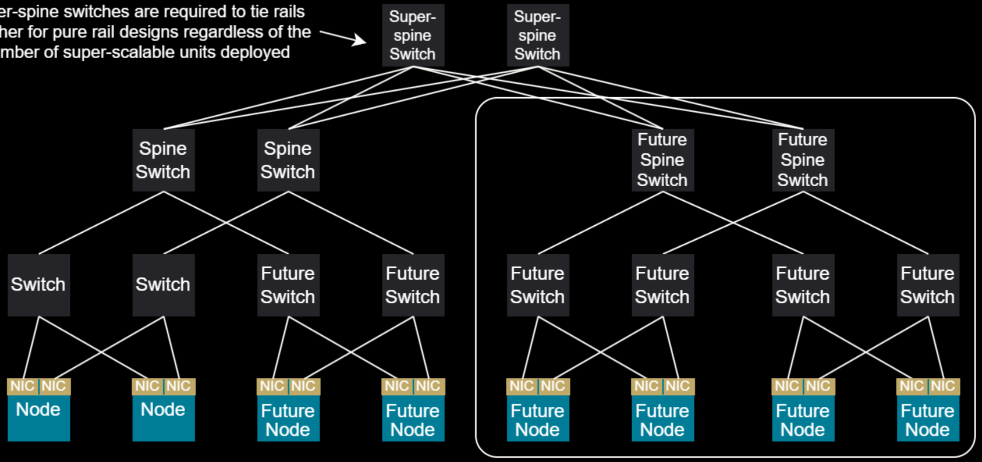
Cluster Backend Deployment Strategy for 3-tier Rail Networks

Super-spine switches are not required for hybrid-rail designs until second super-scalable unit is to be deployed because rails are tied at the spine layer. At which point entire super-spine layer should be deployed

Entire spine layer for first super-scalable unit is still needed to tie rails together regardless of number of scalable units



Super-spine switches are required to tie rails together for pure rail designs regardless of the number of super-scalable units deployed



Network Subscription

Subscription is the relationship between what is provided by the upstream network and what is required by the downstream network in demand side.

It is typically represented as a ratio:

$$Downstream\ Demand : Upstream\ Capacity$$

1.1 subscribed network would be equal downstream capacity to upstream capacity.

Example: 1:1.16 = (1 downstream demand : 1.16 upstream capacity. In this example there is .16 more upstream capacity)

Or as a percentage:

$$Subscription\ Rate = \frac{Downstream\ Demand}{Upstream\ Capacity}$$

80% subscription ratio could be referred to as “20% undersubscribed”, or a 120% subscription ratio could be referred to as “20% oversubscribed”.

Network Design Examples

Network Topology Design Notations:

Design note:

Designs included are based on either Jericho / Ramon switch type (Arista, Ciena, Nokia) or 51.2T switch type (Arista, Cisco, Dell, Juniper)

Vendors/switch models vary for port count and features – please consult desired vendor port count directly to confirm.

Scalable Units/PODs note:

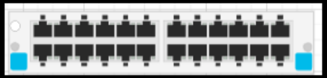
Diagrams presented are designed around a Scalable Unit or POD – which can determine overall network end to end latency and AI use cases.

Certain ML/AI workloads may require change of scalable unit size. Please consult with AMD Architecture as required.

Network Layout - 128 GPU Cluster 16 Nodes – Single Switch

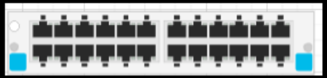
Example 128 port switch

64 uplink, 64 downlink



1:1 CFG

67 uplink, 56 downlink, 5 unused



16% Undersubscription

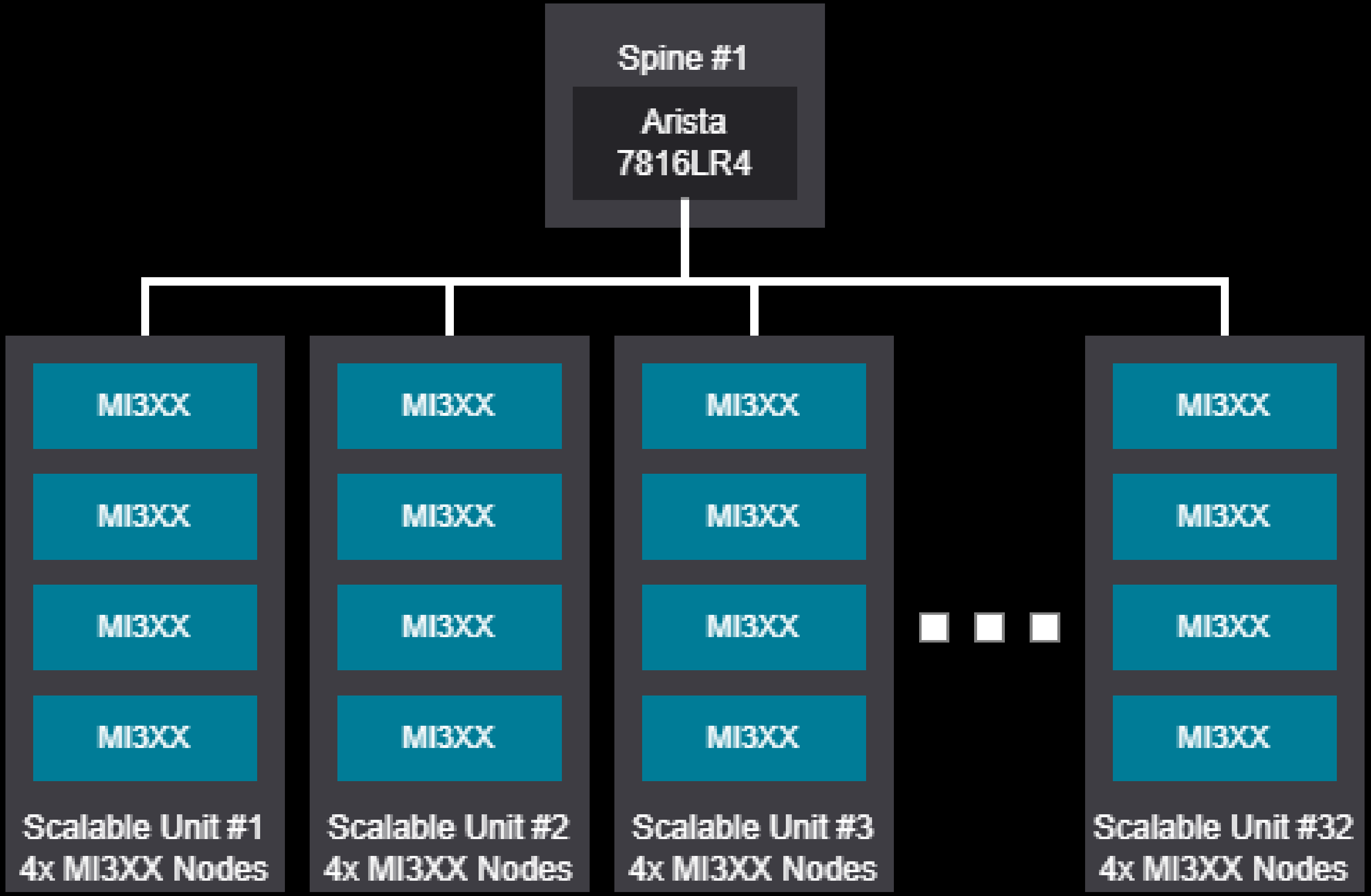
Fat Tree / Rail

Tier	Leaf switch	Spine switch	Superspine switch
3 tier	16%	1-1	All

Network Diagram - 128 GPU to 1024GPU (16 to 128 Nodes)

Tree Design

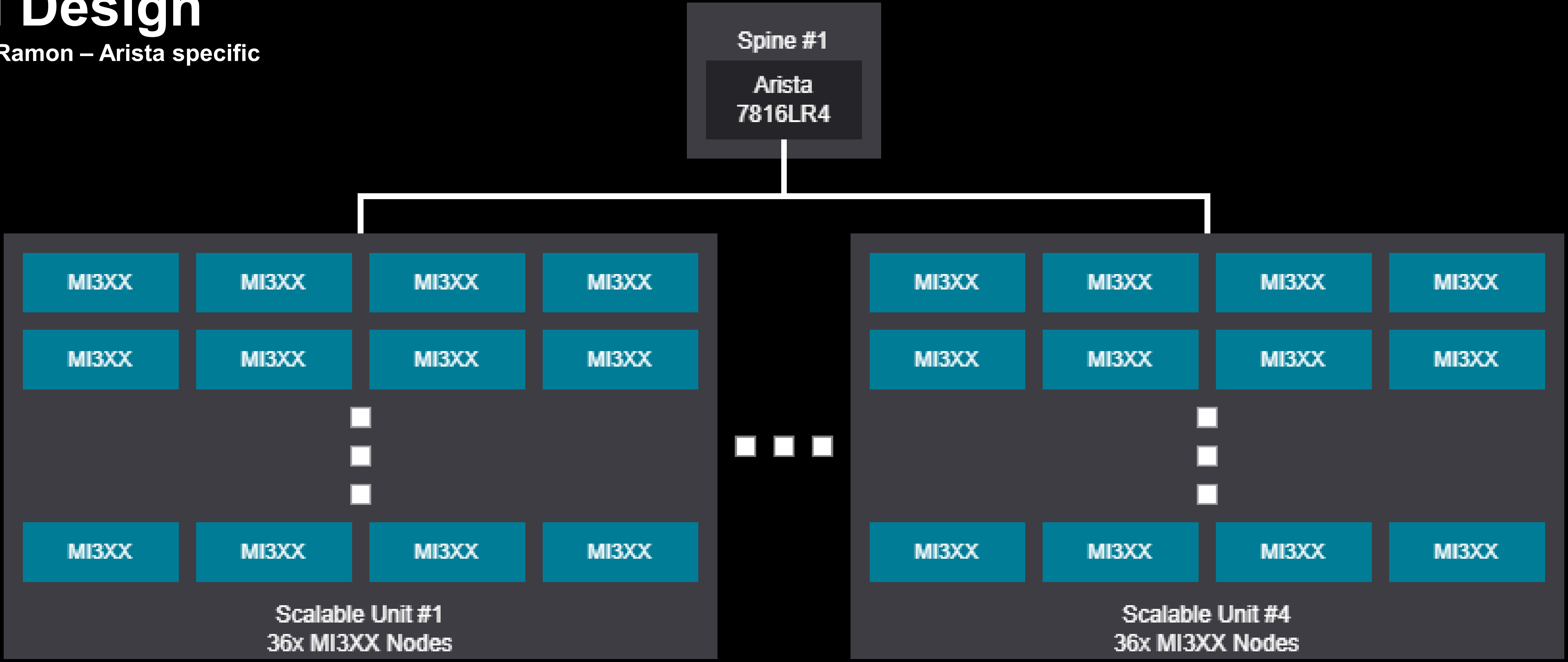
Jericho/Ramon – Arista specific



Network Diagram - 128 GPU to 1152GPU (16 to 144 Nodes)

Rail Design

Jericho/Ramon – Arista specific



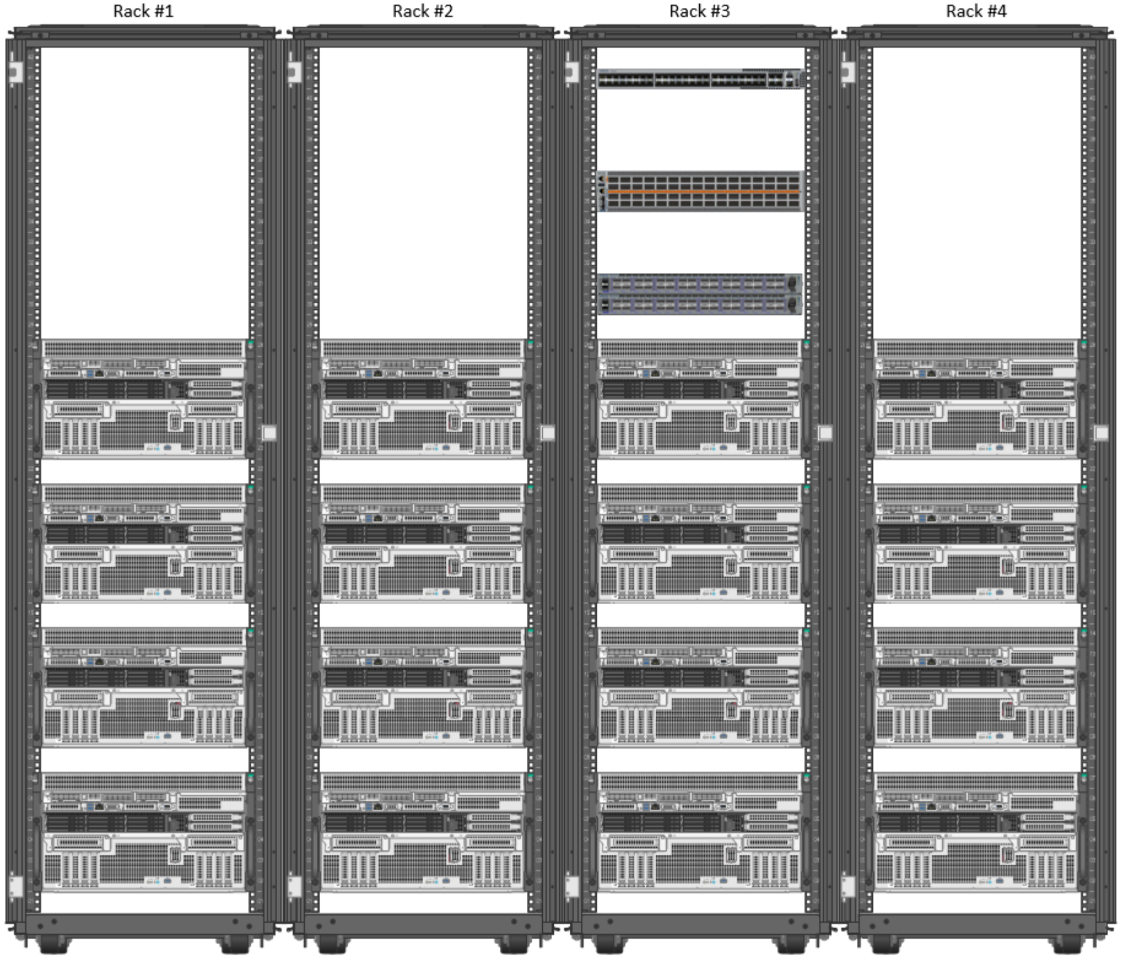
AMD
together we advance_

AMD

Appendix 1 – Vendor Specific Designs

HPE – Arista Design – Sample Racks (POD)

Sample Rack Elevation - 128 GPU (16 Nodes) - Single Switch



<u>Rack #1, #2, #4 – GPU Racks</u>
4 x HPE XD685 MI325 GPU Nodes
Rack Power Estimate: 52 KW

<u>Rack #3 – GPU + Network Rack</u>
4 x HPE XD685 MI325 GPU Nodes
1 x Arista TH5 800G 64p Switch – L1 BE Net
2 x Arista 32p 100G Switch – FE Net
1 x Arista 48p 1G Switch – OOB MGMT
Rack Power Estimate: 55 KW

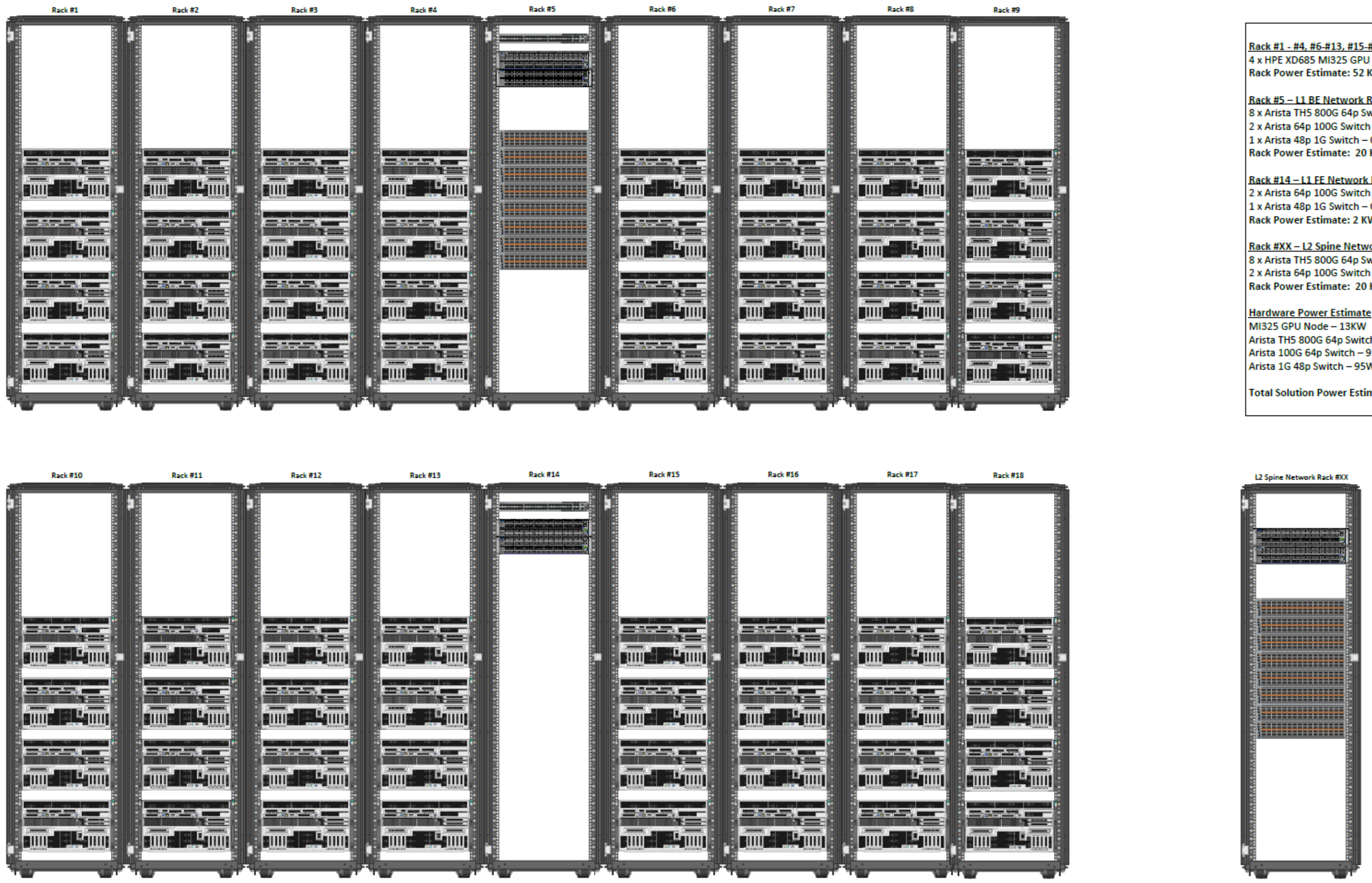
<u>Hardware Power Estimate</u>
MI325 GPU Node – 13KW
Arista TH5 800G 64p Switch – 2250W
Arista 100G 632p Switch – 350W
Arista 1G 48p Switch – 95W

Total Solution Power Estimate: 211 KW

128 GPU MI325X Cluster

	INTL	DATE	AMD
DRAWN BY:	BGM	03/01/2025	
APPROVED BY:	---	--/--/----	
SFDC ID: OPPORTUNITY #		SHEET OF N	

Sample Rack Elevation - 512GPU (64 Node) POD- Rail-Optimized



<u>Rack #1 - #4, #6-#13, #15-#18 – GPU Racks</u>
4 x HPE XD685 MI325 GPU Nodes
Rack Power Estimate: 52 KW

<u>Rack #5 – L1 BE Network Rack</u>
8 x Arista TH5 800G 64p Switch – L1 BE Net
2 x Arista 64p 100G Switch – FE Net
1 x Arista 48p 1G Switch – OOB MGMT
Rack Power Estimate: 20 KW

<u>Rack #14 – L1 FE Network Rack</u>
2 x Arista 64p 100G Switch – FE Net
1 x Arista 48p 1G Switch – OOB MGMT
Rack Power Estimate: 2 KW

<u>Rack #XX – L2 Spine Network Rack</u>
8 x Arista TH5 800G 64p Switch – L2 BE Net
2 x Arista 64p 100G Switch – FE Net Spine
Rack Power Estimate: 20 KW

<u>Hardware Power Estimate</u>
MI325 GPU Node – 13KW
Arista TH5 800G 64p Switch – 2250W
Arista 100G 64p Switch – 927W
Arista 1G 48p Switch – 95W

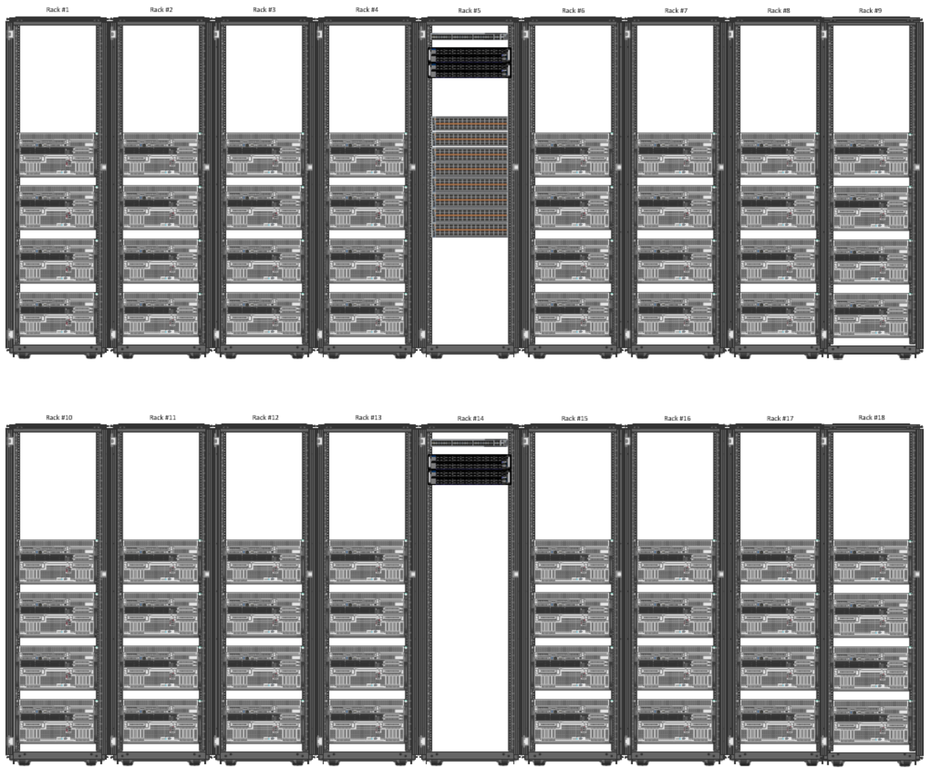
Total Solution Power Estimate: 888 KW

512 GPU MI325X Cluster POD			
	INTL	DATE	AMD
DRAWN BY:	BGM	1/10/2025	
APPROVED BY:	---	--/--/----	
SFDC ID: OPPORTUNITY #		SHEET OF N	

AMD
together we advance_

Sample Rack Elevation - 1024GPU (128 Node) POD- Rail-Optimized

512 GPU – POD #1 – Rack #1-#18



Rack #1 - #4, #6-#13, #15-#22, #24-#31, #33-#36 – GPU Racks
4 x HPE XD685 MI325 GPU Nodes
Rack Power Estimate: 52 KW

Rack #5, #23 – L1 BE Network Rack
8 x Arista TH5 800G 64p Switch – L1 BE Net
2 x Arista 64p 100G Switch – FE Net
1 x Arista 48p 1G Switch – OOB MGMT
Rack Power Estimate: 20 KW

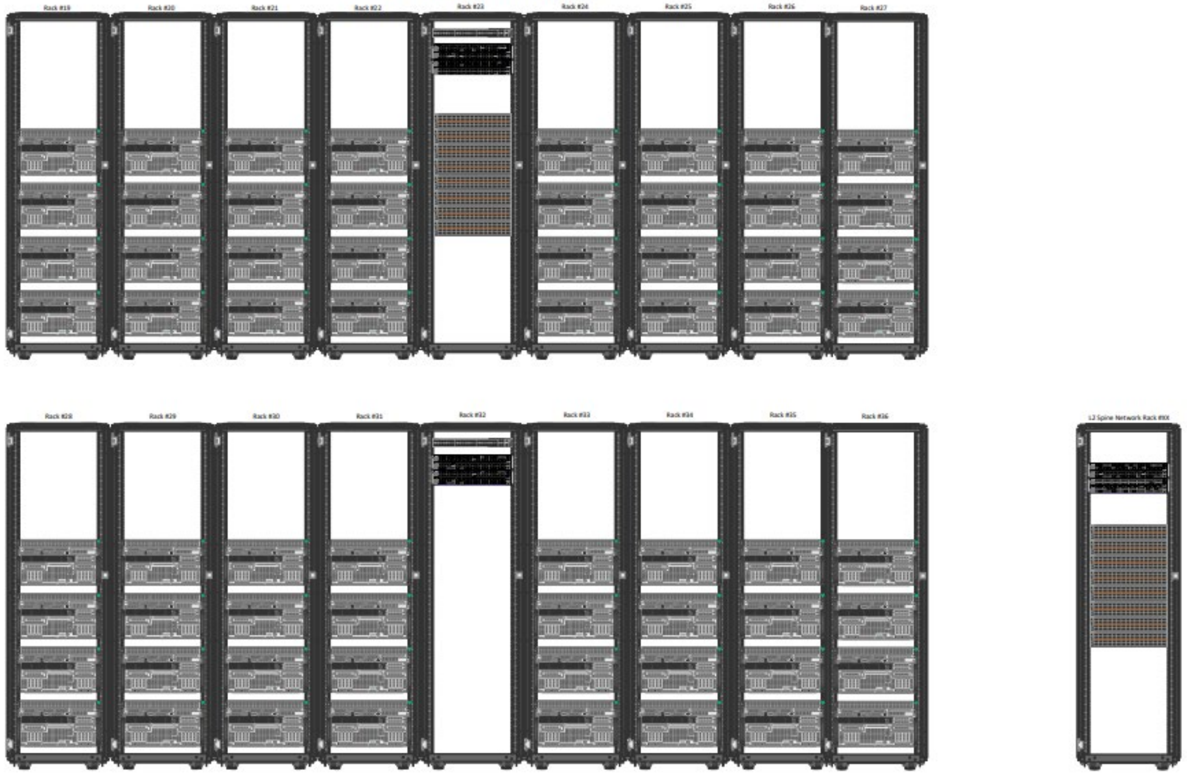
Rack #14, #32 – L1 FE Network Rack
2 x Arista 64p 100G Switch – FE Net
1 x Arista 48p 1G Switch – OOB MGMT
Rack Power Estimate: 2 KW

Rack #XX – L2 Spine Network Rack
8 x Arista TH5 800G 64p Switch – L2 BE Net
2 x Arista 64p 100G Switch – FE Net Spine
Rack Power Estimate: 20 KW

Hardware Power Estimate
MI325 GPU Node – 13KW
Arista TH5 800G 64p Switch – 2250W
Arista 100G 64p Switch – 927W
Arista 1G 48p Switch – 95W

Total Solution Power Estimate: 1,704 KW

512 GPU – POD #2 – Rack #19-#36



1024 GPU MI325X Cluster POD

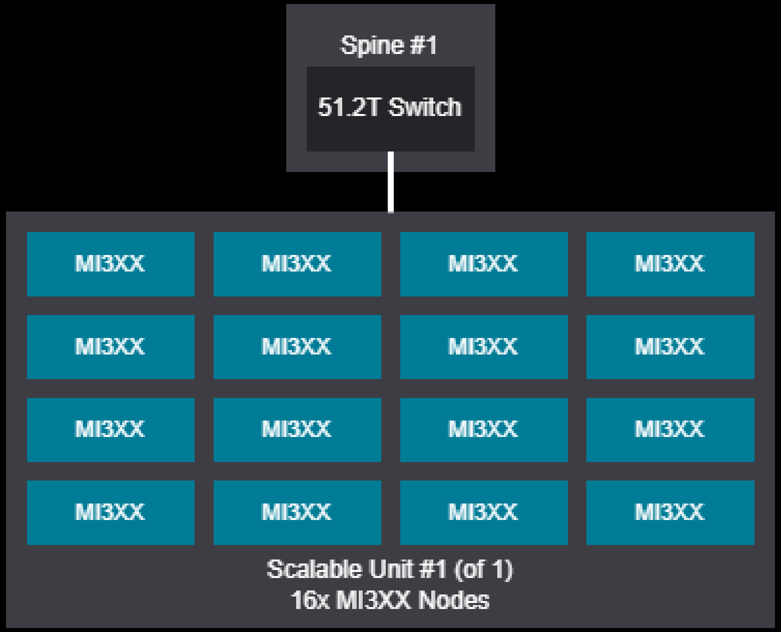
	INTL	DATE	
DRAWN BY:	[illegible]	[illegible]	AMD
APPROVED BY:	--	--/--/--	
SFDC ID: OPPORTUNITY #	SHEET	OF N	

AMD
together we advance_

Appendix 2 - Topological Variations

128 GPU Generic Topology Design Examples

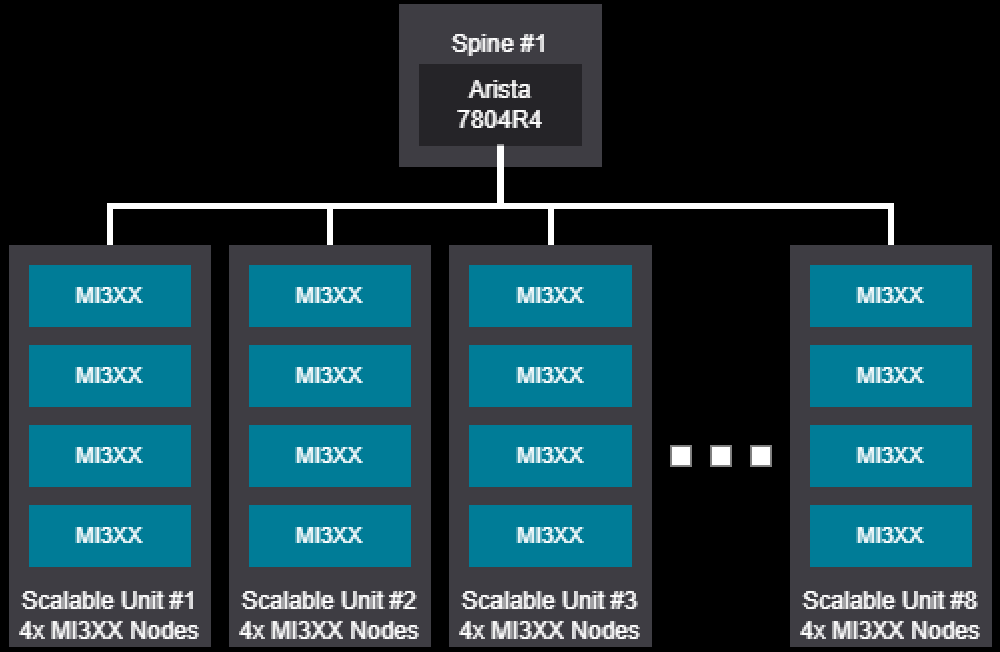
8-128 GPU Single Switch Design



- Can use Cisco G200 or Broadcom Tomahawk 5
- Single switch design will give the best performance possible compared to any alternative
- Extremely simple configuration and deployment

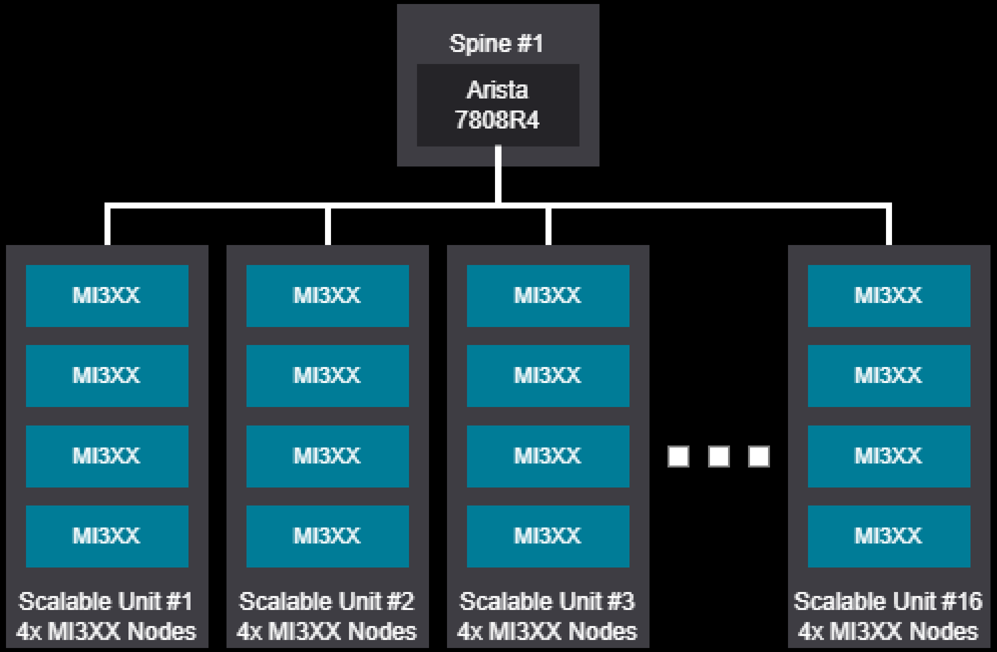
129-1024 GPU Tree Designs – Jericho/Ramon Specific

129-256 GPU Tree Design


Spine #1
Arista 7804R4
MI3XX
Scalable Unit #1 4x MI3XX Nodes
Scalable Unit #2 4x MI3XX Nodes
Scalable Unit #3 4x MI3XX Nodes
Scalable Unit #8 4x MI3XX Nodes

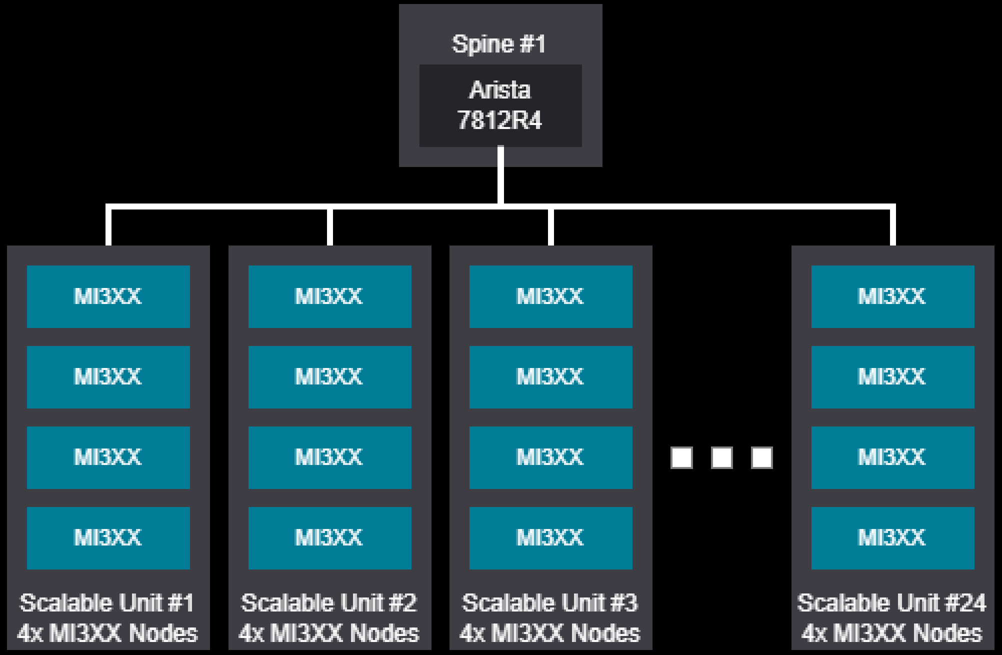
- Built around the Arista 7800R4 Jericho3-AI/Ramon3 series chassis switch
- Functionally a 2-tier network, but managed as a single switch
- Fully scheduled fabric
- Extremely simple configuration and deployment
- Tree vs rail topology is accomplished by adjusting which ports a node is connected to the switch
- Tree can scale to a slightly smaller cluster than rail

257-512 GPU Tree Design


Spine #1
Arista 7808R4
MI3XX
MI3XX
MI3XX
MI3XX
Scalable Unit #1 4x MI3XX Nodes
MI3XX
MI3XX
MI3XX
MI3XX
Scalable Unit #2 4x MI3XX Nodes
MI3XX
MI3XX
MI3XX
MI3XX
Scalable Unit #3 4x MI3XX Nodes
MI3XX
MI3XX
MI3XX
MI3XX
Scalable Unit #16 4x MI3XX Nodes

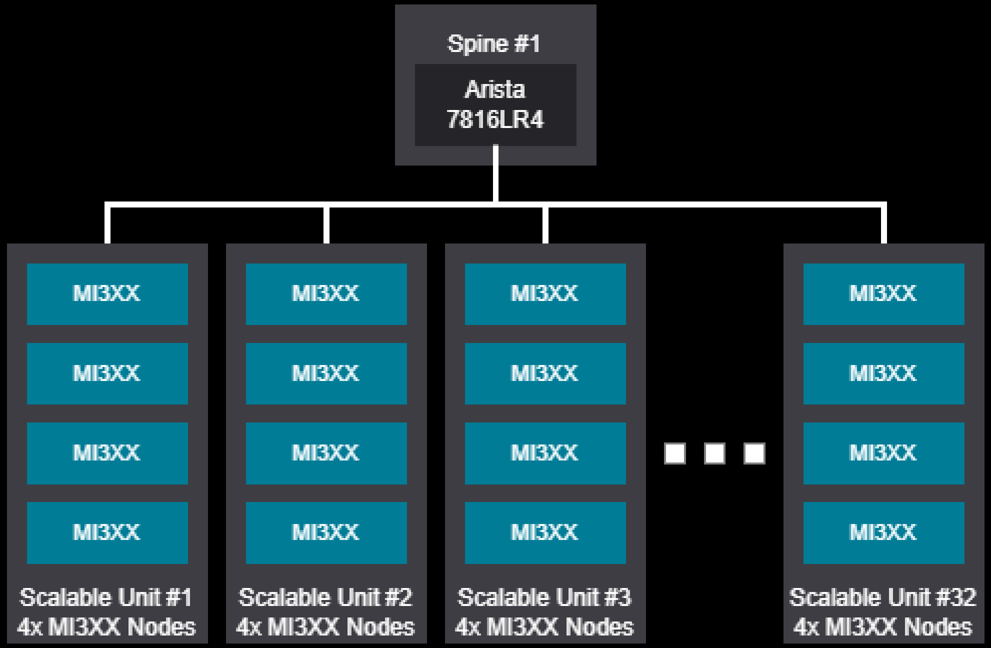
- Built around the Arista 7800R4 Jericho3-AI/Ramon3 series chassis switch
- Functionally a 2-tier network, but managed as a single switch
- Fully scheduled fabric
- Extremely simple configuration and deployment
- Tree vs rail topology is accomplished by adjusting which ports a node is connected to the switch
- Tree can scale to a slightly smaller cluster than rail

513-768 GPU Tree Design


Spine #1
Arista 7812R4
MI3XX
Scalable Unit #1 4x MI3XX Nodes
Scalable Unit #2 4x MI3XX Nodes
Scalable Unit #3 4x MI3XX Nodes
Scalable Unit #24 4x MI3XX Nodes

- Built around the Arista 7800R4 Jericho3-AI/Ramon3 series chassis switch
- Functionally a 2-tier network, but managed as a single switch
- Fully scheduled fabric
- Extremely simple configuration and deployment
- Tree vs rail topology is accomplished by adjusting which ports a node is connected to the switch
- Tree can scale to a slightly smaller cluster than rail

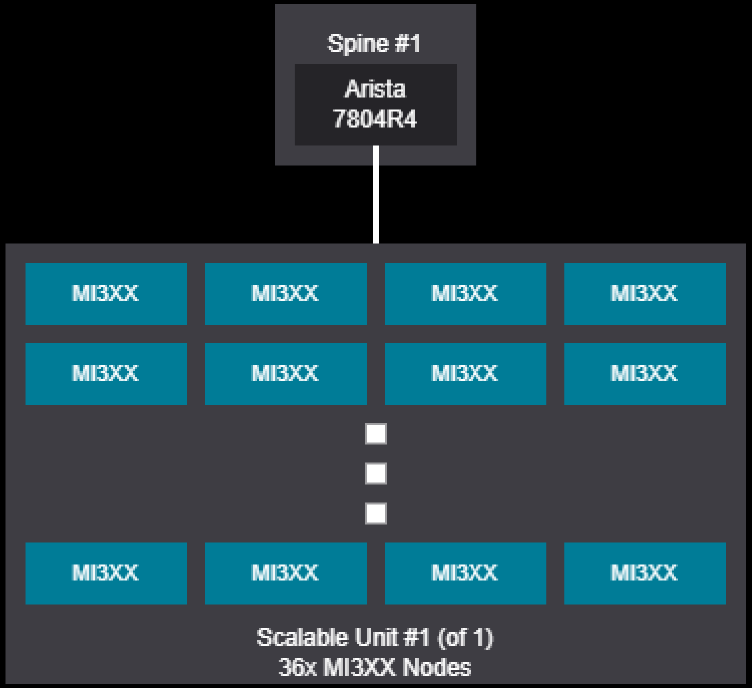
769-1024 GPU Tree Design


Spine #1
Arista 7816LR4
MI3XX
Scalable Unit #1 4x MI3XX Nodes
Scalable Unit #2 4x MI3XX Nodes
Scalable Unit #3 4x MI3XX Nodes
Scalable Unit #32 4x MI3XX Nodes

- Built around the Arista 7800R4 Jericho3-AI/Ramon3 series chassis switch
- Functionally a 2-tier network, but managed as a single switch
- Fully scheduled fabric
- Extremely simple configuration and deployment
- Tree vs rail topology is accomplished by adjusting which ports a node is connected to the switch
- Tree can scale to a slightly smaller cluster than rail

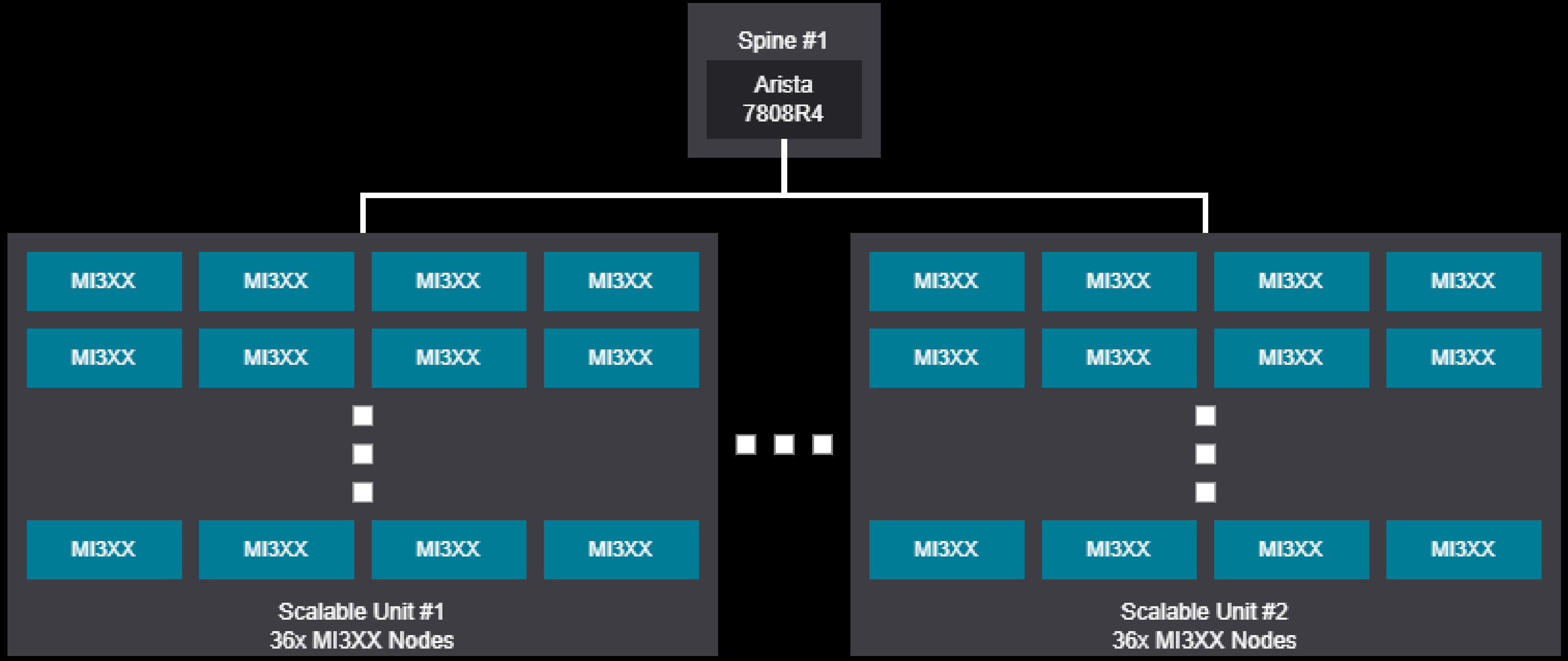
129-1152 GPU Rail Designs – Jericho/Ramon Specific

129-288 GPU Rail Design



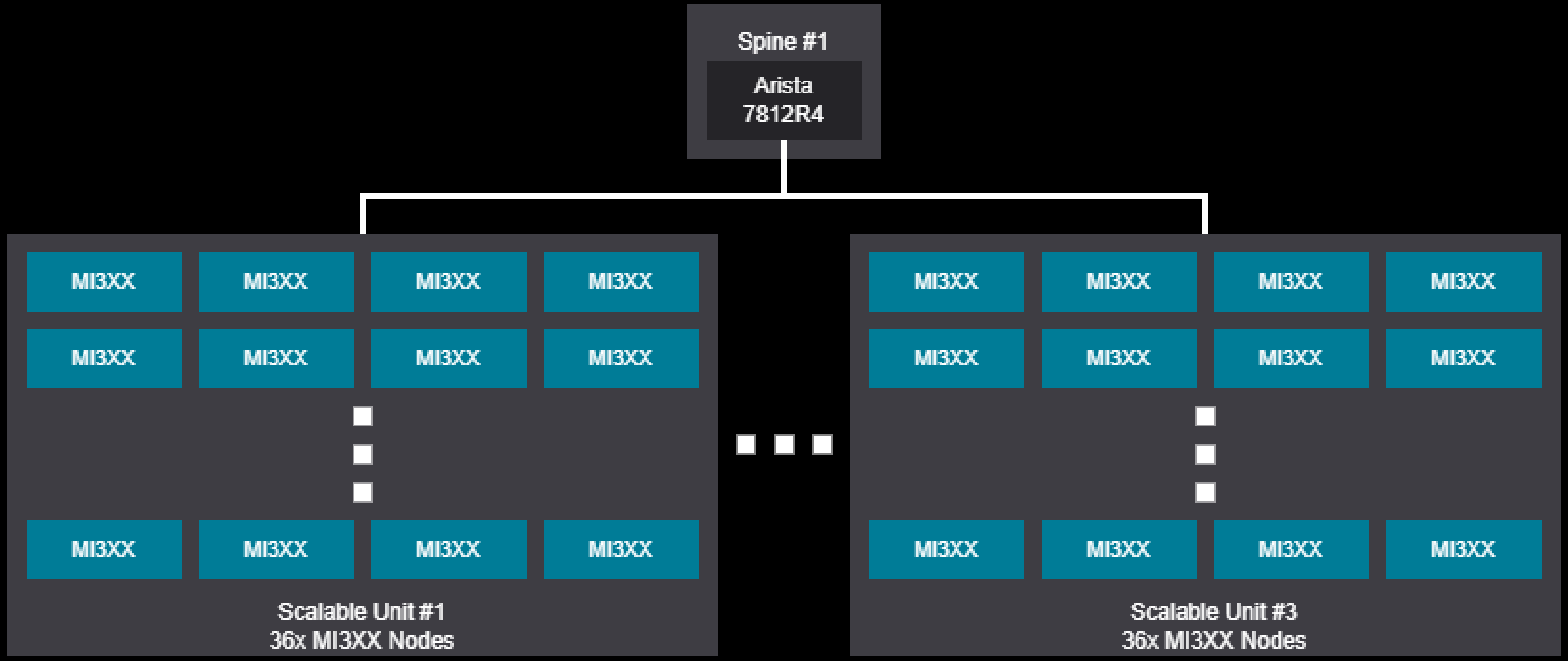
- Built around the Arista 7800R4 Jericho3-AI/Ramon3 series chassis switch
- Functionally a 2-tier network, but managed as a single switch
- Fully scheduled fabric
- Extremely simple configuration and deployment
- Tree vs rail topology is accomplished by adjusting which ports a node is connected to the switch
- Rail can scale to a slightly larger cluster than tree

289-576 GPU Rail Design



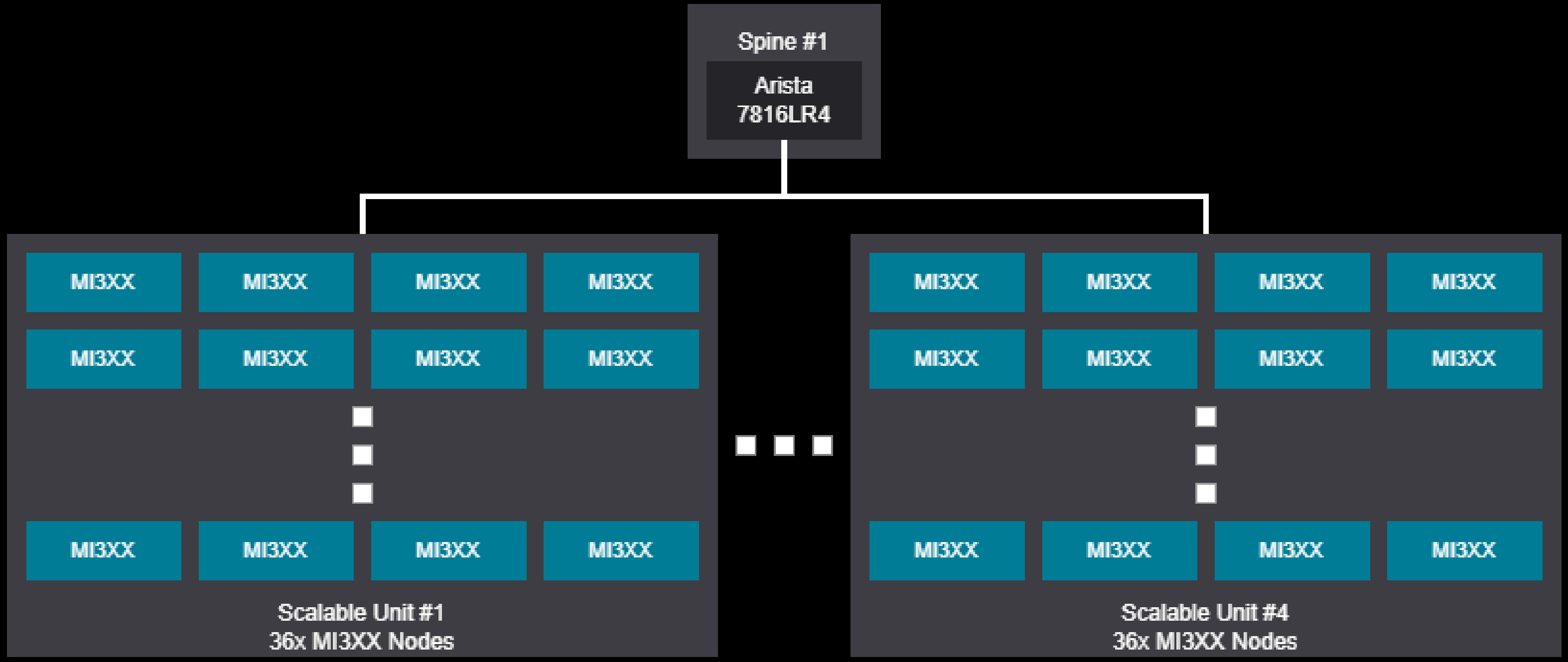
- Built around the Arista 7800R4 Jericho3-AI/Ramon3 series chassis switch
- Functionally a 2-tier network, but managed as a single switch
- Fully scheduled fabric
- Extremely simple configuration and deployment
- Tree vs rail topology is accomplished by adjusting which ports a node is connected to the switch
- Rail can scale to a slightly larger cluster than tree

577-864 GPU Rail Design



- Built around the Arista 7800R4 Jericho3-AI/Ramon3 series chassis switch
- Functionally a 2-tier network, but managed as a single switch
- Fully scheduled fabric
- Extremely simple configuration and deployment
- Tree vs rail topology is accomplished by adjusting which ports a node is connected to the switch
- Rail can scale to a slightly larger cluster than tree

865-1152 GPU Rail Design



- Built around the Arista 7800R4 Jericho3-AI/Ramon3 series chassis switch
- Functionally a 2-tier network, but managed as a single switch
- Fully scheduled fabric
- Extremely simple configuration and deployment
- Tree vs rail topology is accomplished by adjusting which ports a node is connected to the switch
- Rail can scale to a slightly larger cluster than tree

DISCLAIMER AND ATTRIBUTIONS

DISCLAIMER

The information contained herein is for informational purposes only, and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD’s products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18

COMPLIANCE WITH LAWS

Customer shall adhere to all applicable export laws and regulations including, without limitation, those administered by the U.S. Department of Commerce – Bureau of Industry and Security (U.S. Export Administration Regulations 15 CFR 730 et seq.) and those administered by the U.S. Department of State in accordance with the U.S. International Traffic in Arms Regulations (ITAR) set forth in Subchapter M, Title 22, Code of Federal Regulations, Parts 120 through 130 (22 CFR 120-130), as the same may be amended from time to time, and shall not export, re-export, resell, transfer, or disclose, directly or indirectly, any Products or technical data, or the direct product of any Products or technical data, to any proscribed person, entity, or country, or foreign national thereof, unless properly authorized by the U.S. government and/or any other applicable or relevant government or regulatory body, including the export authorities of all respective countries. For the avoidance of doubt, Customer shall not use Products in, or re-export Products to Belarus, Russia and the Donetsk (DNR) or Luhansk (LNR) regions of Ukraine, regardless of the applicable export laws and regulations. Customer shall impose upon its customers terms at least as restrictive as those contained in this Clause 14 with respect to any sale, distribution or export of Products.

Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.