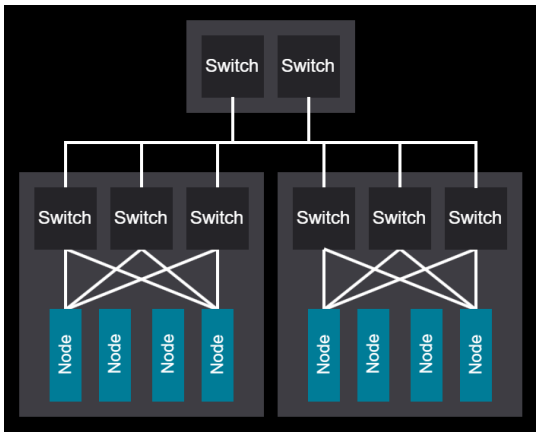
Reference Network Design Architecture MI3XX series

8K GPU

Version 4.1 – November 2025

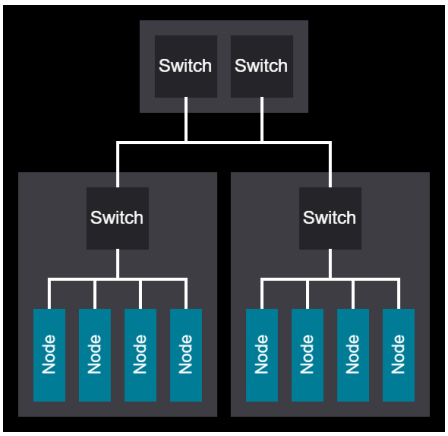
Basic network topologies

2-Tier rail network



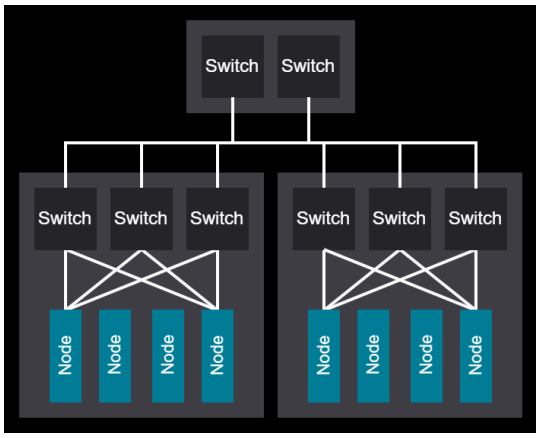
- Enables large scalable unit sizes for large jobs or replica sizes.
- Efficient for workloads favoring ring-based collectives.
- Higher infrastructure costs.

2-Tier tree network



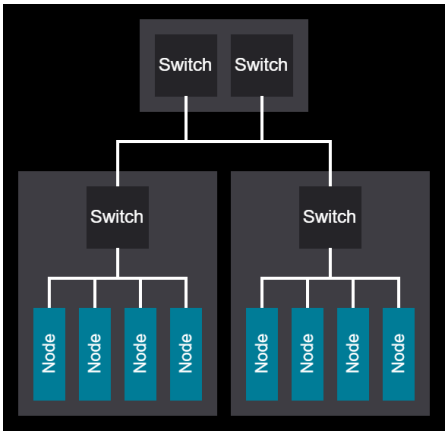
- Efficient for small workloads or replicas.
- Easy to add capacity with proper planning.
- Potentially allows for lower infrastructure costs.
- Limited blast radius compared to rail networks.

3-Tier rail TH5/J3 network



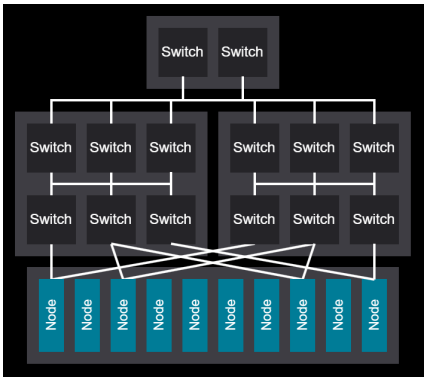
- Spine switches replaced with a 2-tier Jericho3-AI/Ramon3 fabric for increased maximum cluster size.
- Deep buffers and scheduled fabric aid greatly with congestion issues in large clusters at a small latency penalty.

3-Tier tree TH5/J3 network



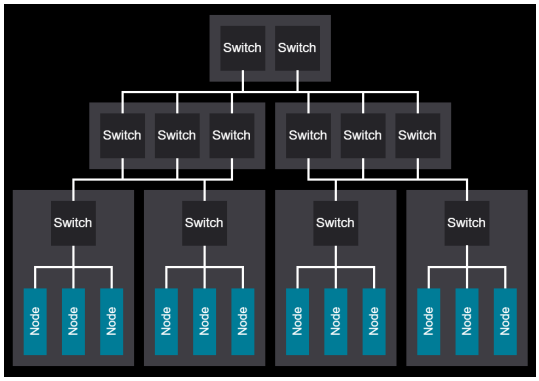
- Same benefits from switch to scheduled spine fabric as with rail, but retains the primary characteristics of tree networks.

3-Tier rail optimized network



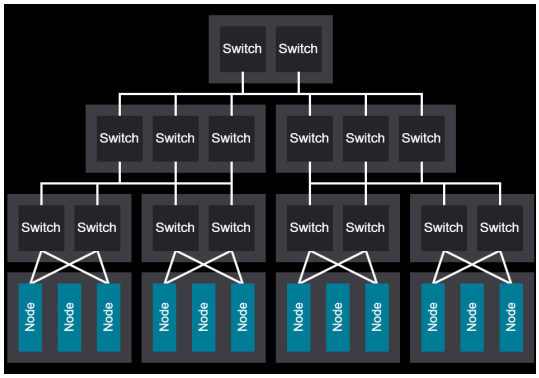
- Allows for massive scalable unit sizes.
- Best ring-based collective performance at scale; at the cost of poor any-any performance.

3-Tier tree network



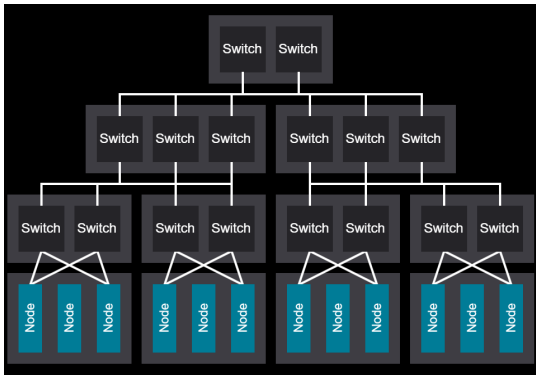
- Allows for massive cluster sizes.
- Best any-any performance at scale.
- Suitable for a “campus style” deployment.

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.

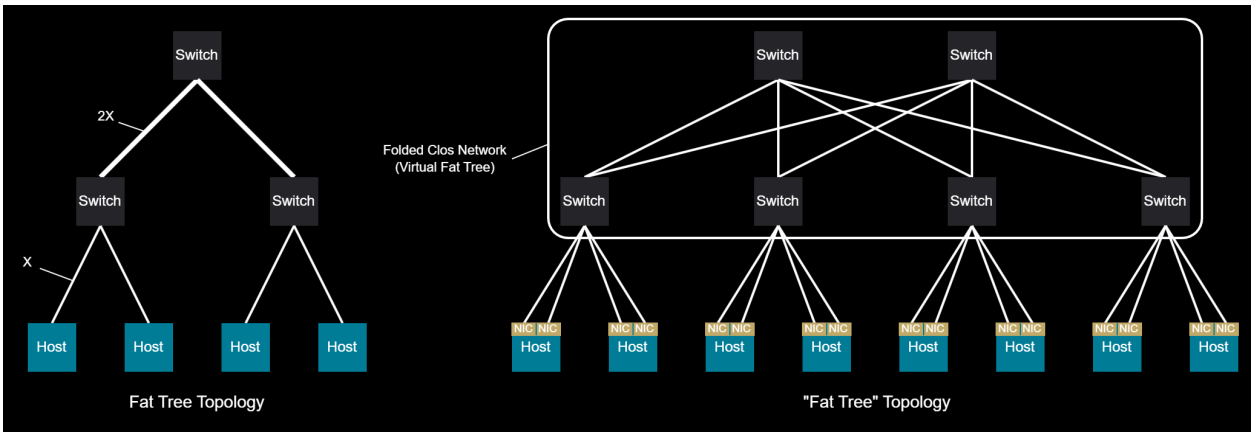
3-Tier fully scheduled rail network



- 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 designs

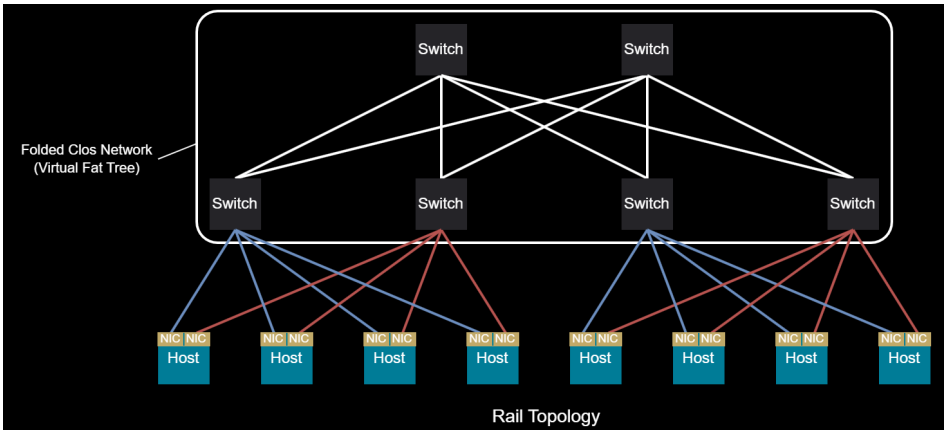
Fat tree networks



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

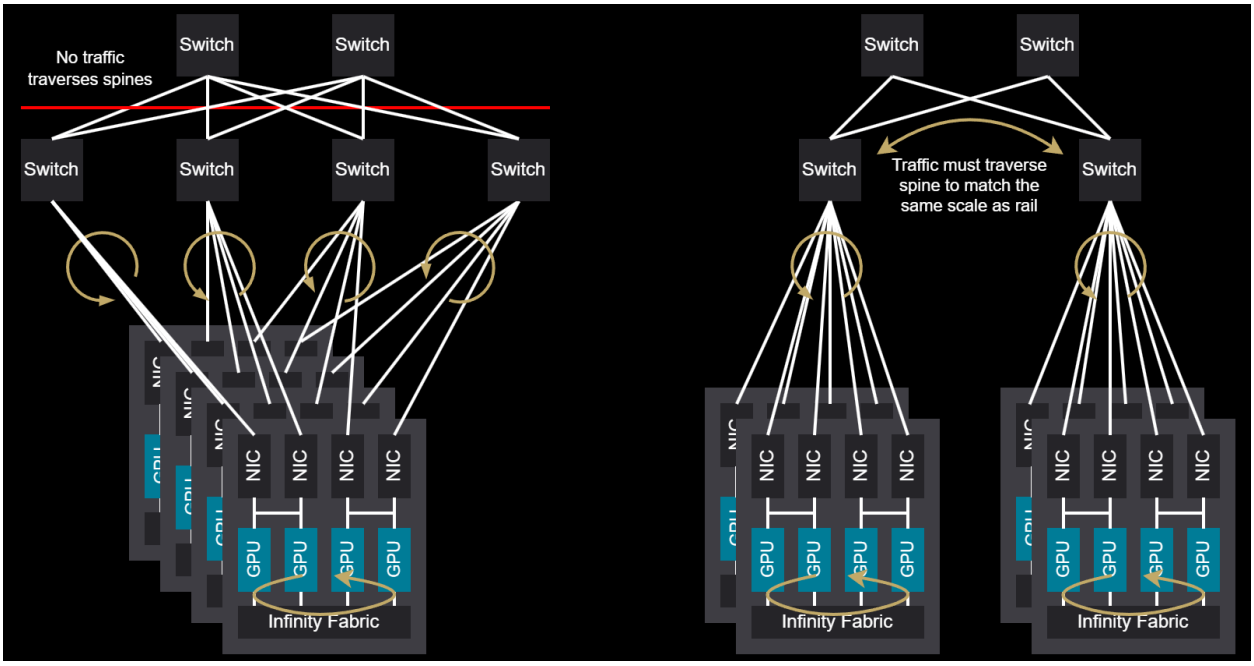


Rail networks leverage the same folded Clos network as tree networks, but host connections are instead aggregated onto switches based on NIC rank. These shared ranks are referred to as rails and allow the network to provide preferential latency for connections which share the same rail. The downside to this design is any traffic which

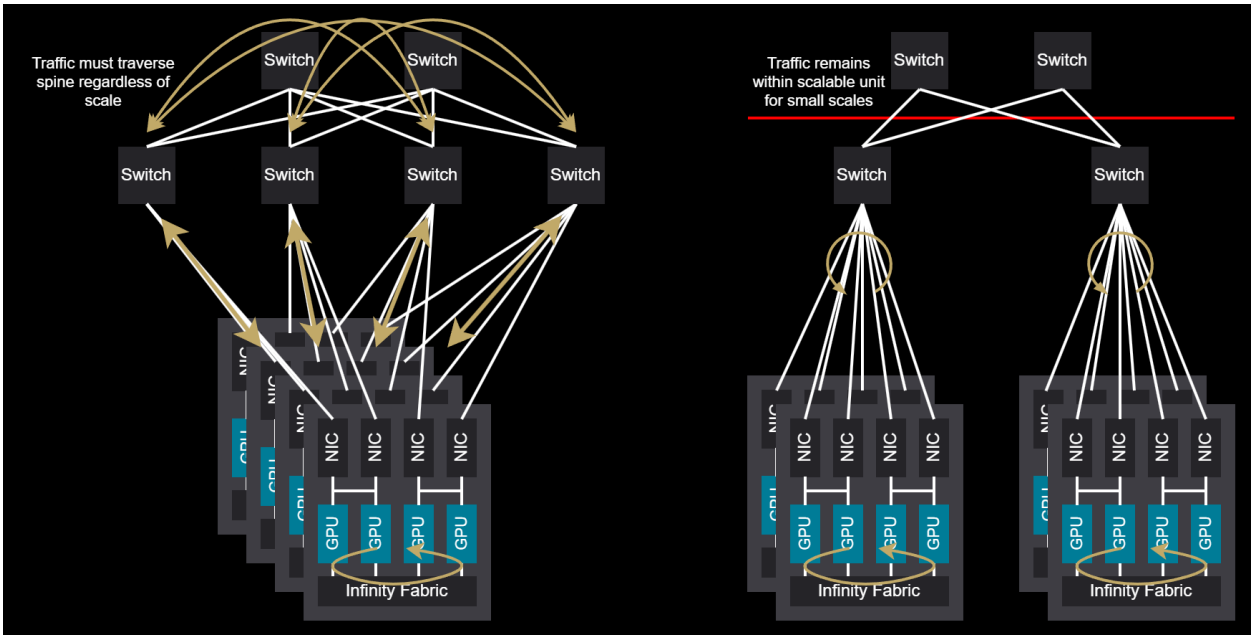
needs to cross rails/ranks must traverse either the network spine layer, or Infinity Fabric (PXN).

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:



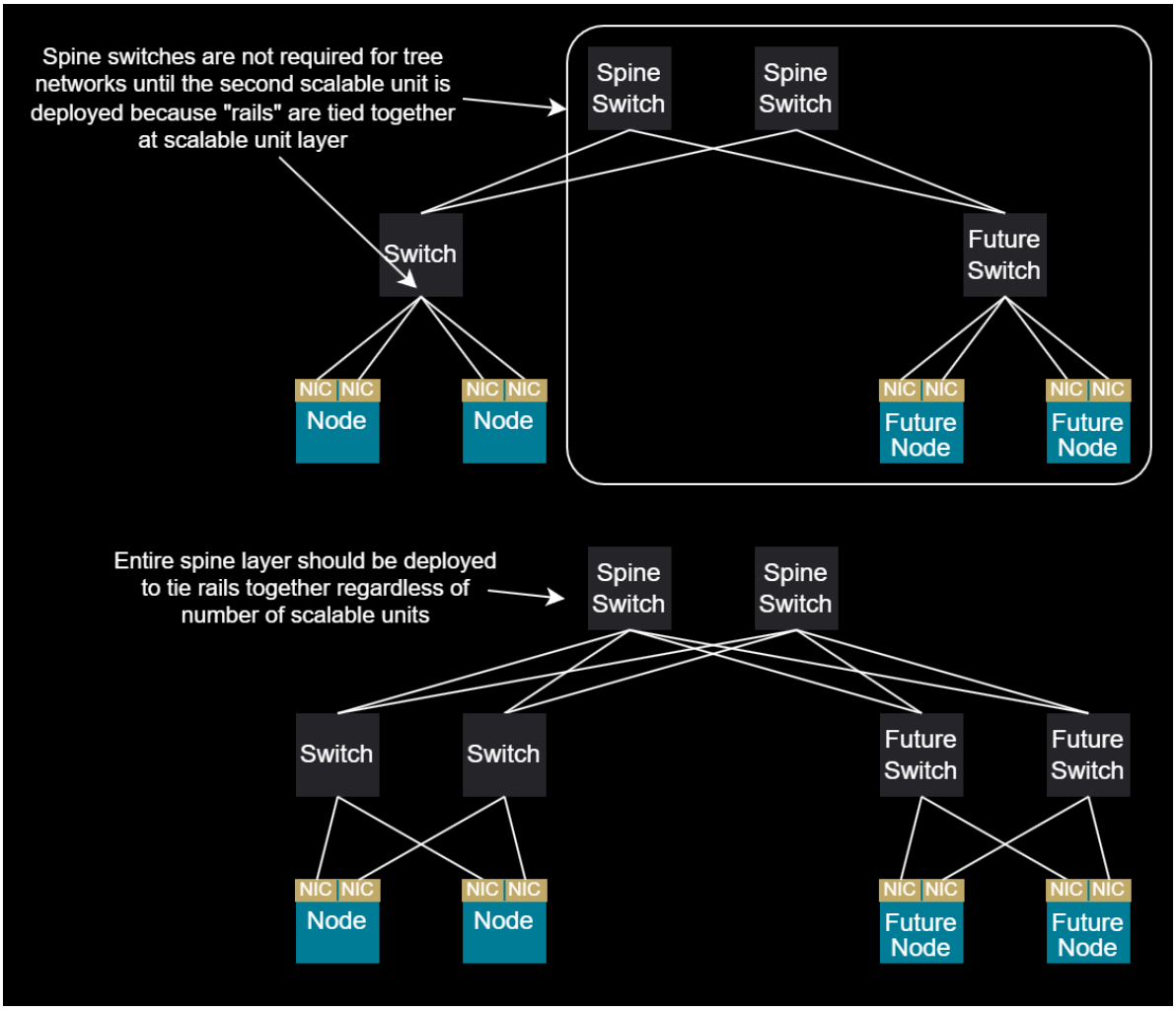
While tree handles cross-rank traffic better:



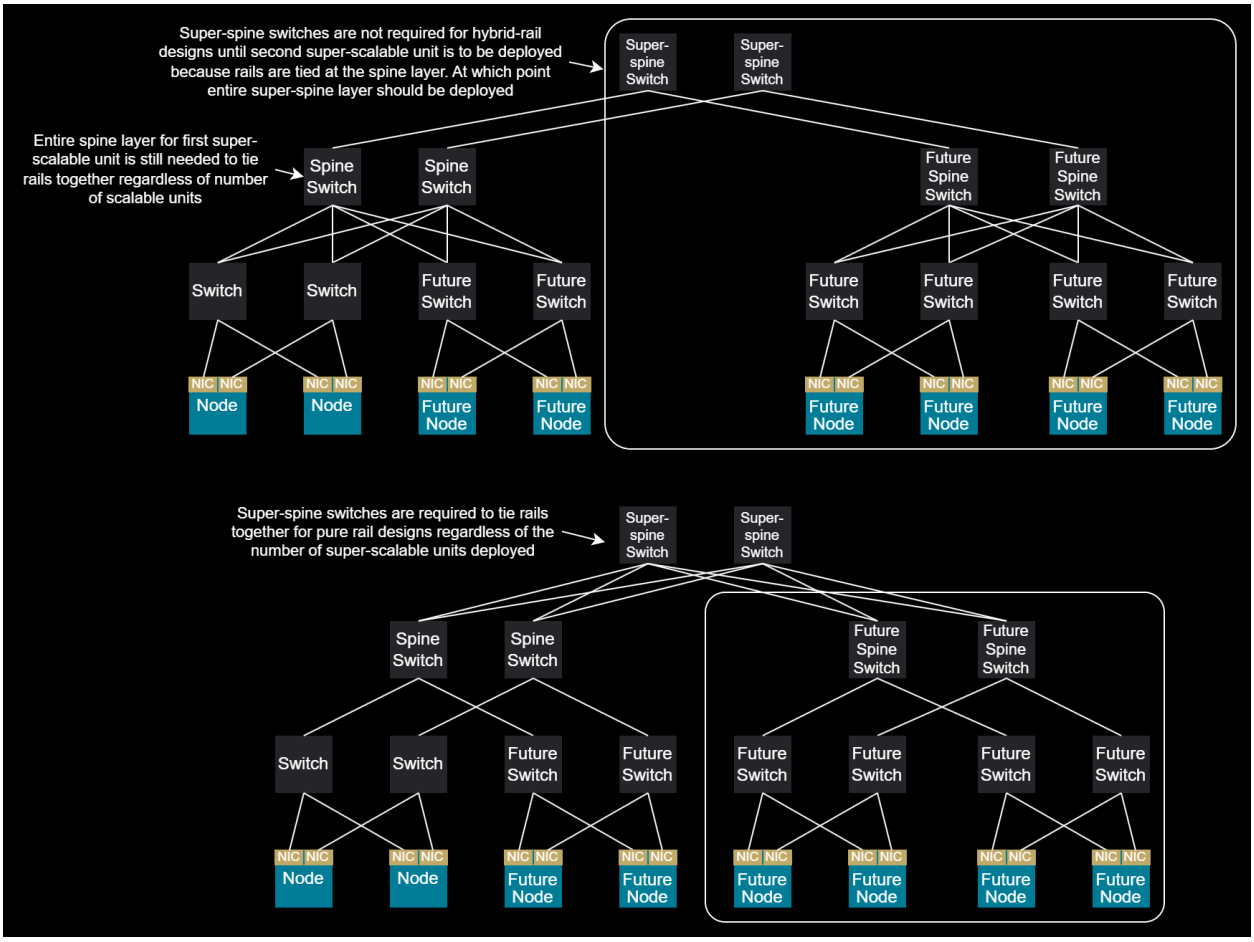
Scaling networks

Cluster backend deployment strategies

2-Tier networks



3-Tier tree networks



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 topologies

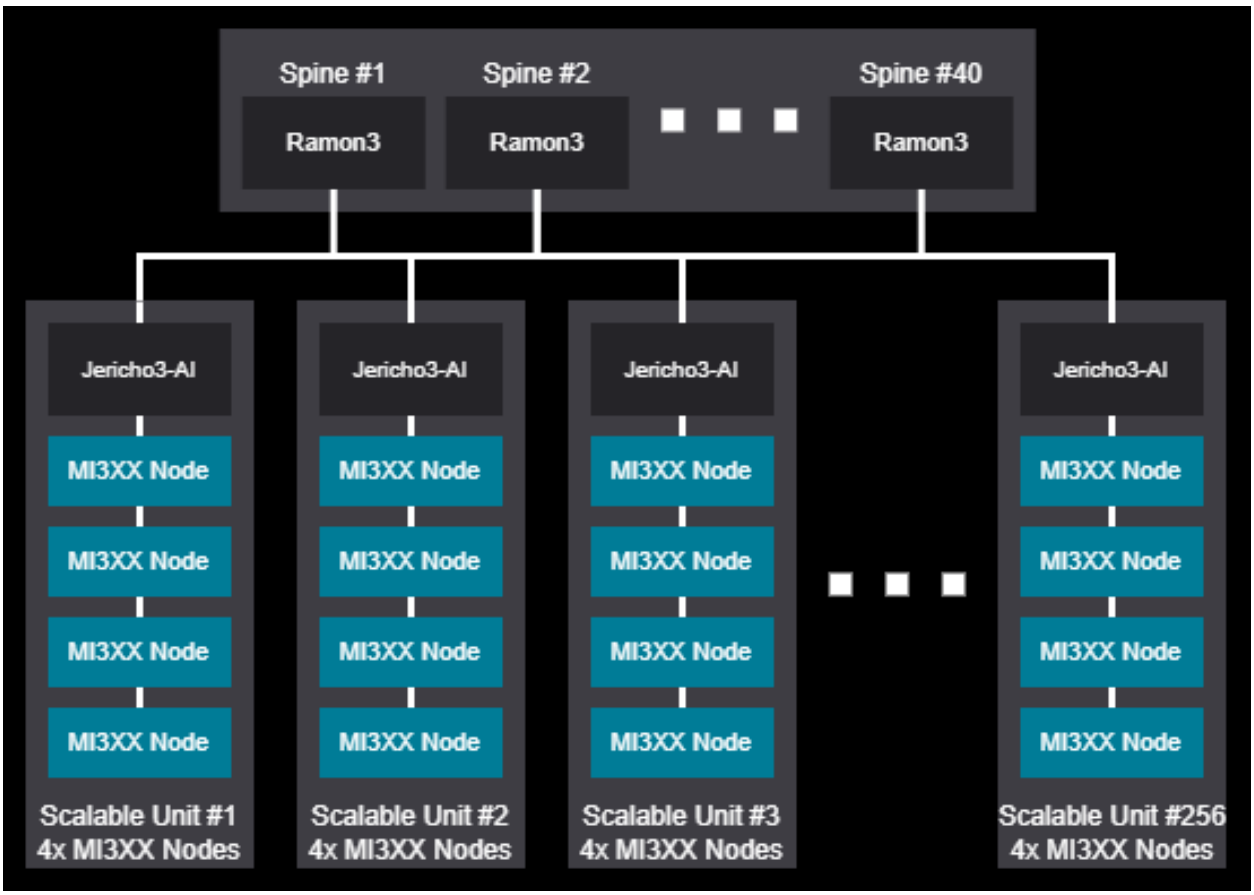
Topologies listed are based on Jericho / Ramon switch type (Accton, Arista, Ciena, Nokia) – please consult desired vendor port count directly to confirm.

Some of the 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.

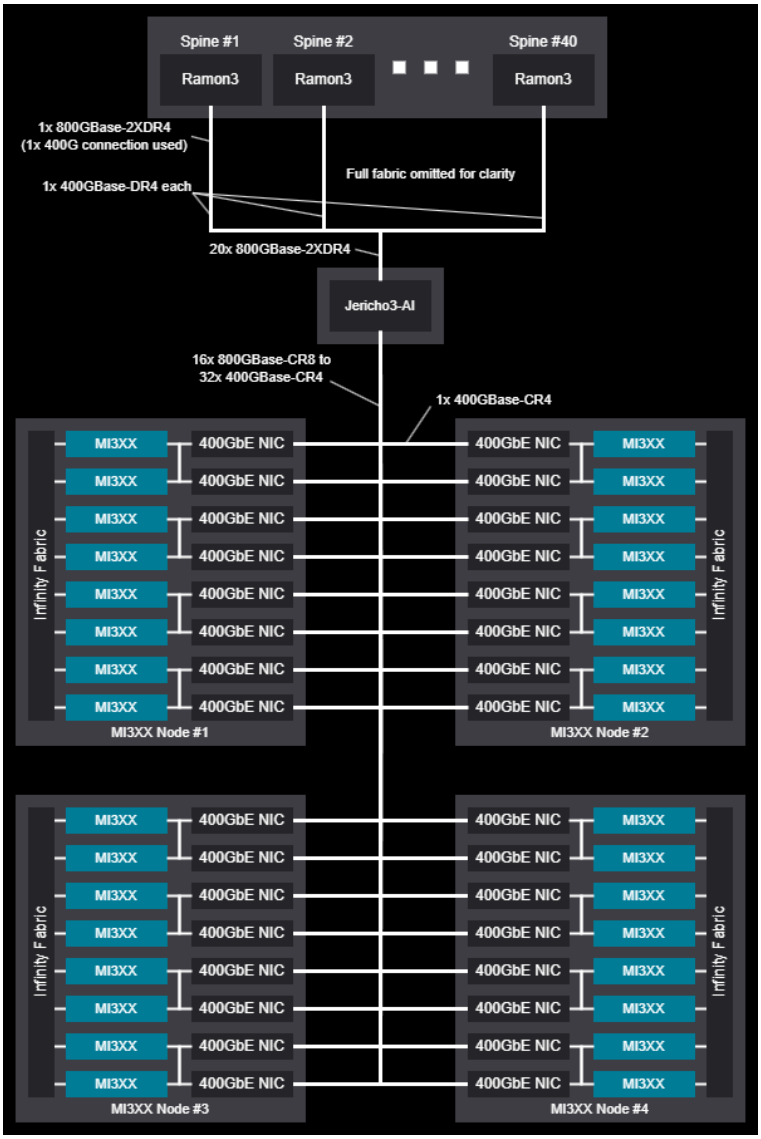
8K GPU topology design examples

Jericho/Ramon network diagrams

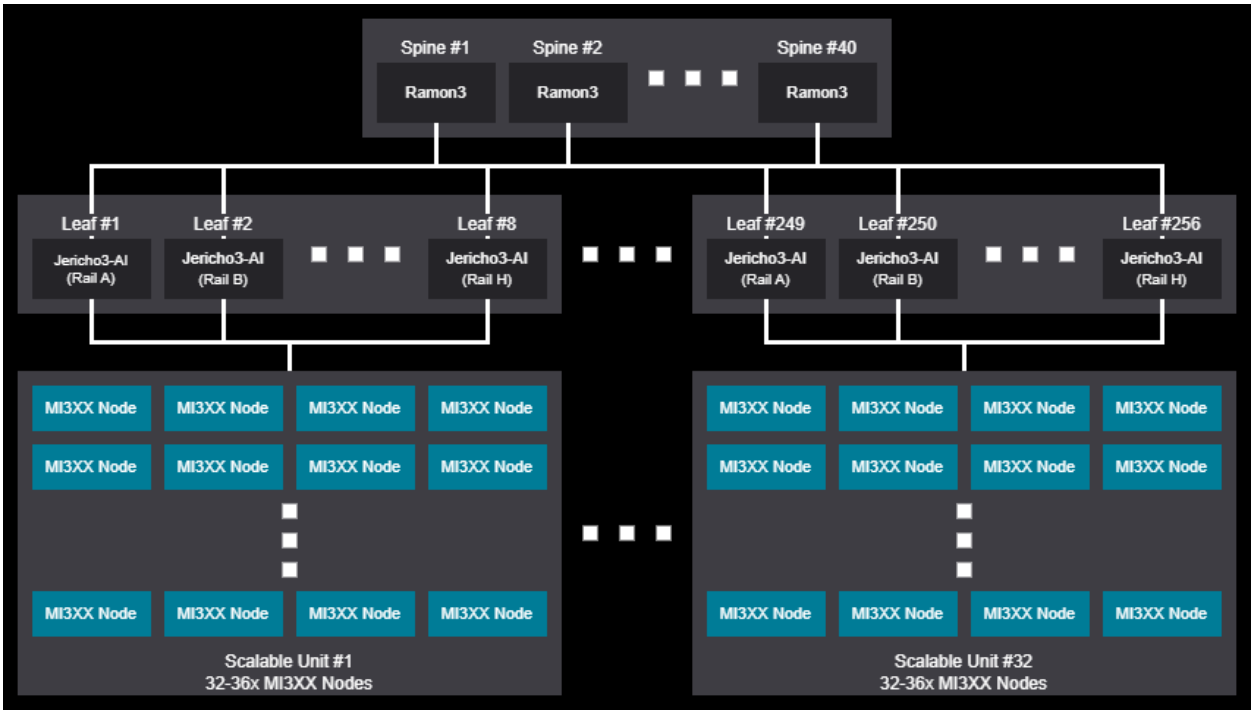
8192 GPU (1024 nodes) tree design



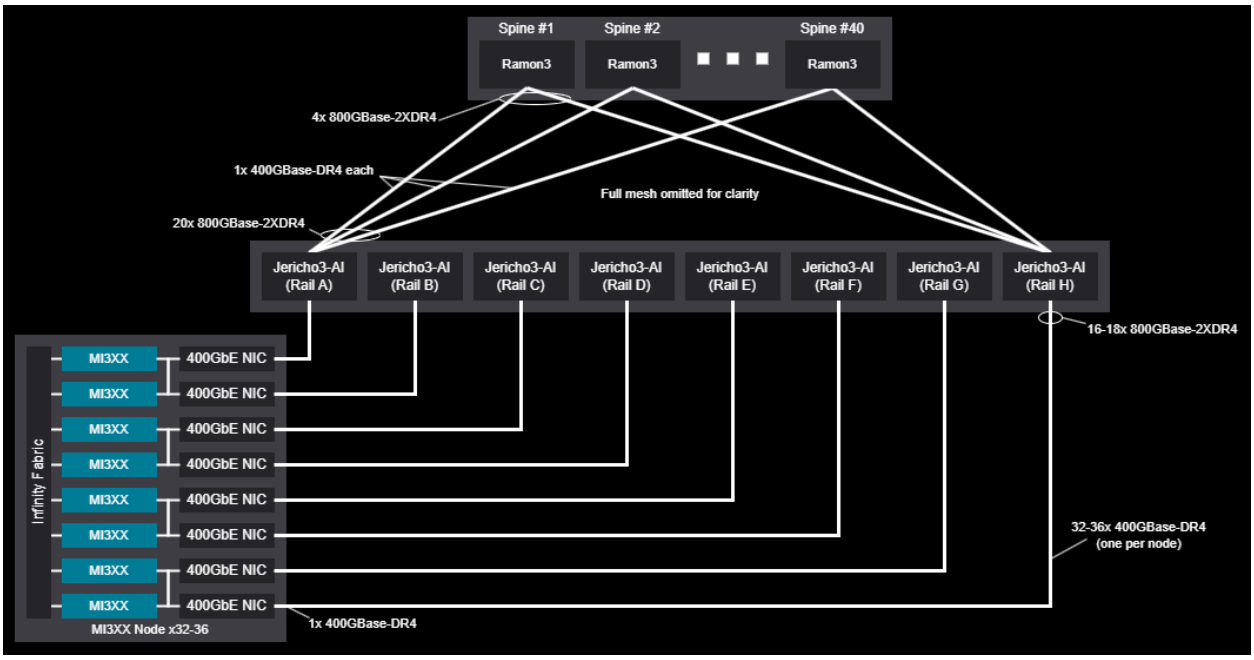
8192 GPU (1024 nodes) tree scalable unit



8192-9216 GPU (1024-1152 nodes) rail design



8192-9216 GPU (1024-1152 nodes) rail scalable unit



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.