VIAVI Solutions

This Former Spirent Business is Now Part of VIAVI

Contact Us +1 844 GO VIAVI | (+1 844 468 4284)

To learn more about VIAVI, visit viavisolutions.com/en-us/spirent-acquisition

Validating high-speed Ethernet for next-gen AI networking

Data center networks must transform to support new AI workloads

The scale of AI applications is skyrocketing with the number of model parameters growing at a blistering 1000x pace every 3 years*. The growing complexity and size of AI applications is driving a need for increasing numbers of AI accelerators, which in turn dictate the type and scale of the networking fabric required to interconnect the accelerators with each other and compute and memory resources.

As AI applications move from small apps with hundreds of AI accelerators to large apps that require tens of thousands of AI accelerators, the networking architectures grow more complex. Up to hundreds of accelerators leveraging CXL, NVLink and/or PCIe can effectively manage connectivity needs. A cluster with thousands of accelerators requires moving to multiple racks to host the accelerators. Clusters of resources running in these racks must be interconnected with an AI leaf layer. As we grow to 10k+ accelerators, a data center scale architecture is required with a routable fabric and AI spine.

AI XPU Size	Server IO 10-100s of XPU	Rack Scale 1000s of XPUs	DC Scale 10K+ of XPU
AI Network Options	CXL NVLink PCIe	AI Leaf Ethernet or IB	AI Spine Ethernet or IB
Type of Applications	Small AI Apps	Moderate AI Apps	Large AI Apps

Scale-up design option

Scale-out design option with a new AI network that is different from traditional network

Source: Dell'Oro Group

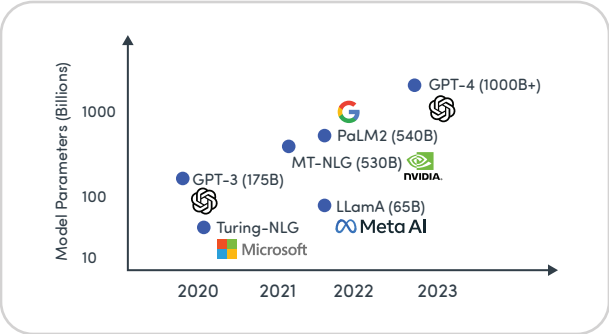
With most AI growth in moderate to large AI apps, AI backend networks leveraging AI leaf and spine approaches are critical for accelerator interconnectivity. This represents a major transformation from traditional front-end data center networks connecting general purpose servers.

Today, AI traffic is increasing by a factor of 10x every 2 years*. With expected continued growth in model complexity, data center network bandwidth per AI accelerator needs to increase from 200/400/800 Gbps today to 1.6 Tbps in the very near future. Given how long it takes to build data centers, the industry must plan ahead for the future growth needs for AI workloads as it designs new data center networks or risk deploying designs that are nearly obsolete as soon as they go live.

* Source: Dell'Oro Group

AI networking challenges

Bandwidth and Scale



High data volumes

AI workloads need high-speed transfers; however, traditional networks may have bottlenecks which limit transfers between compute and accelerator nodes.

Distributed training

Growing workload complexity demands more nodes for AI model training. Traditional networks may not scale efficiently as models move to trillions of parameters and beyond.

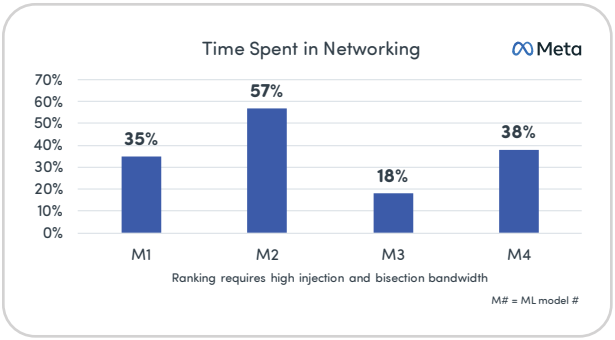
Network Latency

Network radix

Network topology/switch radix influences latency with higher radix providing shorter paths and lower latency for east-west traffic loads.

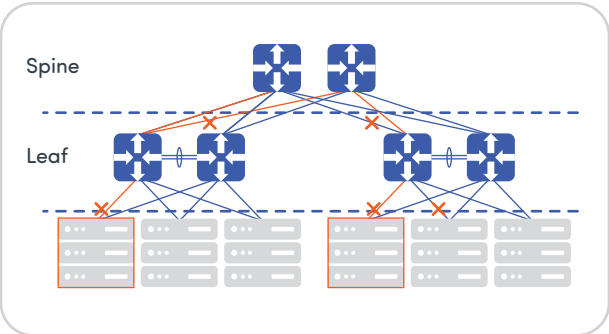
Congestion hotspots

Network congestion slows data transfers between accelerators nodes or into the AI cluster itself, causing lags in training and inference.



Source: 2022 OCP Keynote by Alexis Bjorlin, VP, Infrastructure at Meta

Packet Latency



Packet delivery

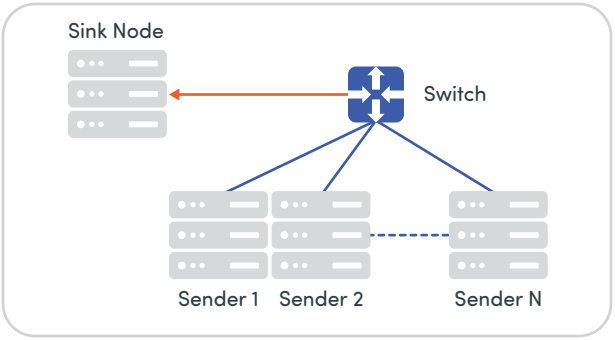
Out-of-order packets can increase latency significantly due to buffering or retransmission.

Long-tail latency

Many AI operations such as model training cannot complete until all required packets are received. The highest 1% or 0.1% of packet latencies can cause extreme delays that disrupt critical AI operations.

Challenges of using traditional Ethernet for AI networking

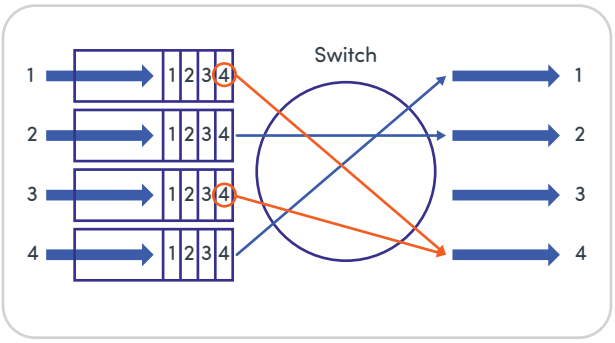
Running Ethernet as it exists today to serve AI/ML workloads is not well-suited to the unique requirements of such traffic.



Bandwidth Limitation

Incast Use Case

- Traffic spike in the network overwhelms the switch from Sender to Sink Node due to bandwidth bottleneck
- Traffic rate from senders gets reduced to avoid congestion



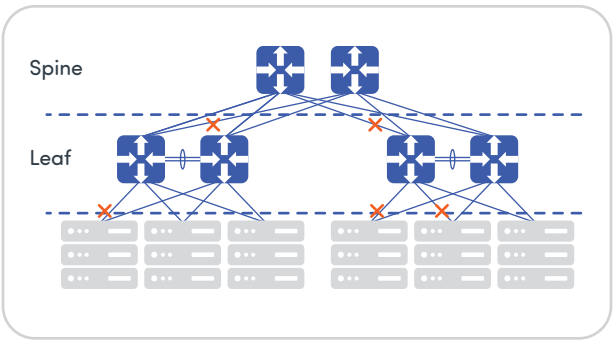
Multi-Path Packet Spraying

Head-of-Line (HOL) blocking

- Packets get blocked because the first packet in a queue is held up

Go back N routing algorithm

- Out of order packets will be dropped and all the subsequent frames will need to be retransmitted



Congestion Control

Priority Flow Control (PFC) storm and deadlock

- Packets paused in a loop across multiple switches block all packets from being transmitted

New Ethernet technologies for better AI networking

The following technologies will help in alleviating network issues when serving AI workloads.

- **Higher bandwidth networking**
  - 400G, 800G, and 1.6T Ethernet on AI data center switch
  - 200G, 400G, and 800G interfaces on SmartNIC/GPU/DPU/TPU
  - 400G and 800G low-latency and low-power optics transceiver
- **Advanced protocols for achieving a low latency, congestion-free network**
  - Optimized RoCEv2 – RoCE selective repeat
  - Multi-pathing packet spraying
  - AI optimized congestion control
  - High scale – High-radix fabric enabling the building of massive flat networks

About Spirent

Spirent Communications (LSE: SPT) is a global leader with deep expertise and decades of experience in testing, assurance, analytics and security, serving developers, service providers, and enterprise networks. We help bring clarity to increasingly complex technological and business challenges. Spirent's customers have made a promise to their customers to deliver superior performance. Spirent assures that those promises are fulfilled.

For more information visit: www.spirent.com

Test and validation solutions

Test solutions help validate that the industry is leveraging the cost-intensive AI/ML infrastructure to its maximum capabilities.

Ethernet Validation

- Up to 800G, multi-speed appliances
- Modular chassis based solutions
- License just the speeds and port blocks you need
- Full support, automation, per port reservation
- Breakouts supported
- L1 to L3 statistics and performance
- RoCEv2 hardware-based congestion control
- Bidirectional RoCEv2 traffic
- Collective Communications Library supported (on A1 and B3 appliances only)

B3 800G 2U Appliances

A1 400G 2U Appliance

A2 400G 2U Appliance

MX2 / MX3 Series
Multiprotocol Performance

FX2 / FX3 Series
Flexible Functionality

Americas 1-800-SPIRENT
+1-800-774-7368 | sales@spirent.com

Europe and the Middle East
+44 (0) 1293 767979 | emeainfo@spirent.com

Asia and the Pacific
+86-10-8518-2539 | salesasia@spirent.com

Rev D | 08/25 | www.spirent.com