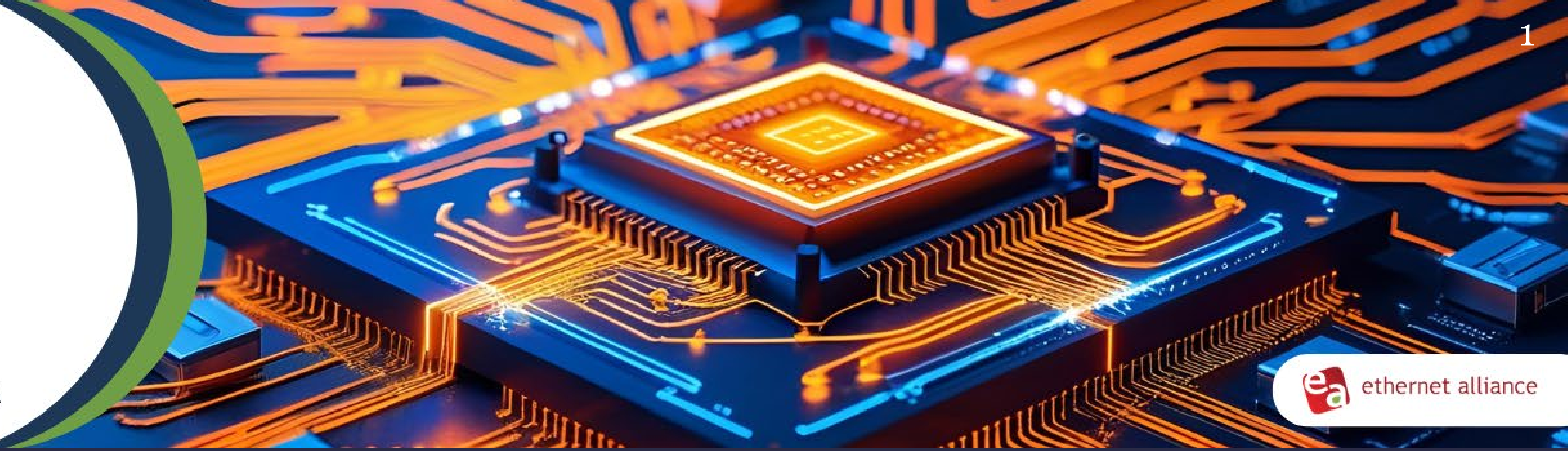
TEF 2025
Ethernet for AI

December 2-3, 2025
Hyatt Centric Mountain View, CA, USA



400 Gb/s Signaling for AI Networks From A System Perspective

December 2-3, 2025

This presentation has been developed within the Ethernet Alliance, and is intended to educate and promote the exchange of information. Opinions expressed during this presentation are the views of the presenters, and should not be considered the views or positions of the Ethernet Alliance

Setting the Stage for Networking in an AI World

Alan Weckel, Founder and Technology Analyst – 650 Group

AI Waves

Wave 1
Academic Research

- Pre 2022
- <$10B in equipment spend

Wave 2
Foundational Models and Content Creation

- 2022 - 2025
- $300B in equipment spend

Wave 3
AI Agents

- 2025-2028
- ~$1T in equipment spend

Wave 4
Autonomous Transportation and Robots

- 2027-2035
- >~1T in equipment spend

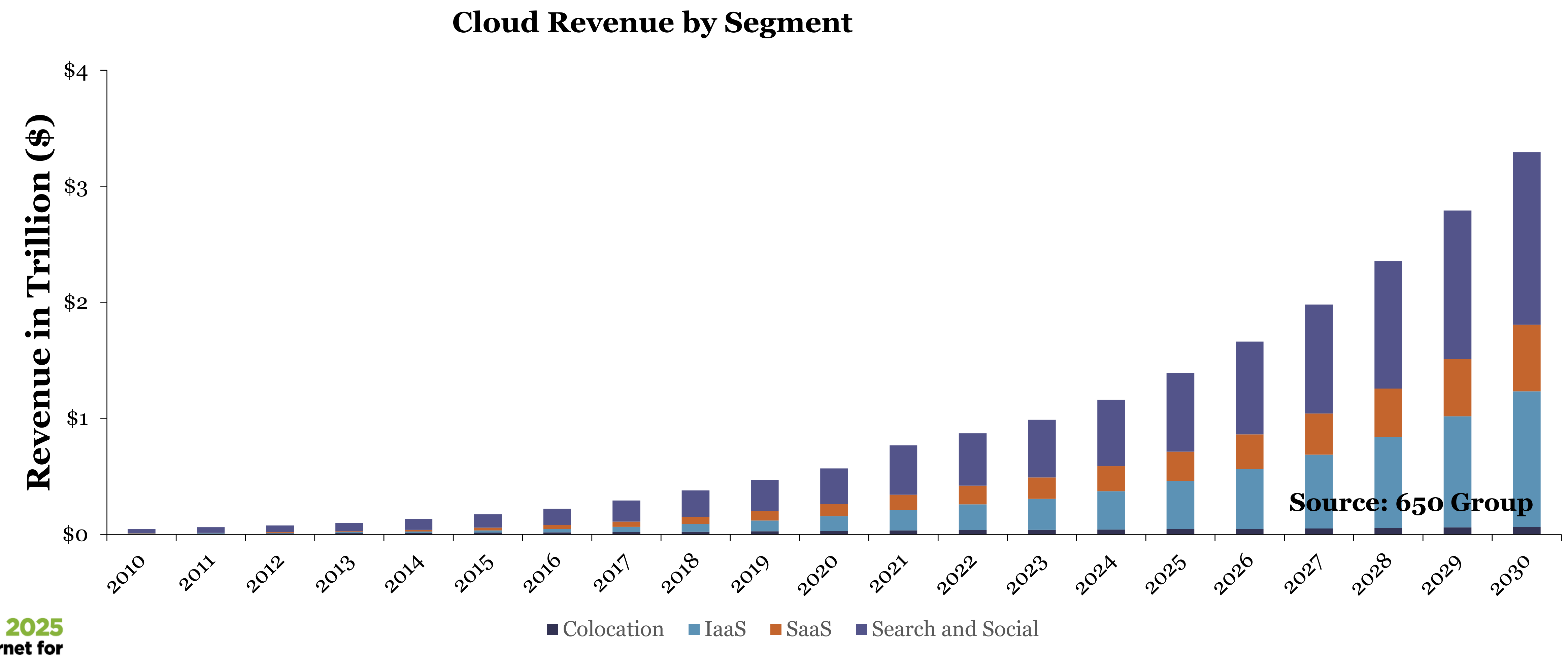
What is $1T of equipment spend to support AI really mean?

- $1T = ~$1M a minute in spend
- 300 DC Switch ports shipping a minute
- 200,000 Tb of bandwidth shipping each minute

Source: 650 Group

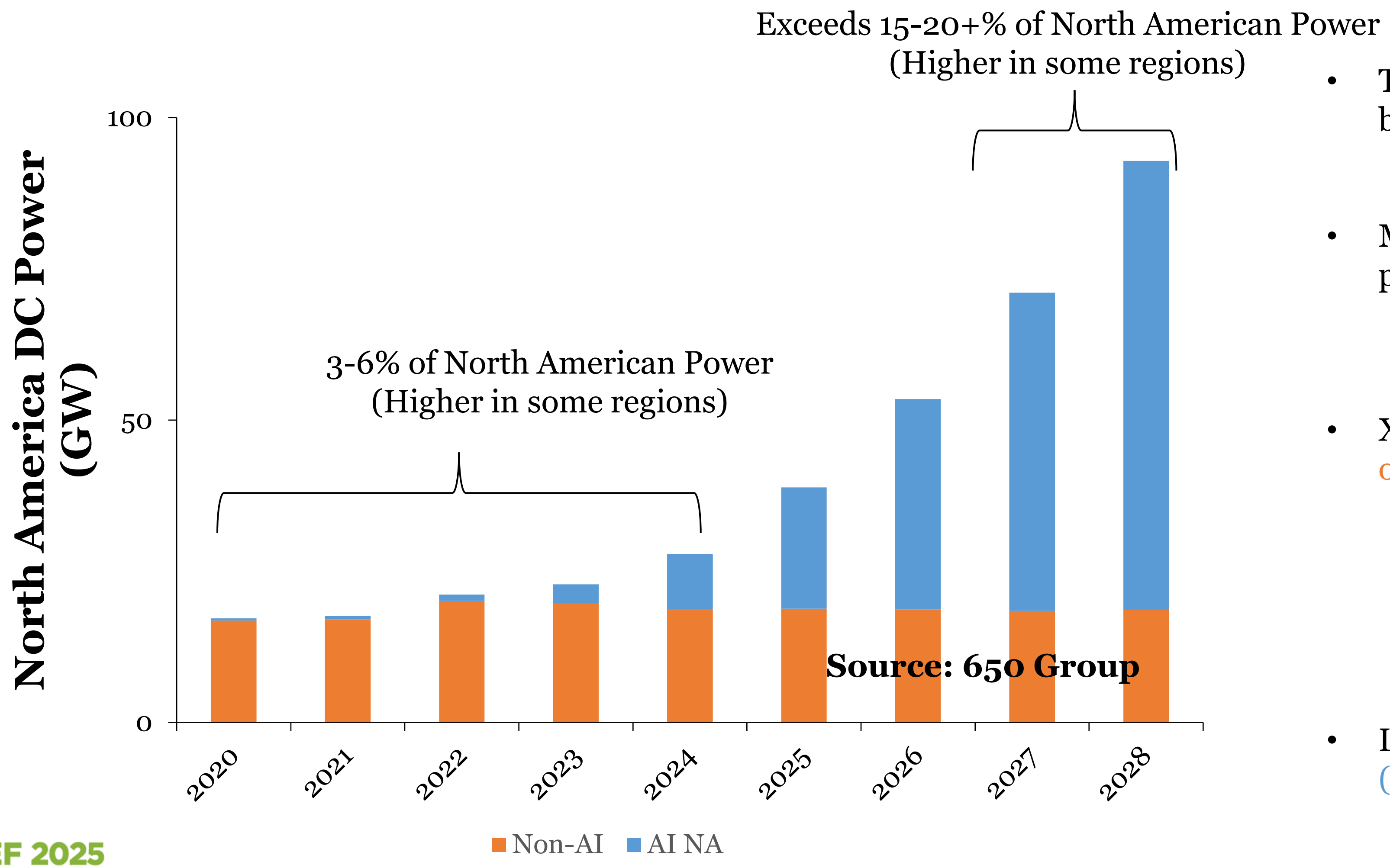
TEF 2025
Ethernet for AI

Cloud Revenue Support Continued AI Spend



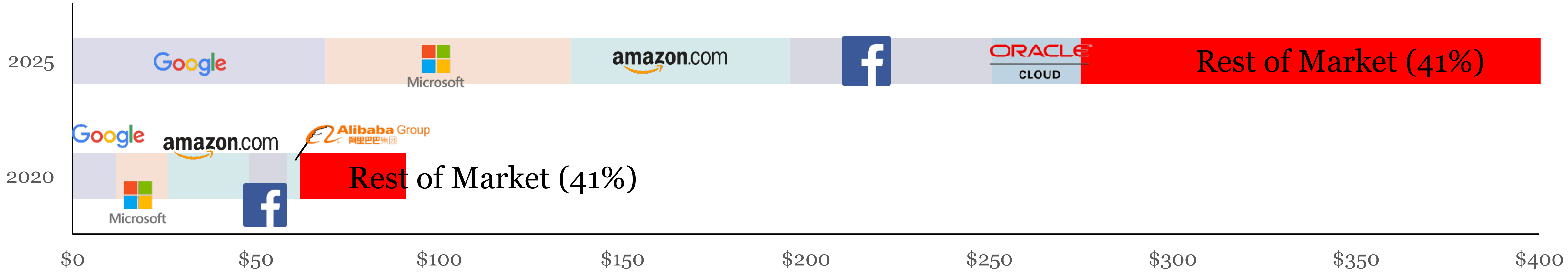
TEF 2025 Ethernet for AI

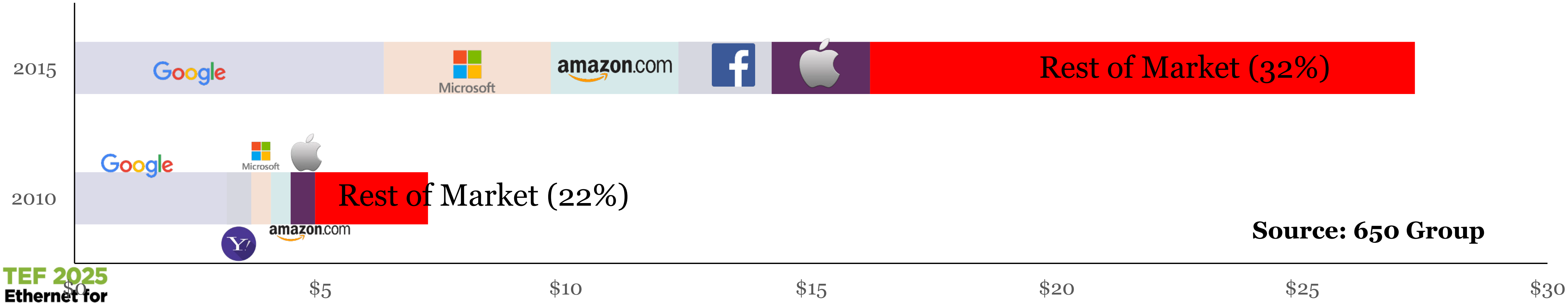
DC Related Power in North America



- Tier-2 Training and Inference will need purpose-built lower power ASICs (lowers blue bar)
  - The right ASIC for the right workload
- May move workloads to other continents where power is more readily available (lowers blue bar)
  - Similar to how most of Japan's DCs sit in the Pacific Northwest
- X86 Server refresh can push down Non-AI (lowers orange bar)
  - 1M older servers can be replaced with ~600K to get the same level of compute
  - Only a one-time savings, but can cover almost all of one years shortfall in new power generation
- Liquid Cooling reduces power consumption (lowers blue bar)
  - Cold Plate cooling is the technology for current data centers
  - Immersion cooling is not ideal in many facilities

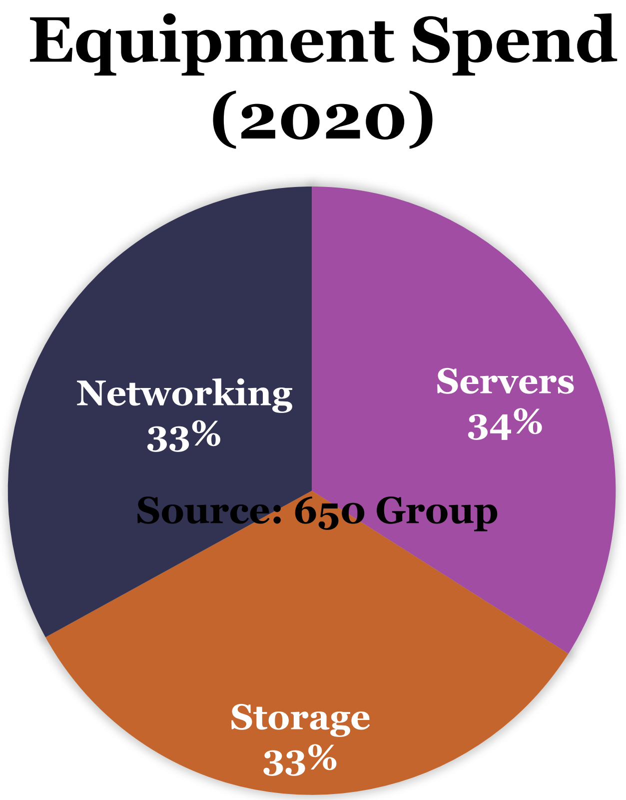
DC Equipment CAPEX and Players Shift

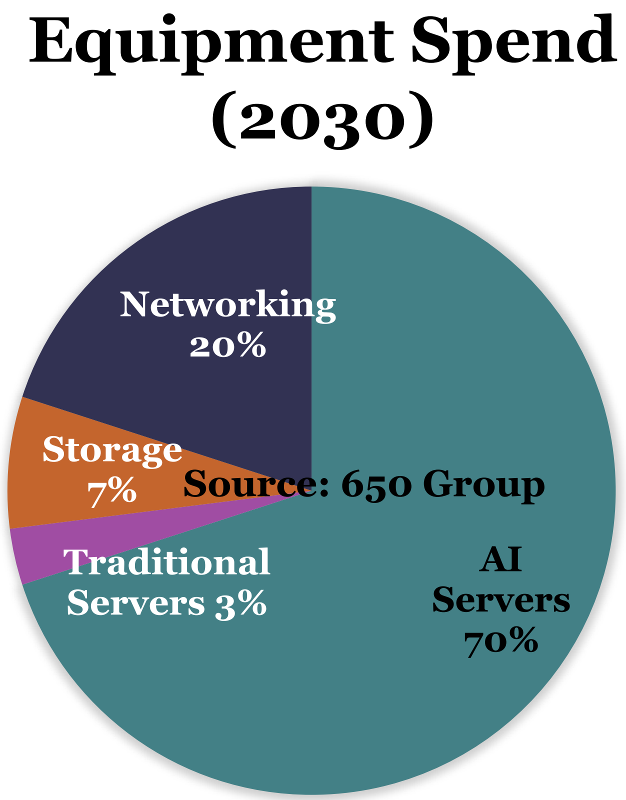




TEF 2025 Ethernet for AI

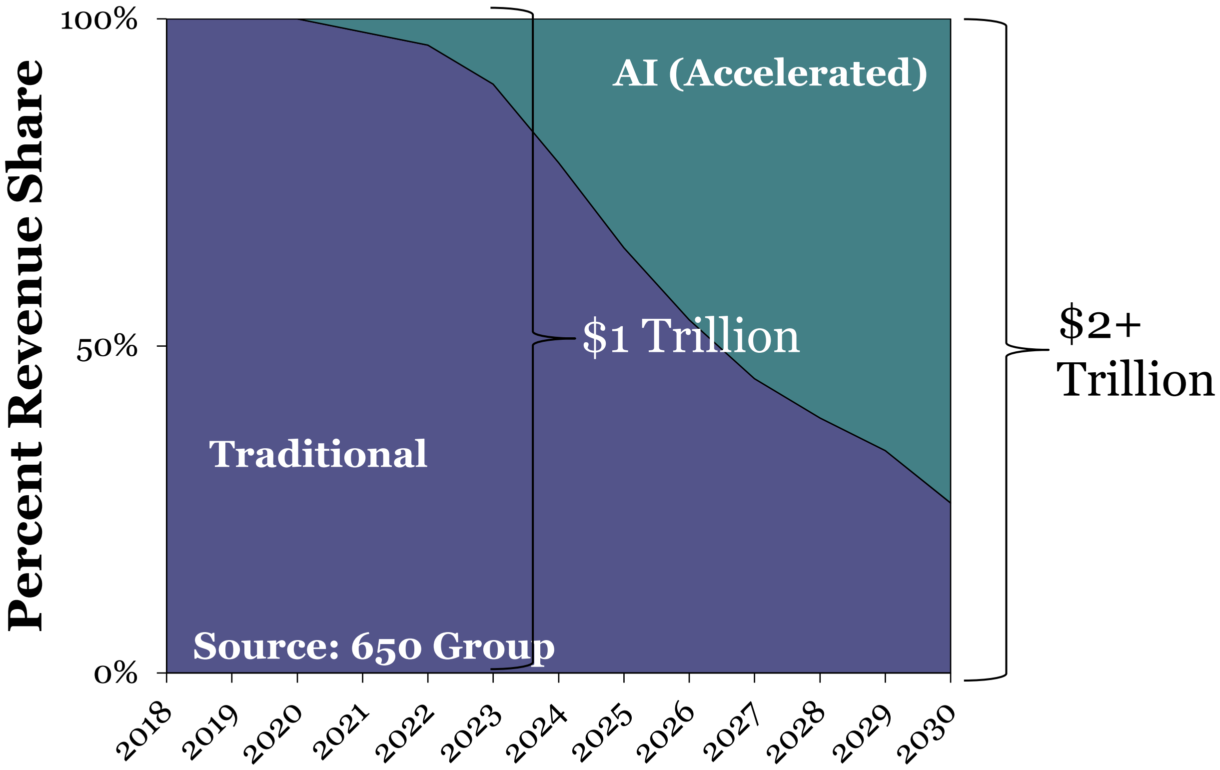
US Top 5 Equipment CAPEX Spend



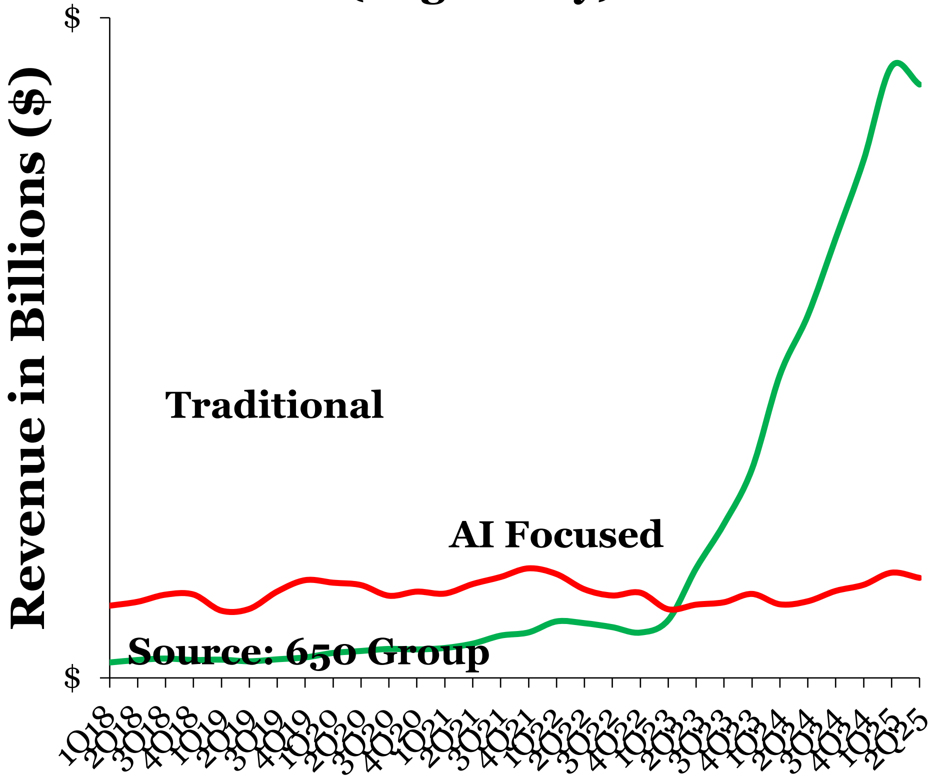


Market Transition to AI/ML

DC Installed Base of Equipment

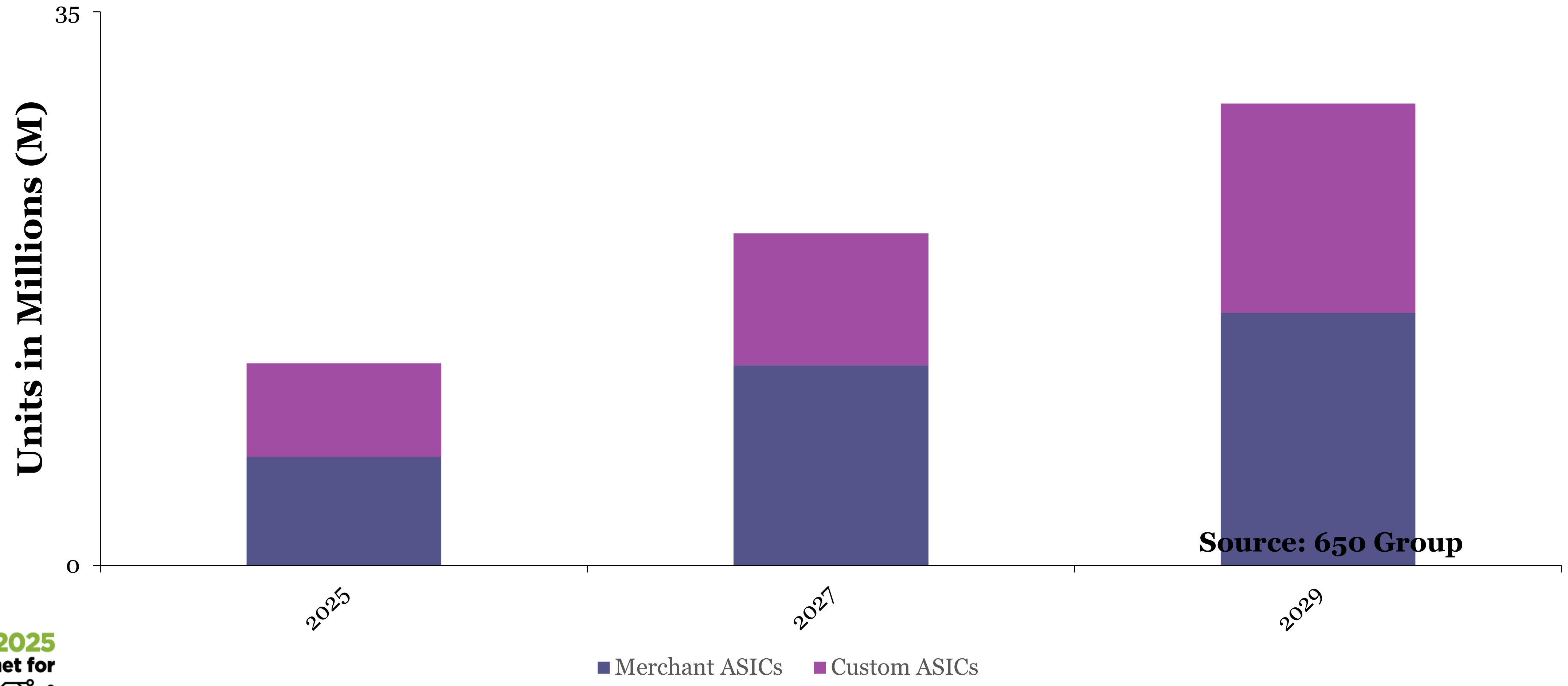


DC Semiconductor Revenue (Logic Only)



TEF 2025 Ethernet for AI

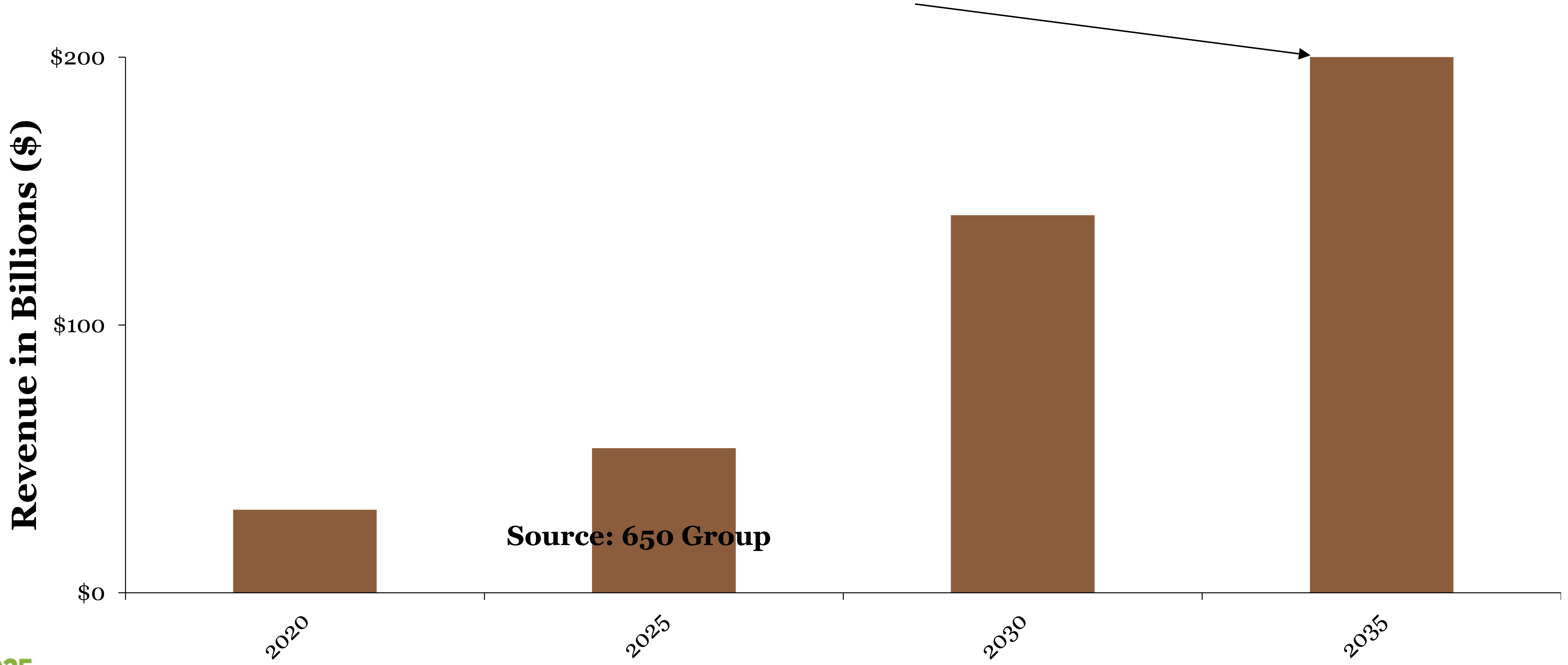
GPU and XPU Shipments Converge



TEF 2025
Ethernet for AI

The Evolution of the Ethernet Switch Market

By 2035, led by AI, The Ethernet Switch Market will Exceed $200B



Includes Total Ethernet Switching Market Campus and DC (Scaleup, Scaleout, Frontend, Scale Across)
Does not include NICs, InfiniBand, NVLink, PCIe

TEF 2025 Ethernet for AI

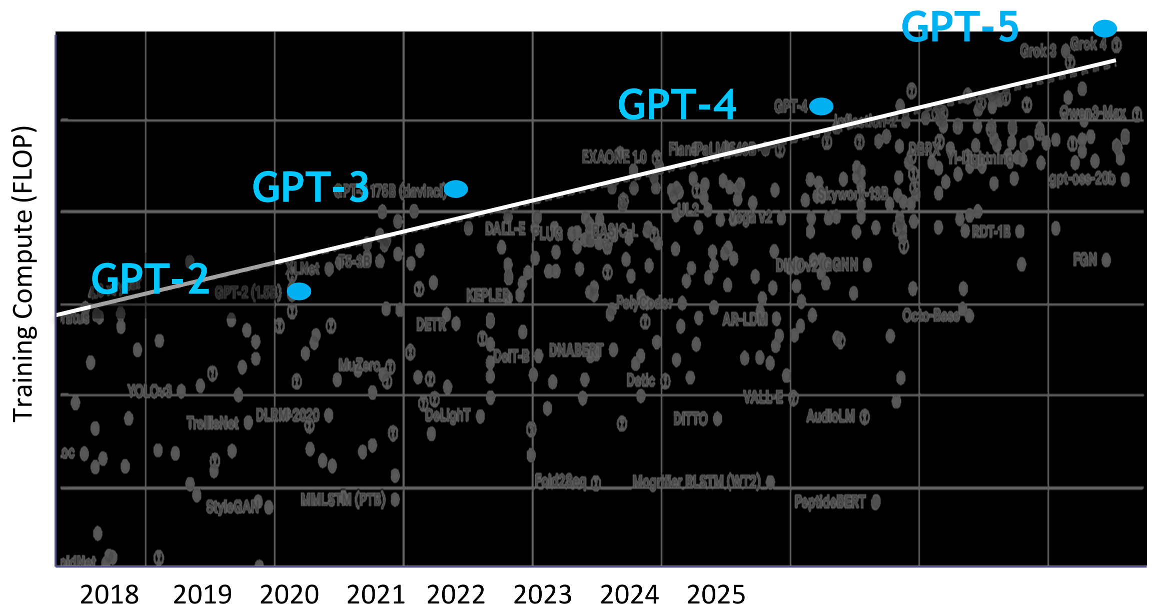
Building systems for AI deployments

Brian Welch, Distinguished Engineer, Cisco

Acknowledgements: This presentation would not exist without the inputs, expertise, and patience of many Cisco colleagues! Thanks to Anthony Torza, Mark Nowell, David Nozadze, Mike Sapozhnikov and Rakesh Chopra

TEF 2025
Ethernet for AI

Unlocking AI potential | Scaling AI clusters

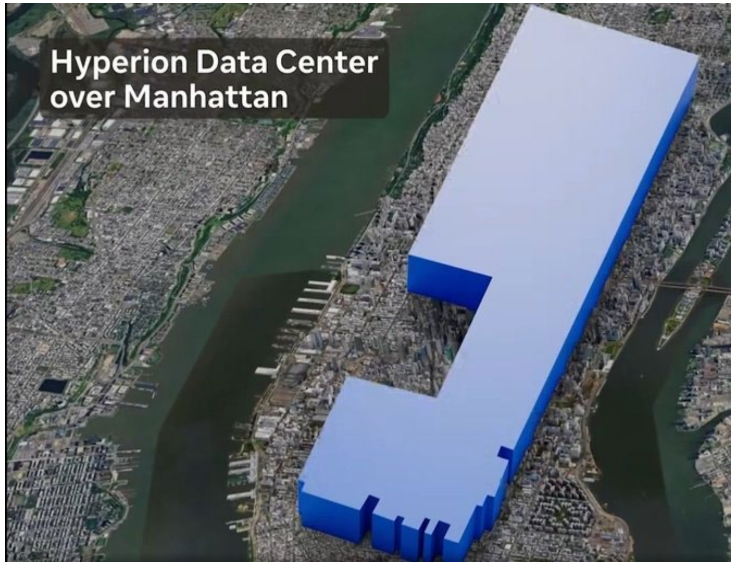


Baseline graph from Epoch AI, September 2025, added GPT5 point manually

Other major models like Llama, Gemini, Grok, others have excellent performance. Only using GPT to simplify trends.

Meta Hyperion Data Center

Up to 5GW



Meta's planned Hyperion data center (Credit:Meta)

Limited real estate & high electricity costs

Data centers must "Scale Across" to other sites

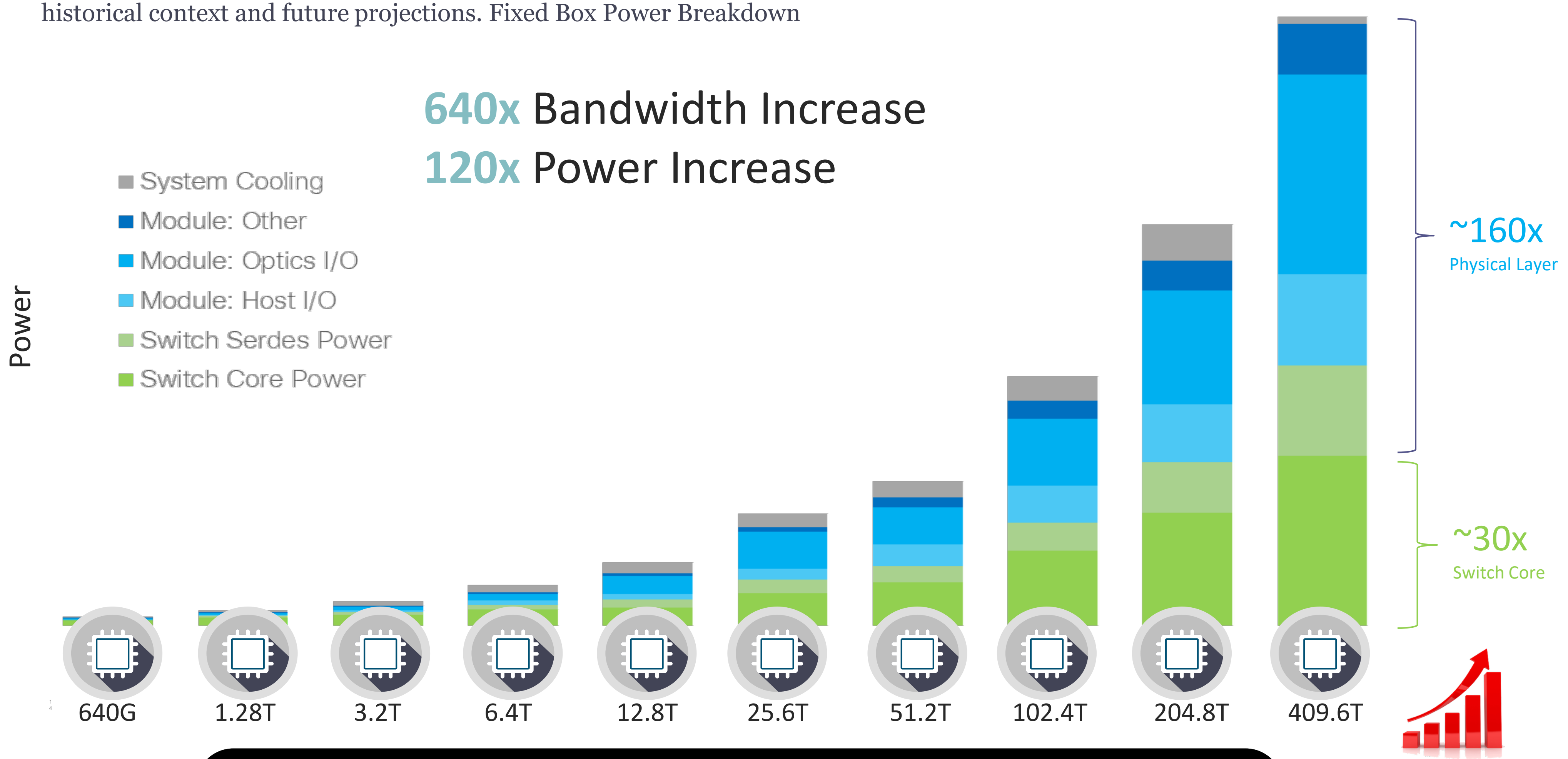
Equipment power efficiency highest priority

Cluster size unlocks intelligence

Power limits the ability to scale

Interconnect power increasingly dominates

Represents a combination of multiple chip families and architectures to provide historical context and future projections. Fixed Box Power Breakdown



Increase vs. 2010

Interconnect power is increasingly becoming a dominant factor

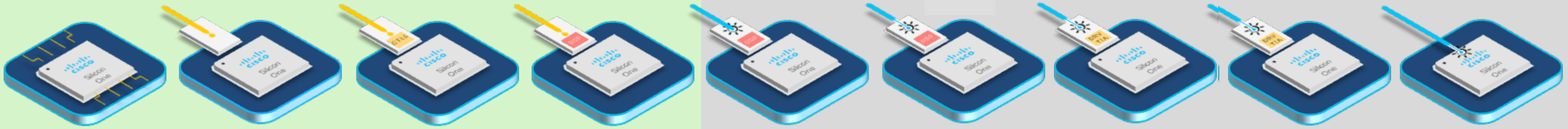
High-Density Interconnect options: Power vs. Reach Tradeoff

Power and Reach @ 1.6T

← Maximize use within a rack →

← Use when inter-rack →

Backplane Channel	Direct Attach Copper (DAC)	Active Copper Cable (Linear) (ACC)	Active Electrical Cable (AEC)	Active Optical Cable (AOC)	Retimed Pluggable Optics	Linear Receive Optics (LRO)	Linear Pluggable Optics (LPO)	Co-packaged Optics (CPO)
<1W	<1W	3.5W	17W	25W	25-30W	17-22W	12-15W	<10W
0.5m	1.0m	2 to 3m	3m+	20m	2000m	2000m	2000m	2000m



Copper cables and interfaces still look attractive from power perspective

Longer reach = increased power

Power reduction opportunities coming from CPO, LPO & LRO (i.e. removing the DSP's power impact)

Switch system design implications

Scale Up	Scale Out	Scale Across
High level of GPU integration	Aggregation of Scale-up clusters	Geographical Aggregation
High radix Interconnects ➢ Dense, short reach ➢ Copper (& optics)	Intra- and Inter-building reaches ➢ Optics (& copper) ➢ Pluggable and CPO	Inter-building/DCI ➢ Coherent Optics ➢ Pluggable
Ethernet Physical Layer	Ethernet	Ethernet
Thermal density drives towards liquid cooling	Air-cooled and liquid cooled facilities	Air-cooled and liquid cooled facilities

- System design: One size does not fit all – architecture dependent
- Common technology building blocks are required and will be packaged in many ways

Preparing for 400G SerDes-based systems

Electrical interfaces

High radix required

Many needs to address:

- Chip to Chip
- Chip to module
- Cables (passive & active)

FEC, Coding and Modulation undetermined

- Balance FEC, OH, gain, latency, power, backward compatibility
- Convergence needed soon

New connectors Needed

- CPC
- Pluggable

Optical Interfaces

Early requirements for **optimized high-radix** short reach solution

- Longer reaches to follow

Power efficiency is key

- shared lasers, linear, CPO

FEC & Coding likely reused from 200G

Pluggables vs CPO

Need for CPO

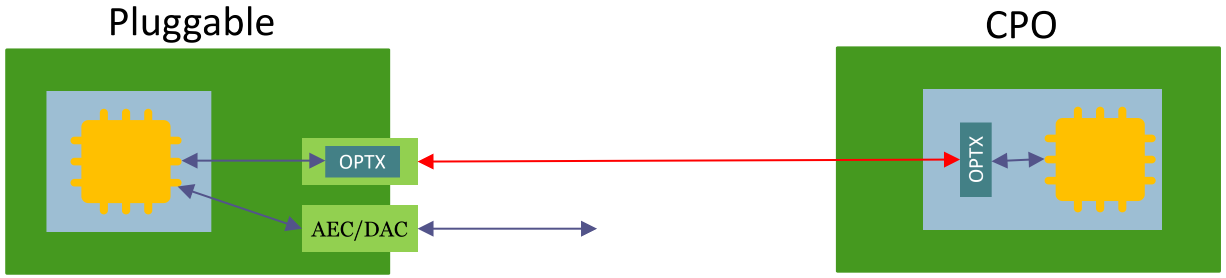
- Power efficiency & density
- Simple AUI

Need for pluggables

- Switch and End-Point
- Interface flexibility

New pluggable connectors

- Current solutions do not have the SI performance



Difficult to separate the optical coding from the electrical coding discussion.

Summary

- AI infrastructure requires scale
- Scale requires power
- Equipment design will be dominated by meeting the requirements of scale with optimum power efficiency
- Power efficiency – a critical factor for feasibility and operational cost – without compromising the essential pillars of performance, cost-effectiveness, and a robust ecosystem.

- The path to 400G will require innovation but must leverage current solutions to ramp to scale quickly.
  - 400G solutions that are very different (technology, architecture, or economics) from 200G may not meet the needs of the market.

AI Scale-Out Network Designs and Interconnects

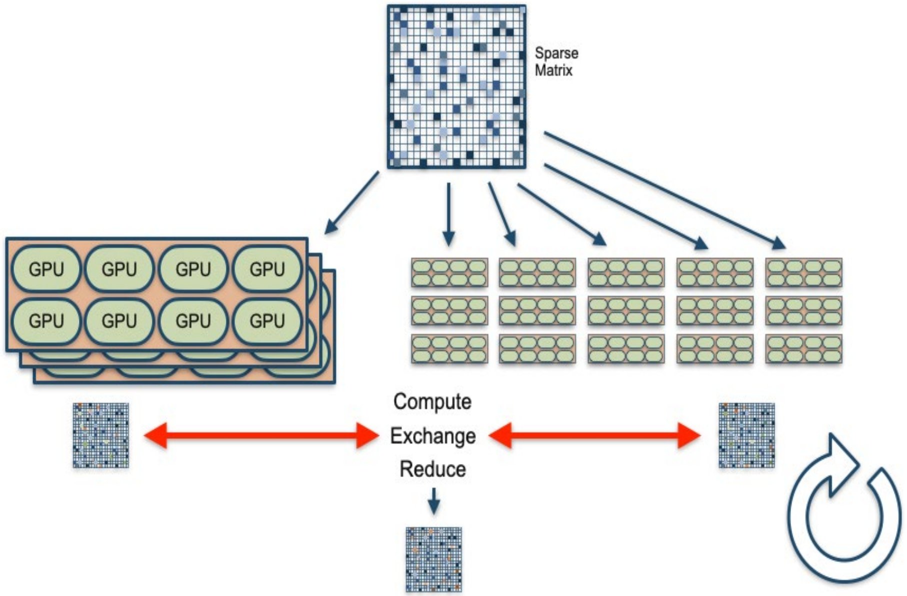
Arihant Jain, Manager-Systems Engineering, Arista

Agenda

- Introduction
- Why Scale-out AI Ethernet Networking
- Architectures
  - Switch Scheduled vs Endpoint Scheduled
  - Rail & Plane Architectures
- Interconnects
- Summary

Why Scale-Out AI Networking ?

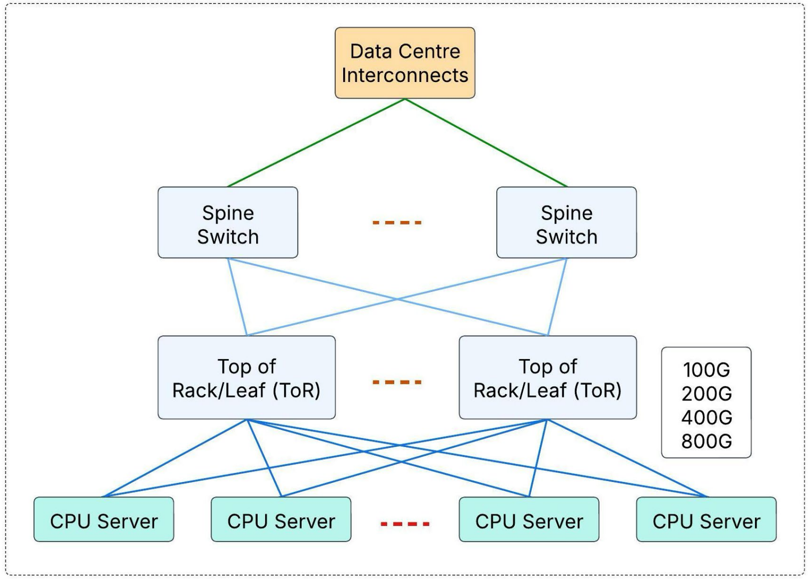
Evolution from an era of information to an era of intelligence is creating unprecedented demand for AI Infrastructure

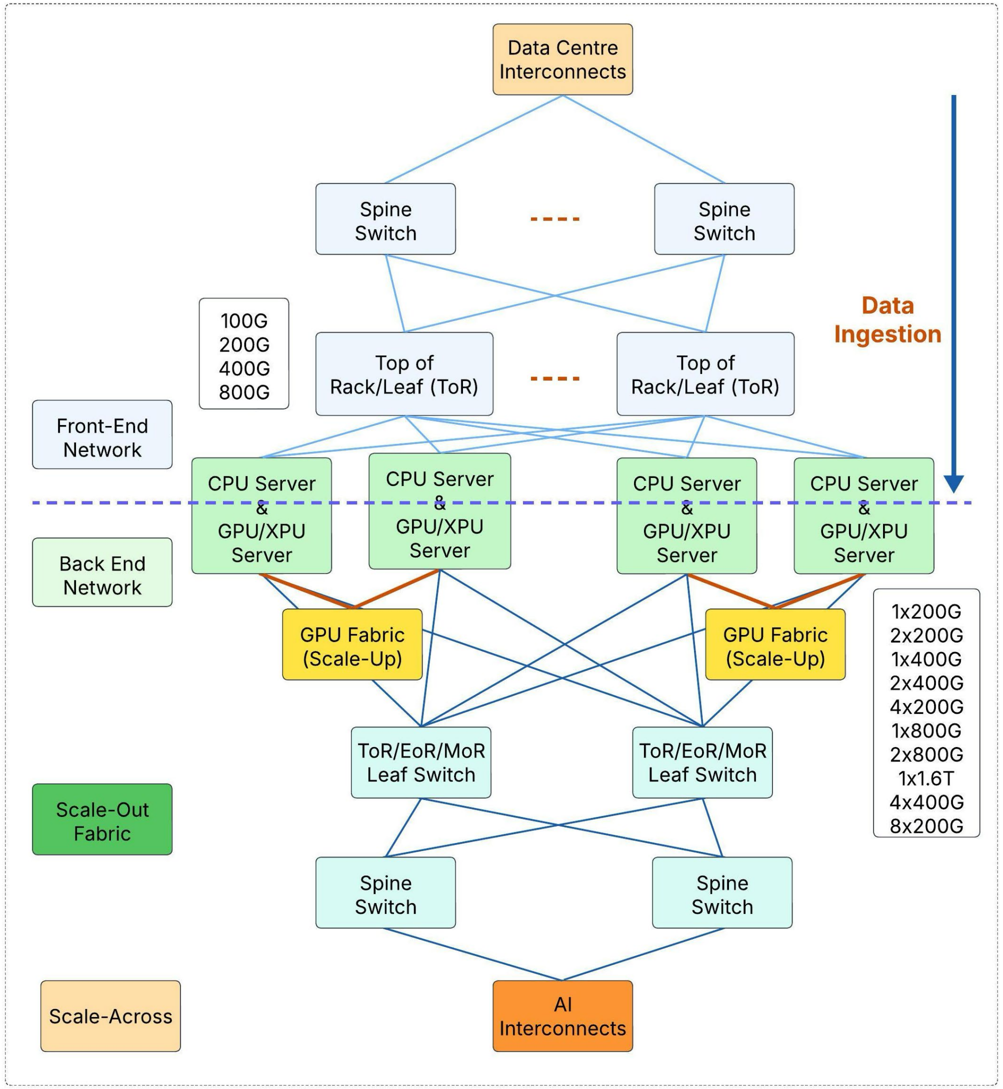


Accelerating the evolution of AI networking to deliver

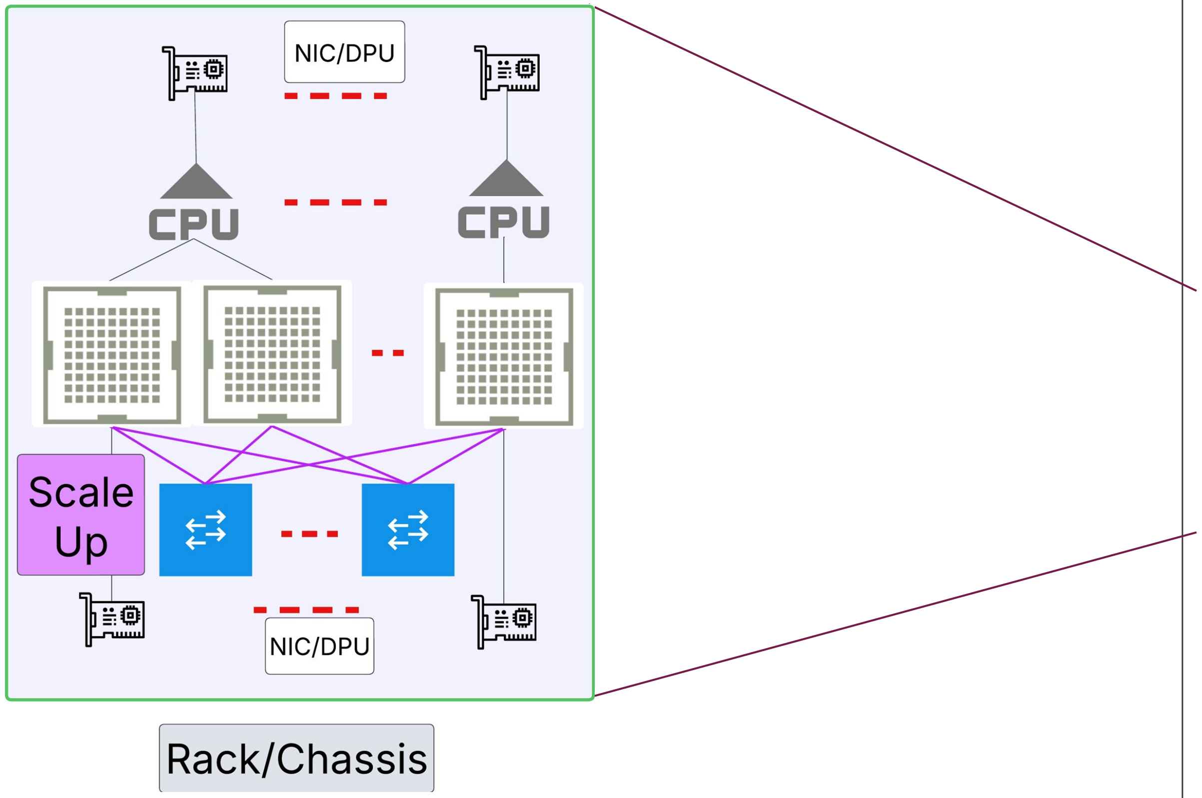
- Unparalleled scale (Mn+ GPU's)
- Efficiency
- Flexibility (Technology, Design/Topology, Deployment)

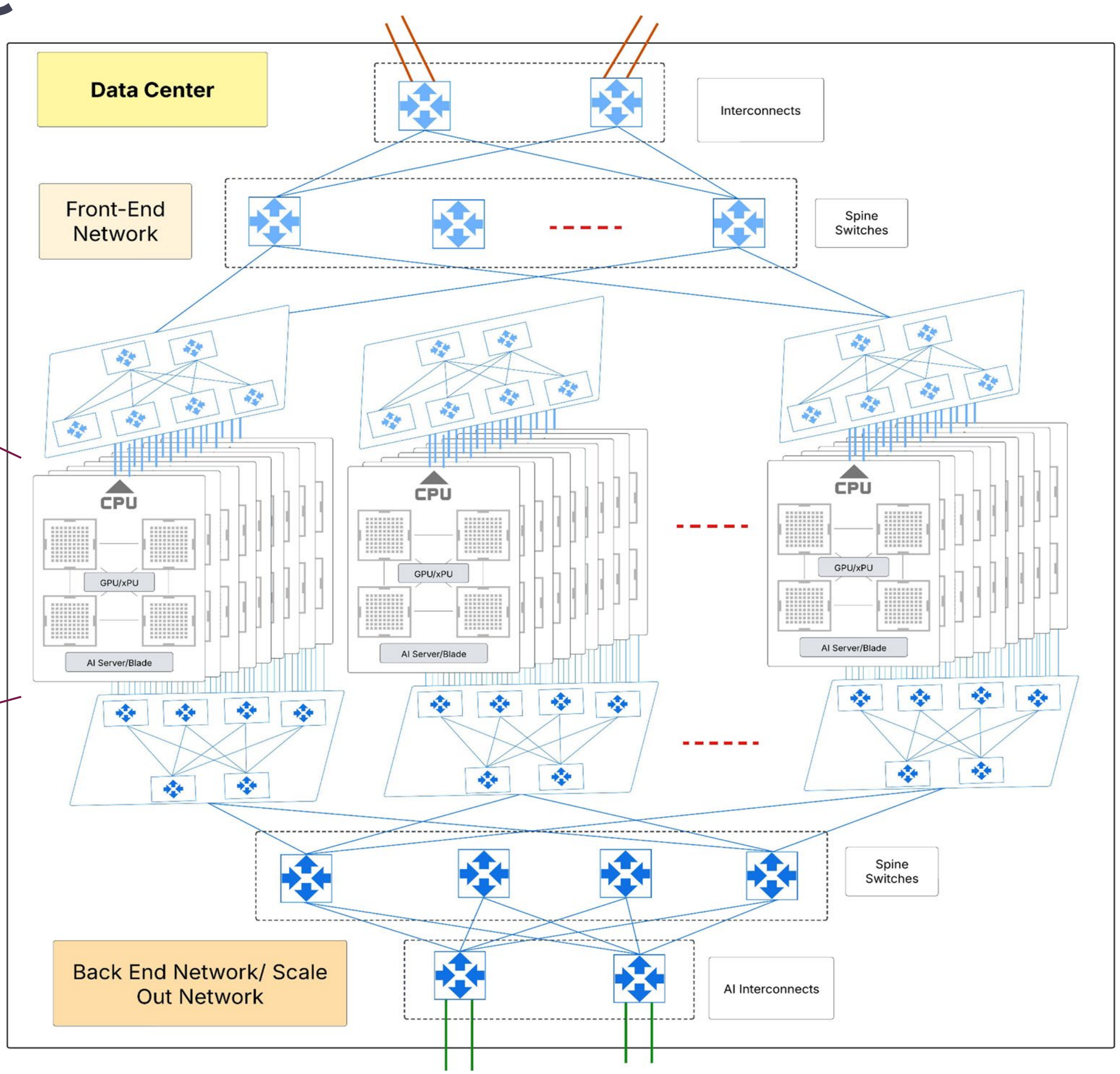
Traditional Clusters vs AI Clusters





AI Networking Big Picture





TEF 2025 Ethernet for AI

AI networks Design Considerations

Building Networks for Supporting varied AI use-cases

Motivation and Design Principles

- Supports varied use-cases (Generative AI, Training, Inference et. al)
- Address Rapid Growth in Computational Density
- Requires efficient and reliable dedicated networks
- Understanding of Workloads (Data, Pipeline, Tensor Parallel)

Network Topology Design requirements

- Dedicated ‘Backend’ networks
- Scale-up, **Scale-Out**, Scale-Across
- Redundancy, Reliability, Flexibility, Latency, and Resilience
- Deployment Heterogeneity

Network Optimization Routing Implementations

- Address Load Imbalance and burstiness
- Achieve Near-Optimal Traffic Distribution
- Congestion Control Management

Operations

- Robust maintenance procedures
- Build Tooling pipeline to Detect and Mitigate issues for these large networks

Design Choices

Technology

- Endpoint Scheduled/Non-Scheduled Fabric (NSF)
    - Low Latency , higher switch capacity
    - Simplified Cabling

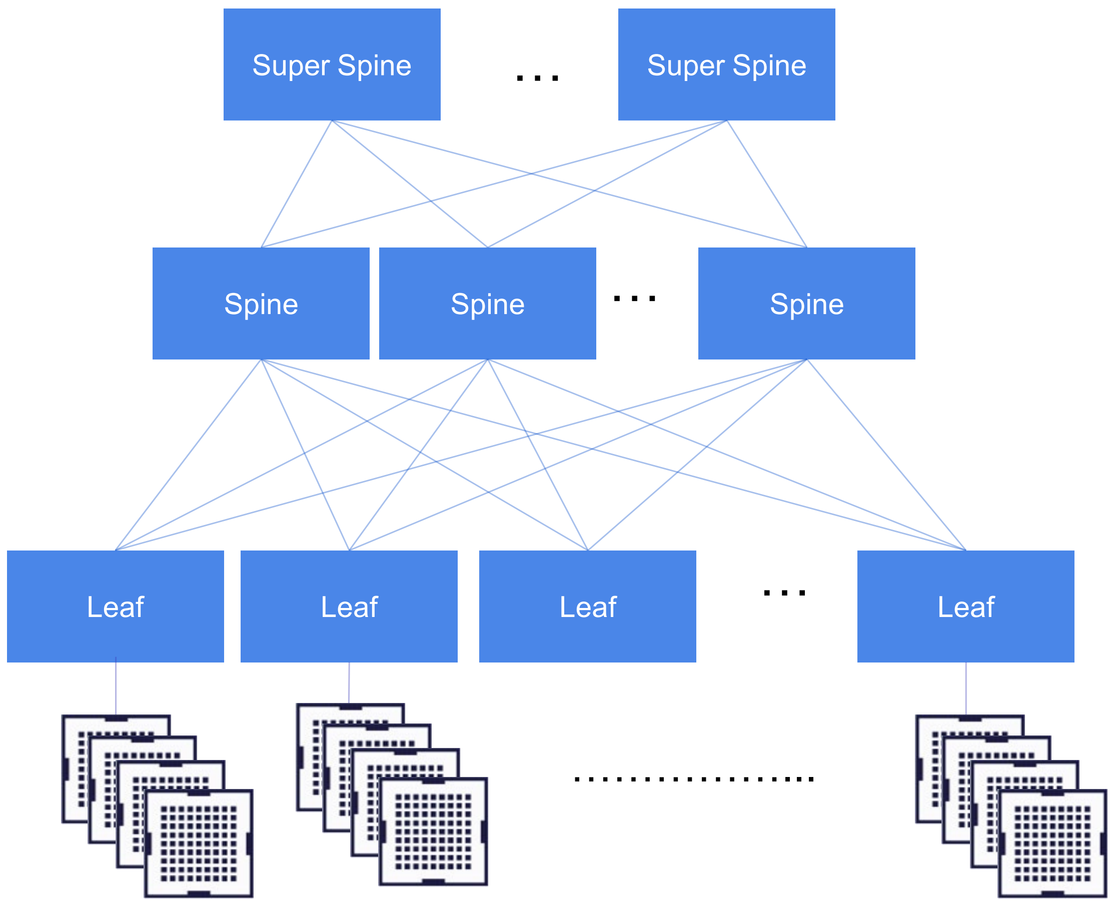
- Switch Scheduled Fabric / Disaggregated Switch Fabric (DSF)
    - NIC Agnostic
    - Deep Buffer

Similar topology choices possible with both Technology Options :

- Top of Rack (ToR)/End of Row(EoR)/Side of Row (SoR)
- Rail Based/Non-Rail, Planar (Single, Dual)
- Single/Two/Three Stage Clos Fabric

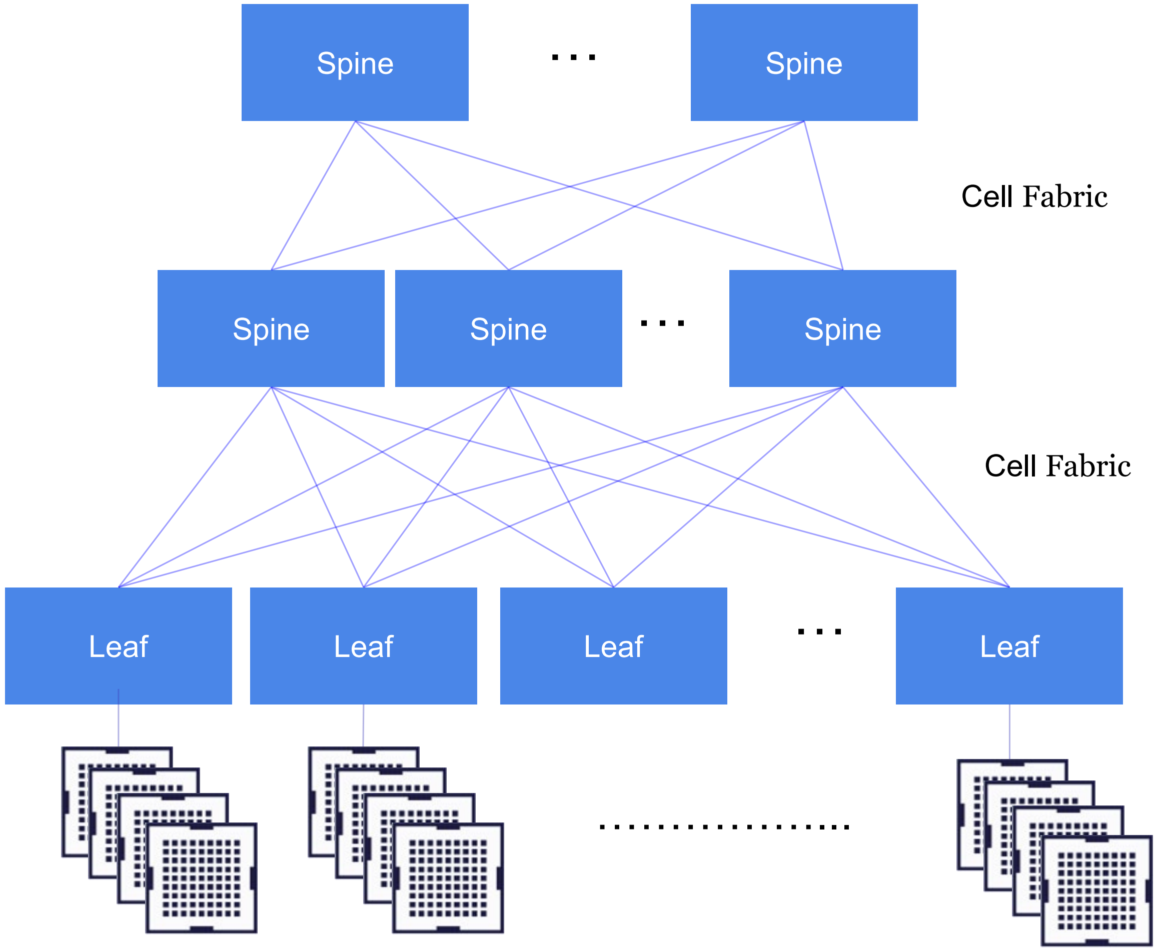
Endpoint Scheduled / Non-Scheduled Fabric

- Standard Ethernet fabric with flat, high-bandwidth, high-radix topology
- Uniform switch type across leaf/spine/super-spine for switching, forwarding, queuing, and scheduling
- Adaptive routing with end-to-end congestion control
- Endpoints implement congestion management and spraying / load balancing
- Uses standard Ethernet for RDMA, leveraging NIC capabilities (e.g., out-of-order handling) without proprietary control-plane signaling
- Standard lossless congestion-control mechanisms (ECN + PFC)
- Simpler deployment, operations, and maintenance; reuses existing non-AI network designs
- Redundant inter-switch links for higher resiliency



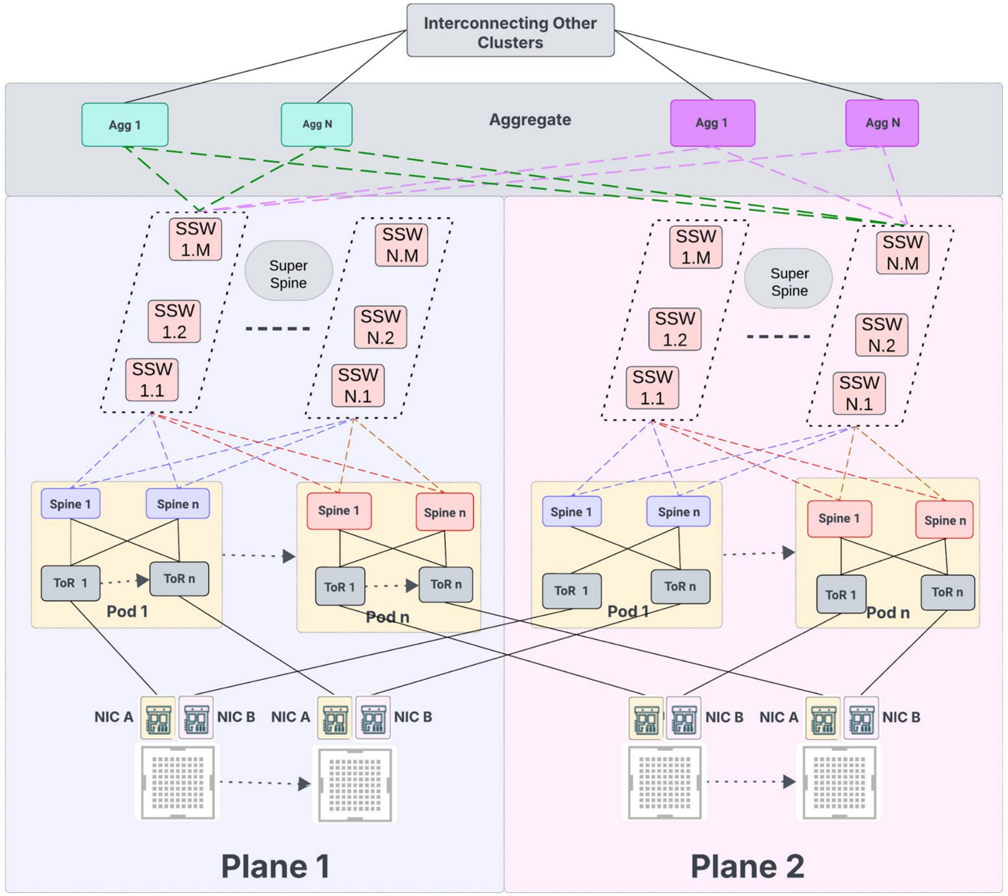
Switch Scheduled / Disaggregated Switch Fabric

- Standard Ethernet network connectivity
- NIC Agnostic Solution
- **Leaf**: switching, forwarding, queuing, scheduling
- **Spine/Super Spine**: forwarding at low power
- Lossless delivery from ingress to egress
- Cell spraying ensures no congestion
- Credit request/grant protocol ensures egress queues do not overflow
- Tiered distributed switching system
  - Scales to 4.6k x 800G or 9.2k x 400G accelerators in single system
  - With two stage can go beyond 32K+ GPU



TEF 2025 Ethernet for AI

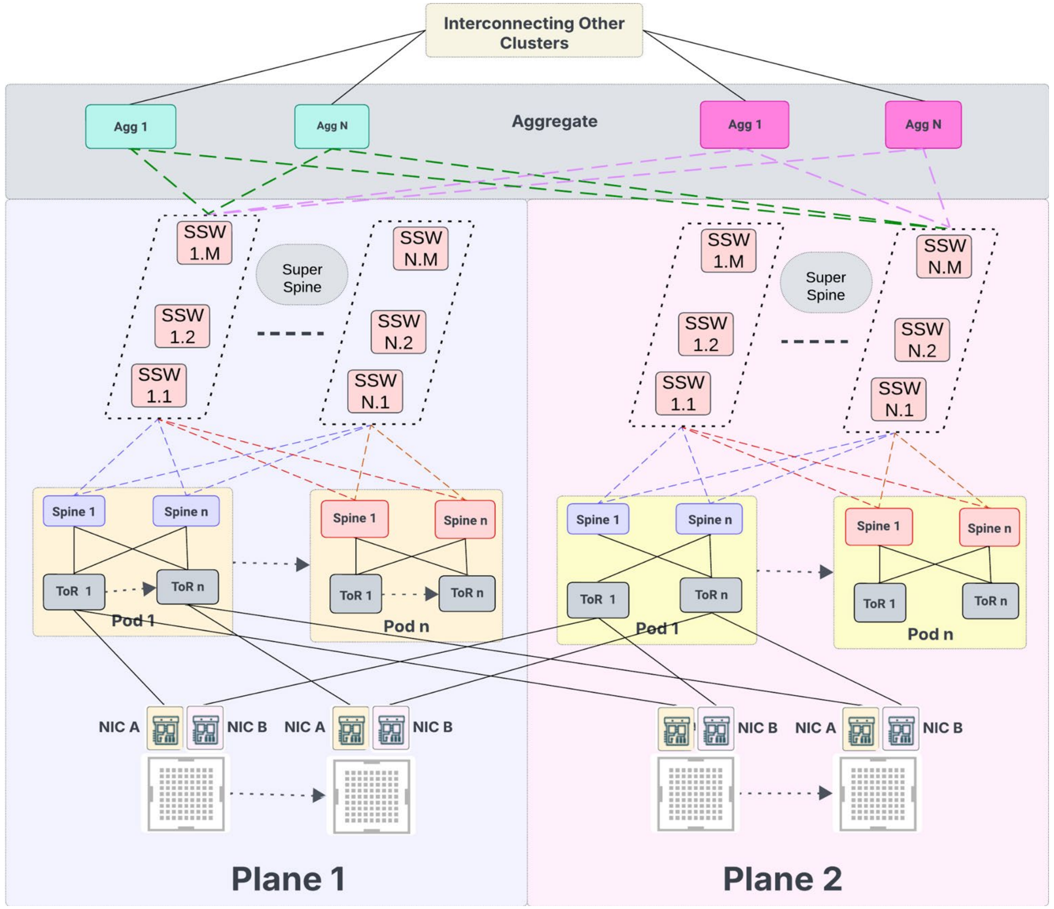
Reference Designs



Planar Design

Scale-Across

Scale-Out



Planar Design with Rail

TEF 2025 Ethernet for AI

Interconnect Options

Interconnect technology	Reach	Radix	Power/ Port	Use cases
DAC (Direct Attach Copper)	1-3m	2x400G	0W	Intra-rack
ACC (Active Copper Cable) Re-driver based	3-4m	2x400G	3W	Intra-rack and adjacent rack
LPO OSFP (Linear Pluggable Optics)	2FR4 (2 km)	2x400G	9W	NIC <> Leaf Leaf <> Spine Spine <> Spine
LRO OSFP (Linear Receive Retimed Transmit)	2DR4 (500m) 2XDR4 (2 km) 2FR4 (2 km)	4x200G 4x200G 4x200G	11W	NIC <> Leaf Leaf <> Spine Spine <> Spine
DSP Optics (Fully Retimed)	2XDR4 (2 km) 2FR4 (2 km)	4x200G	16W	NIC <> Leaf Leaf <> Spine Spine <> Spine
DSP Optics	2xLR4(10km) 2xZR4 (<10km)	4x200G 4x200G	16W/30W	Spine <> Spine Interconnects
DSP Optics (Fully Retimed, 1.6T)	2DR4 2FR4	8x200G 8x200G	25W	NIC <> Leaf Leaf <> Spine Spine <> Spine

Evolving Interconnect Requirements!

- With Interconnects Power going up, alternatives need to be looked at
- Low Power, High Density
- Operations & Serviceability
- Ability to support Multi-Vendor Eco system
- Possible Avenues: CPO, LPO, Cabled/Optical Backplanes

Summary

- Scale changes everything
- Performance is frequently limited by power, physical layout, system reliability, and other physical constraints
- As we continue to scale, a new wave of challenges emerges across cost, reliability, data, and power
- Innovative solutions are required to keep improving both efficiency and flexibility

TEF 2025 Ethernet for AI

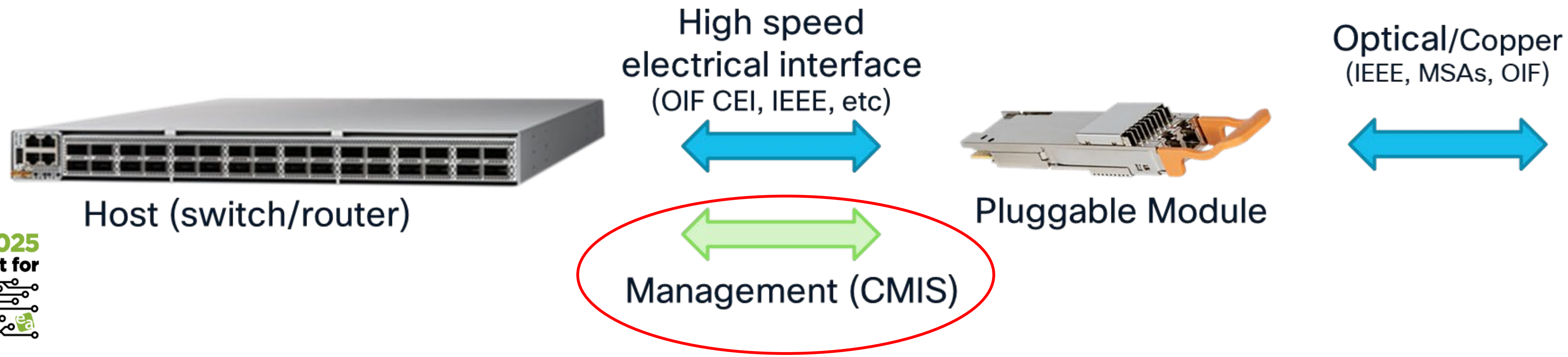
CMIS – The Interface that ties everything together for AI

Gary Nicholl, Distinguished Engineer, Cisco
OIF PLL Working Group, Management Co-Vice Chair

Acknowledgements: This presentation was put together with the help of OIF PLL Working Group, Management Co-Vice Chair, Ian Alderdice, Ciena.

What is CMIS ?

- **CMIS (Common Management Interface Specification)** is an industry standard management interface for high-speed modules, and is defined in a family of (OIF) documents
- It defines how hosts configure, initialize and monitor pluggable optics
- It is implemented (through SW code) on the host and on the module
- It has become ubiquitous across the industry and used in QSFP-DD, OSFP and next generation 800G/1.6T module form factors
- It provides a unified management approach across vendors, module form factors and interface technologies ranging from copper cables to long-reach coherent optical interfaces



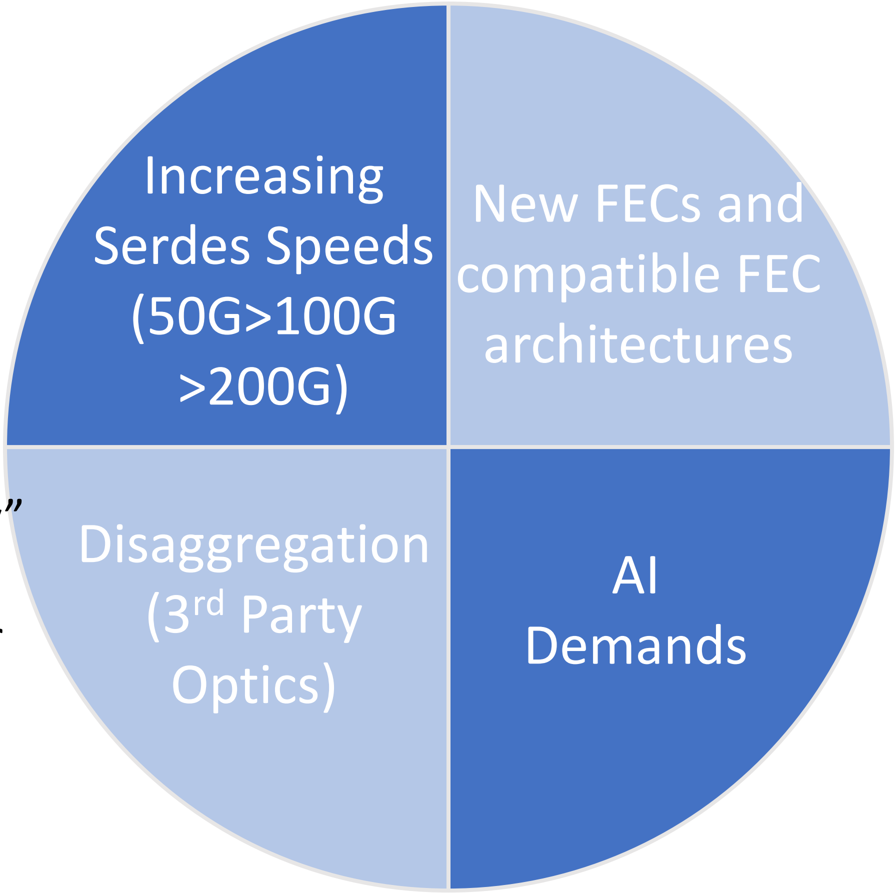
The key drivers behind CMIS ?

CMIS was originally developed by the QSFP-DD MSA to address several industry trends:

Now standardized in the OIF

More complex EQ strategies, demanding closer coordination between host and module to ensure reliable initialization.

More stuff to configure, initialize and monitor within module. Better coordination between host and module (different initialization times…). FW Upgrade.



Management becomes a true "3rd party" interoperability interface. Plug'n'Play expectation - any module should power up, initialize and carry traffic.

Demands higher levels of link quality and performance assurance (monitoring and telemetry are critical)

TEF 2025 Ethernet for AI

Why CMIS is important for AI Networks

- AI clusters require extremely high throughput and low latency
- Distributed training is sensitive to link errors and/or failures
- Predictive monitoring is essential for reliability
- Building out at massive scale (thousands of optical modules), and at speed, demands a consistent management solution

CMIS Benefit #1 – Vendor Interoperability

- Standard behavior across vendors
- Simplifies qualification, buildout and operations
- Reduces development and integration time
- Ensures consistent monitoring and control

AI Impact: Much easier scaling to AI clusters containing thousands of modules with potentially mixed suppliers and interface technologies.

CMIS Benefit #2 – Faster Module Initialization

- Standardized state machines to control both power-up and initialization
- Predictable and reliable transitions (LowPwr > Initialized > Application Ready)
  - Based on advertising: Technology dependent transition times
- Improved stability during provisioning, initialization and restarts

AI Impact: Reduces downtime and speeds up cluster availability

CMIS Benefit #3 – Advanced Telemetry

- Module temperature, voltage and current reporting
- Per-lane optical power (Tx/Rx) monitoring
- BER, preFEC BER, FEC bin histogram and signal integrity counters
- Real time fault flags (LOS, LOL, lane degradation, etc)

AI Impact: Potentially detects failing links before they disrupt training jobs

CMIS Benefit #4 – Support for Advanced Optics

- Breakout support (1x800G > 2x400G > 8 x 100G)
- Supports coherent optics for AI data center interconnect (AI-DCI): **C-CMIS**
- Works with line-drive optics (LPO) for low power / low latency fabrics: **CMIS-VCS**
- A single core management interface for many optical technologies

AI Impact: Flexible network architectures for large-scale AI deployments (scale up, scale out and scale across)

Summary

- CMIS supports AI networks by providing:
  - ❖Advanced telemetry for predictive monitoring
  - ❖Fast, reliable and predictable module initialization
  - ❖Unified management for next generation optics (from copper to long reach coherent)
  - ❖Power and thermal adaptability
  - ❖Interoperability across vendors
- **Essential for:** Scalable, reliable high-performance AI networks

Note: The OIF continues to drive CMIS enhancements to support these evolving industry needs, through its diverse membership and strong technical leadership in the interconnect space.

QUESTIONS?