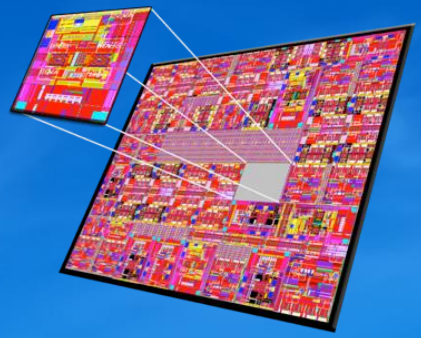
Ethernet for AI Networking

Peter Jones

CISCOU-2038

By Way of Introduction ...



I am a **Distinguished Engineer** in the Cisco Networking Hardware team and have been with Cisco since 2005.

I work on system architecture and standards strategy across the portfolio. I was a key figure in the development of the UADP switching ASIC architecture and the Catalyst switches that use it.

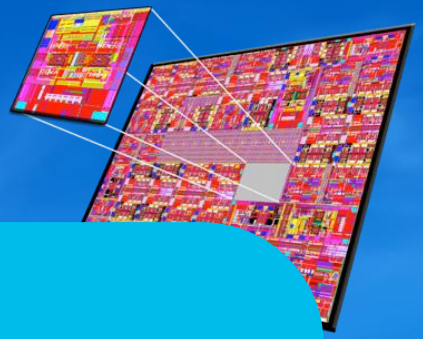
I work in defining and promoting new Ethernet standards in IEEE 802.3 and as Ethernet Alliance Chairman.

I am passionate about Network Evolution and Adoptable Technology.

Peter Jones
Distinguished Engineer

Email: petejone@cisco.com
Bluesky: petergjones.bsky.social
LinkedIn: in/petergjones/

By Way of Introduction …



Too Much Detail! Look at https://petergjones.substack.com/

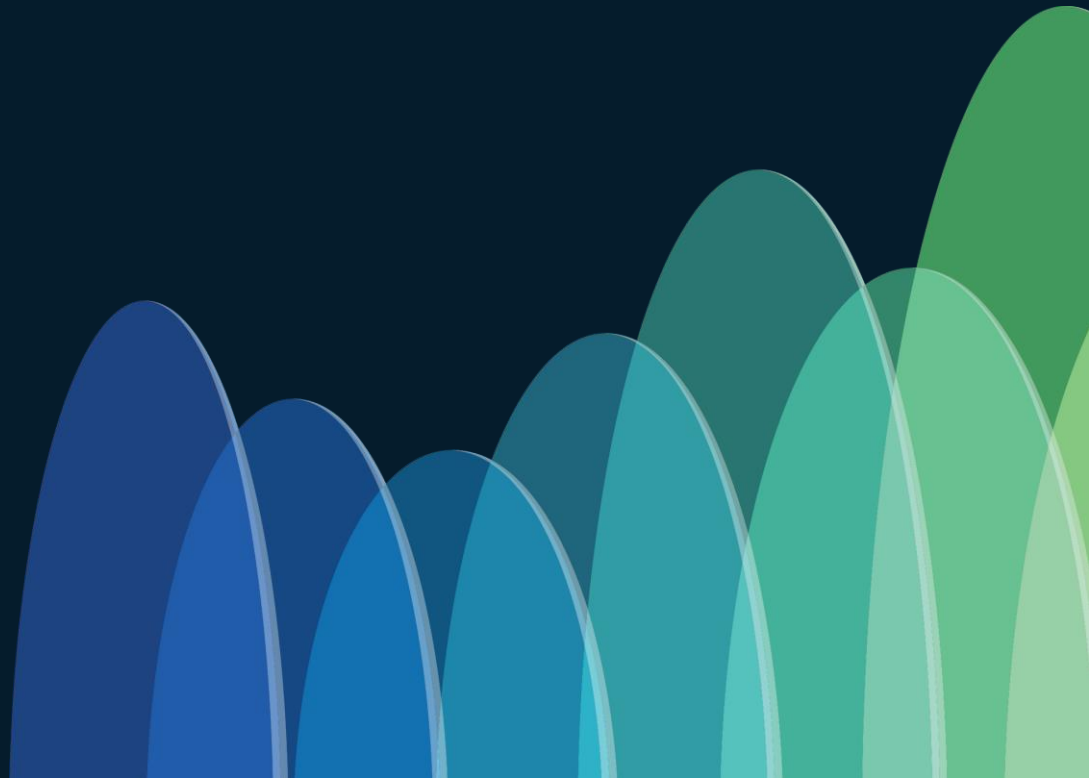
Email: petejone@cisco.com
Bluesky: petergjones.bsky.social
LinkedIn: in/petergjones/

Agenda

- What's the problem?
- Where are we?
- What's next?

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

What's the problem?



GenAI spend is ENORMOUS!

GenAI is upending the global IT spend.

- The Hyperscalers spent ~$180B in infrastructure alone in 2024[1].
- AI accelerator silicon revenue grew 130% in 3Q 2024[2].
- DC switching and NIC markets will double to >$50B in 5 years[3].
- Hyperscalers are investing in renewable and nuclear power[4].
- Blackstone estimates a 40% increase in electricity demand in the United States over the next decade[5].

1. CIO Dive: Big tech on track to pour more than $180B into data centers this year
2. Dell'Oro: US Hyperscalers Set to Deploy Over 5 Million AI Training-Capable Accelerators in 2024
3. Crehan Research: Ethernet switch and NIC market to reach $50 Billion in the next five years
4. D'Ornano + Co: The Great Infrastructure Bottleneck: Why GenAI's Next Phase is About Atoms, Not Bits
5. The Motley Fool: Blackstone (BX) Q2 2024 Earnings Call Transcript

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

Why does the network matter for AI/ML?

Process

Execute instructions on GPU(s)

- **Compute the calculations** to inference / train a model

Synchronise

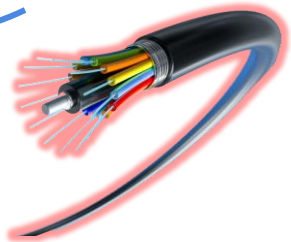
Wait for everyone to complete

- **Creates synchronization between GPUs**
- GPUs are engaged to a 'ready state' for the next stage of the job
- **Computation stalls waiting for the slowest path**
- **Job Completion Time (JCT) is based on the worst-case tail latency**

Notify

Send results of the computation

- **GPU-to-GPU Memory transfer**
  All-to-all Collective
  (Everyone sends to everyone, eg: All-Reduce)
- **High bandwidth compute can saturate network links**



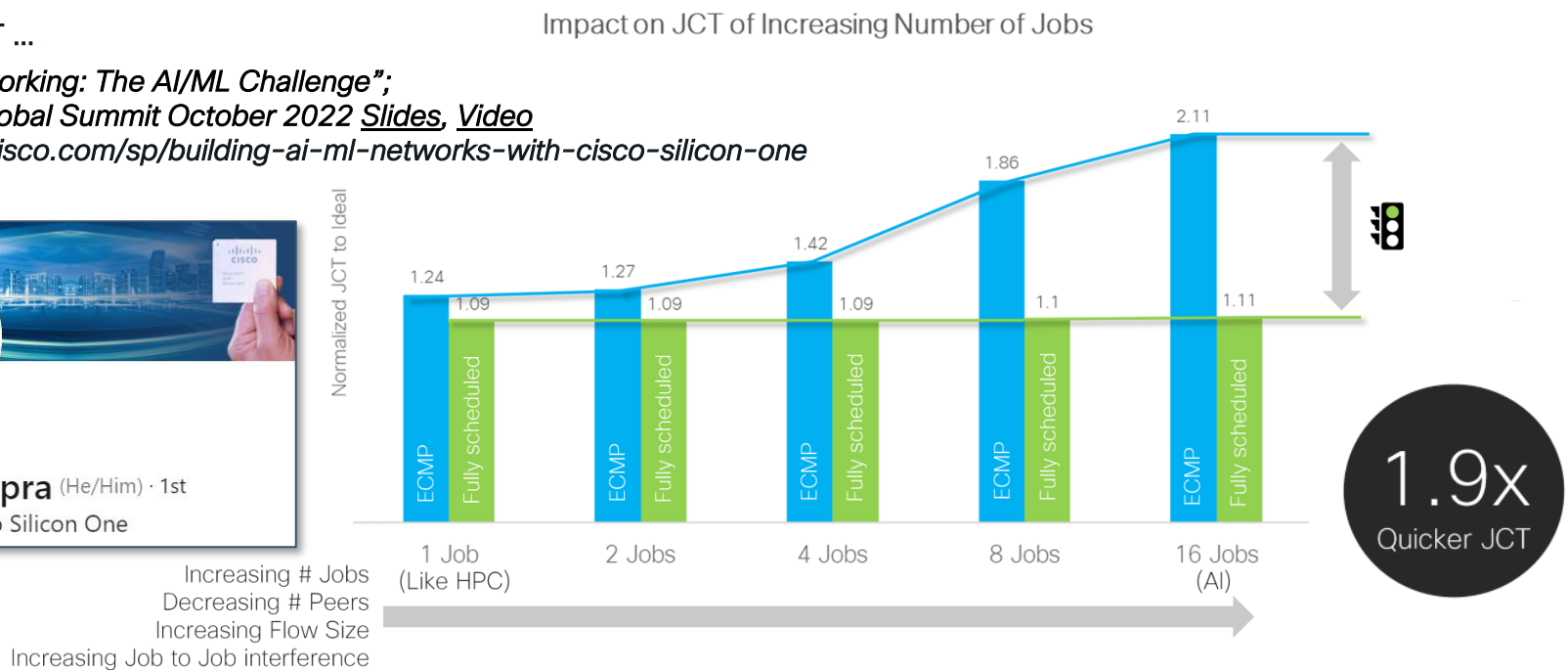
© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

Building Networks for ML/AI Workloads

Optimized Job Completion Time(JCT) with Fully Scheduled Fabric

MORE INFO AT …

"Evolved Networking: The AI/ML Challenge";
OCP Global Summit October 2022 Slides, Video
https://blogs.cisco.com/sp/building-ai-ml-networks-with-cisco-silicon-one



© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

How do I get the most out of $Bs of GPUs

Giorgio Trovato on Unsplash

GDJ on Open Clipart

The network exists to enable the GPUs to stay busy.

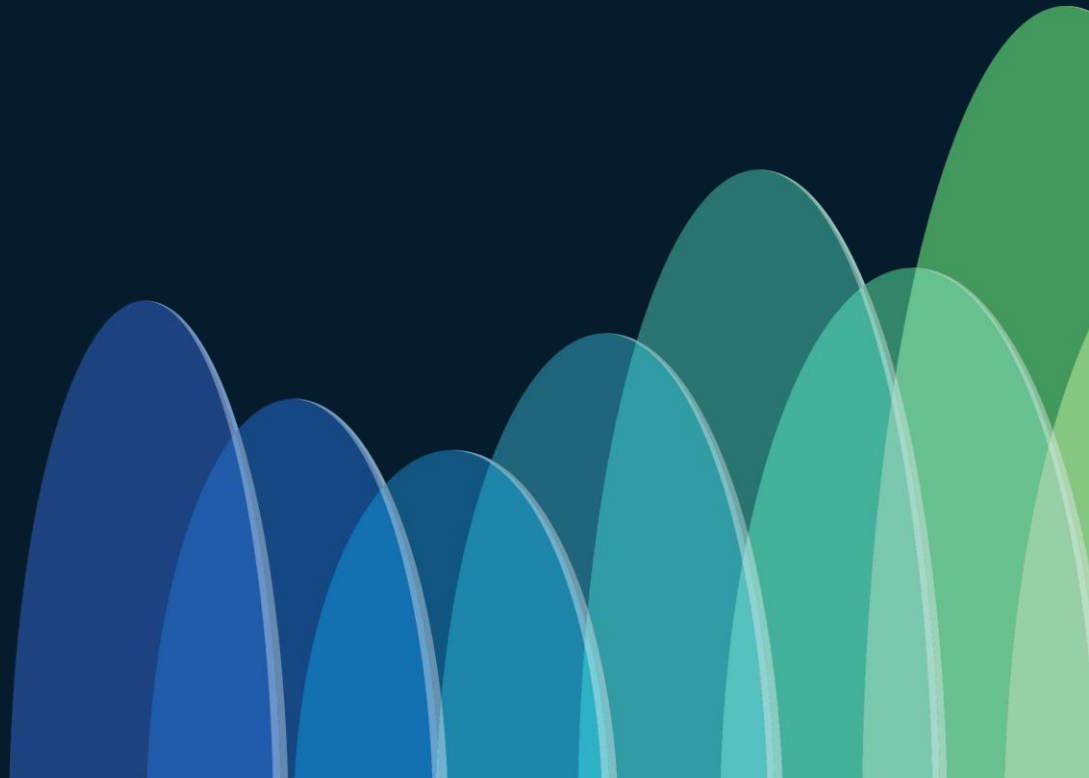
A ***watt*** spent on the network is a ***watt*** not spent on the GPUs

What matters?

- Throughput under full load
- Reliability/Resilience
- Power

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

Where are we



Terminology

- Back End Network
  - Handles data movement between GPUs[1]
- Front End Network
  - Handles the interactions between the training clusters and the rest of the DC[1]
- Scale-Out
  - A network architecture that interconnects AI compute clusters together, typically using standard networking technologies[2]. This is part of the Back End Network.
- Scale-up
  - A network architecture within an AI compute cluster that maximizes intra-cluster communication bandwidth and minimizes latency[3] . This is part of the Back End Network.

1. Sujal Das/Enfabrica: Evolution of Data Center Networking Designs and Systems for AI Infrastructure
2. Ayar Labs: Scale-Out
3. Ayar Labs: Scale-Up

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

Backend Network Technologies

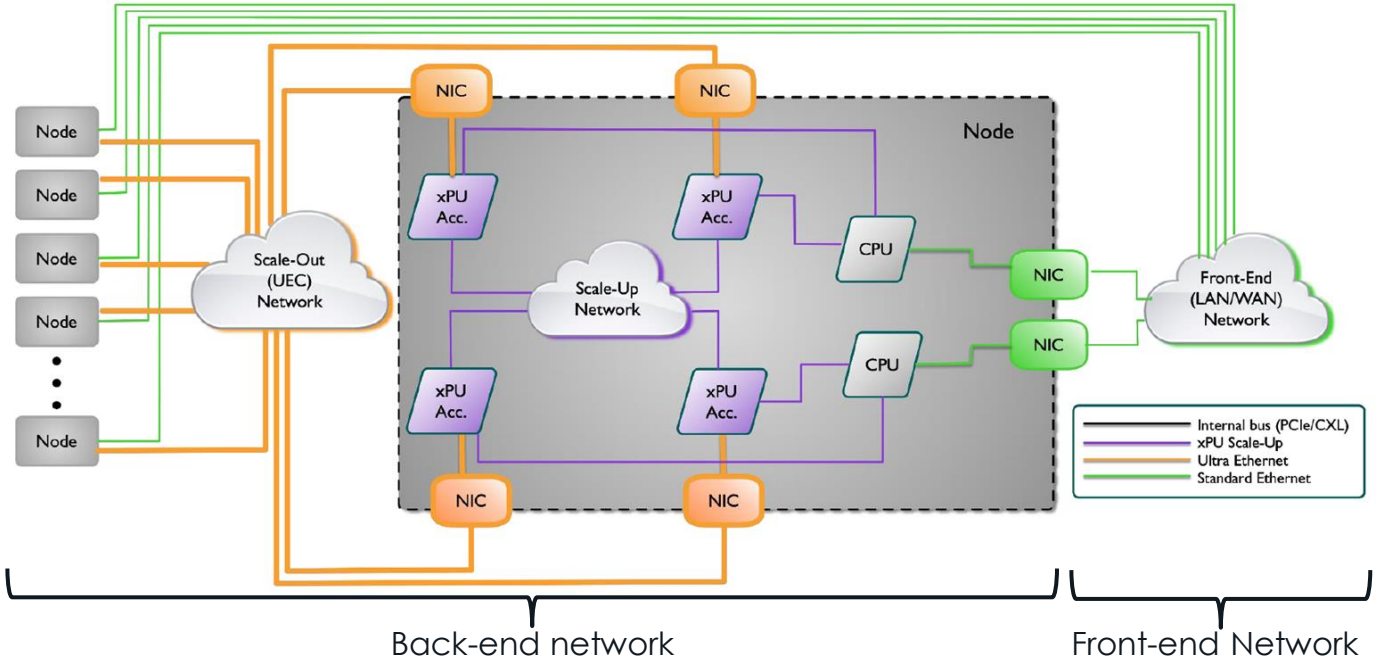
Scale-Up

- NVLink [3]
- Infinity Fabric [4]
- UALink

Scale-Out

- Infiniband [2]
- Ethernet/Ultra-Ethernet

General Purpose vs **Scale-Up** vs **Scale-Out (UEC)** Networks



1. Sujal Das/Enfabrica: Evolution of Data Center Networking Designs and Systems for AI Infrastructure
2. TechTarget: What is InfiniBand?
3. NVIDIA: What Is NVLink?
4. Dell: Understanding the Value of AMDs Socket to Socket Infinity Fabric

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

InfiniBand vs Ethernet: Nvidia[1]

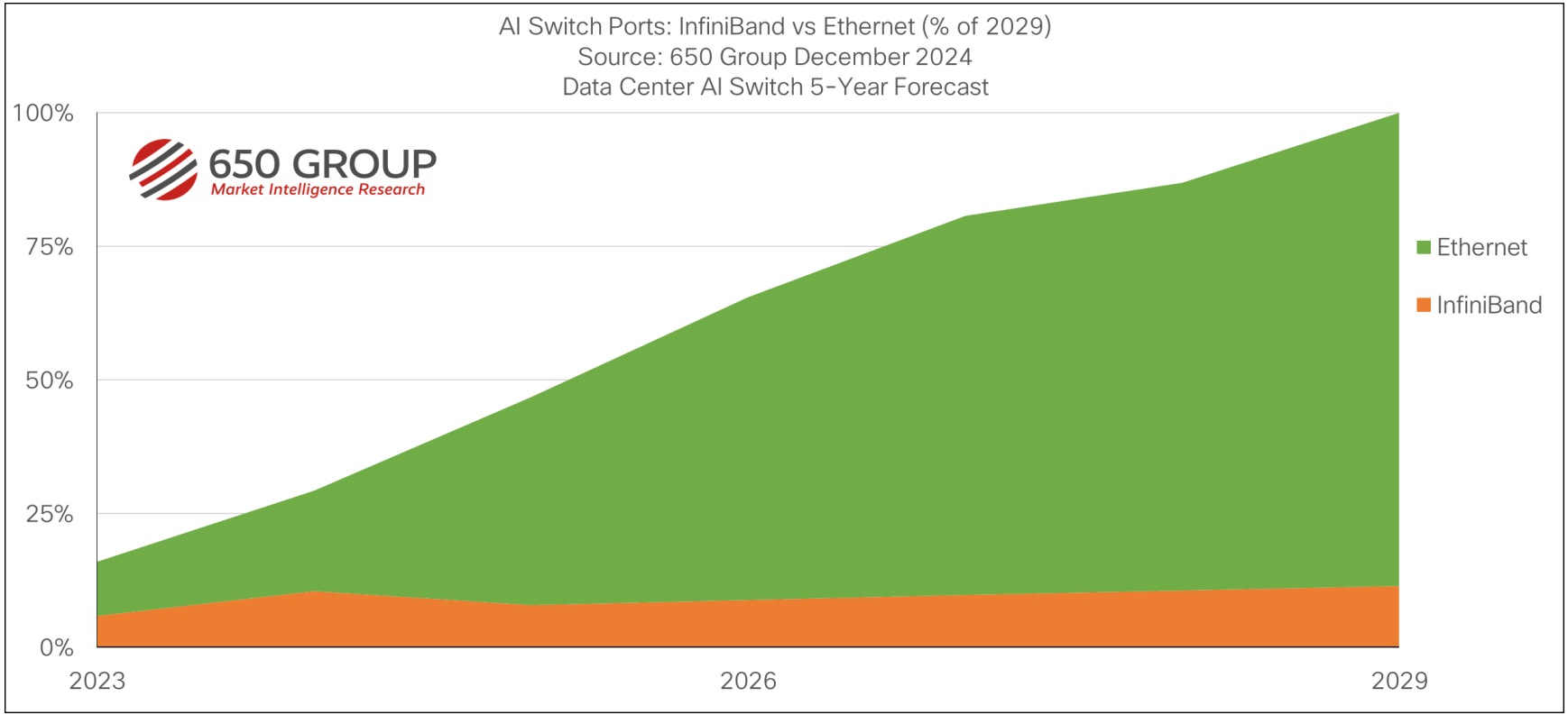
For AI training and inference infrastructure among the hyperscalers and cloud builders, Nvidia will tell you plain and simple that the network represents 20 percent of the cluster cost.

InfiniBand, explains Nvidia co-founder and chief executive officer Jensen Huang, delivers 20 percent better performance at scale at the same bandwidth than Ethernet, however, so "InfiniBand is effectively free."

1. The Next Platform: META PLATFORMS IS DETERMINED TO MAKE ETHERNET WORK FOR AI

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

Ethernet vs InfiniBand – AI Backend Networks



© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

What's driving Ethernet?

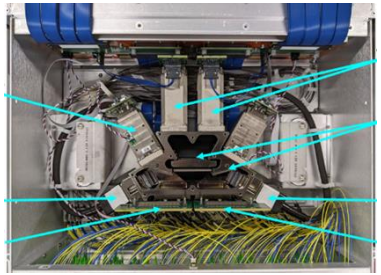
- Scale
  - Hyperscalers are looking to build very large training clusters (300,000+) [1], have clusters span multiple DCs[1], and InfiniBand has scaling limitations.
- Supplier Diversity
  - Nvidia(Mellanox) dominates the InfiniBand market[2].
- Cost of Operations
  - History shows that Ethernet becomes less expensive to own and operate than the technologies it replaces.
  - Everyone has Ethernet, using one technology reduces operational cost

1. SemiAnalysis: Multi-Datacenter Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure
2. NADDOD: Where to Buy Infiniband products

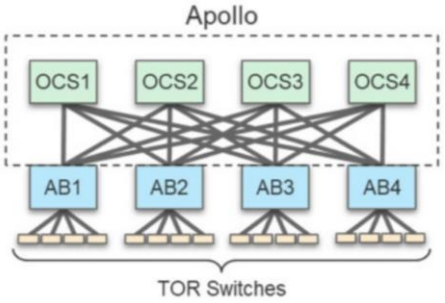
© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

RFC 1925 rule 10 – “One size never fits all”.

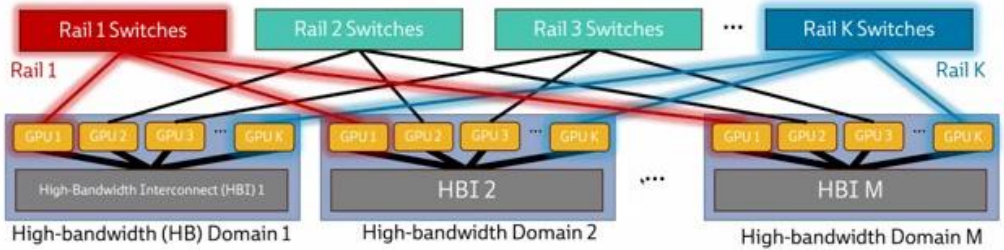
Google uses Custom Optical Switches[1] in its Jupiter network architecture[2].



Paloma OCS



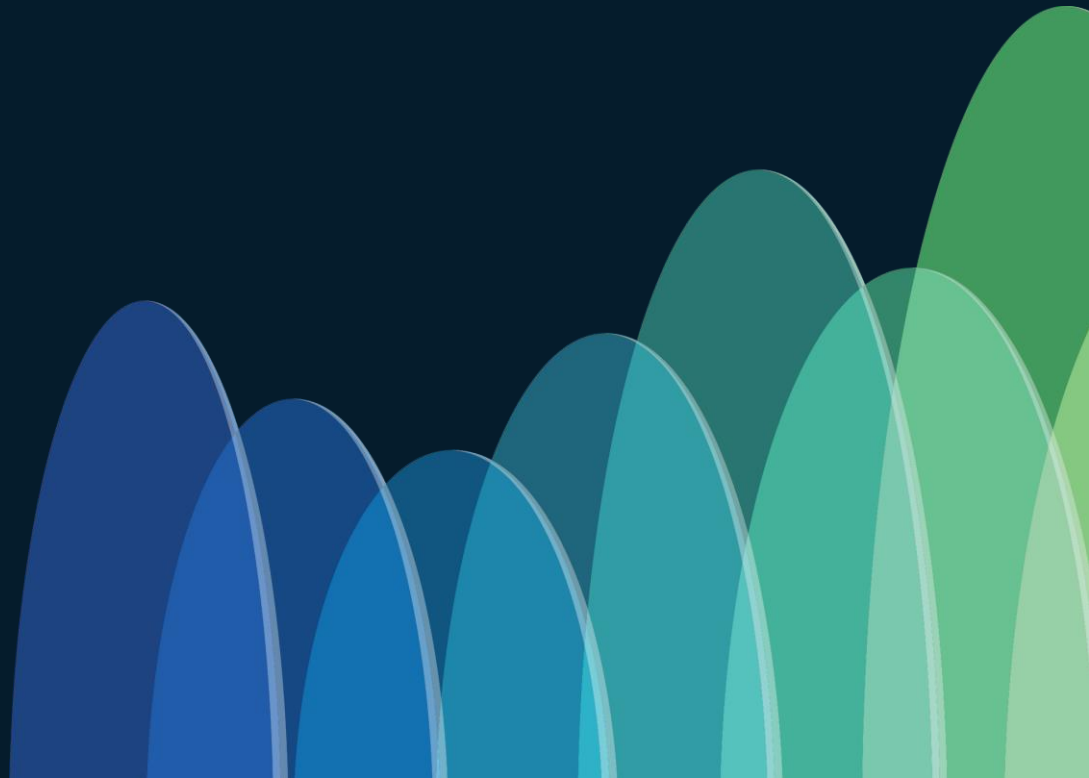
Meta has a “Rail-only” design. [2]



1. SemiAnalysis: Google OCS Apollo: The >$3 Billion Game-Changer in Datacenter Networking
2. Google: Speed, scale and reliability: 25 years of Google data-center networking evolution
3. NextPlatform: This AI Network Has No Spine – And That’s A Good Thing

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

What's next?



Ethernet for AI Networks: Who's doing What

Ethernet Alliance

- Building cross industry consensus, e.g. TEF 2024: Ethernet in the Age of AI.

IEEE 802.3

- IEEE P80
- NEA star

Ultra Ethern

- Optimize networks.

> RFC 1925 rule 12:
> "In ~~protocol~~ network design, perfection has been reached not when there is nothing left to add, but when there is nothing left to take away".

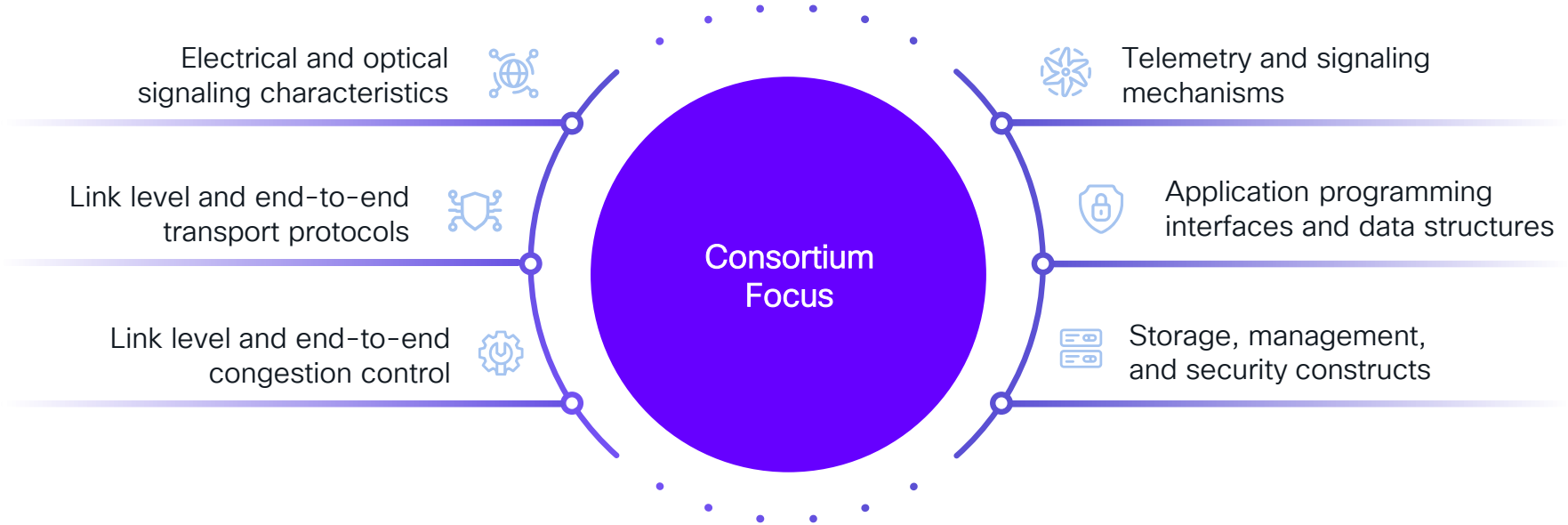
Adjacent: Ultra Accelerator Link™ (UALink™)

- Create an open industry standardized Interconnect for Accelerator-to-Accelerator communication.

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

UEC Focus

Open specifications, APIs, source code for optimal performance of AI and HPC workloads at scale.



1. UEC White Paper

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

Net/Net

Ethernet will become the dominant technology for Scale-Out networks.

Ethernet will add “just enough” to make it viable compared to InfiniBand.

Even though the Hyperscalers want to reduce cost, the network is a small part of the system.

A more efficient network that meets requirements (scale, suppliers, etc) is worth a lot ***IF*** it can increase GPU utilization.

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

Webex App

Questions?

Use the Webex app to chat with the speaker after the session

How

1. Find this session in the Cisco Events mobile app
2. Click “Join the Discussion”
3. Install the Webex app or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until February 28, 2025.

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

Fill Out Your Session Surveys

Participants who fill out a minimum of 4 session surveys and the overall event survey will get a unique Cisco Live t-shirt.

(from 11:30 on Thursday, while supplies last)

All surveys can be taken in the Cisco Events mobile app or by logging into the Session Catalog and clicking the ‘Participant Dashboard’ link at https://www.ciscolive.com/emea/learn/session-catalog.html.

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public

Continue your education

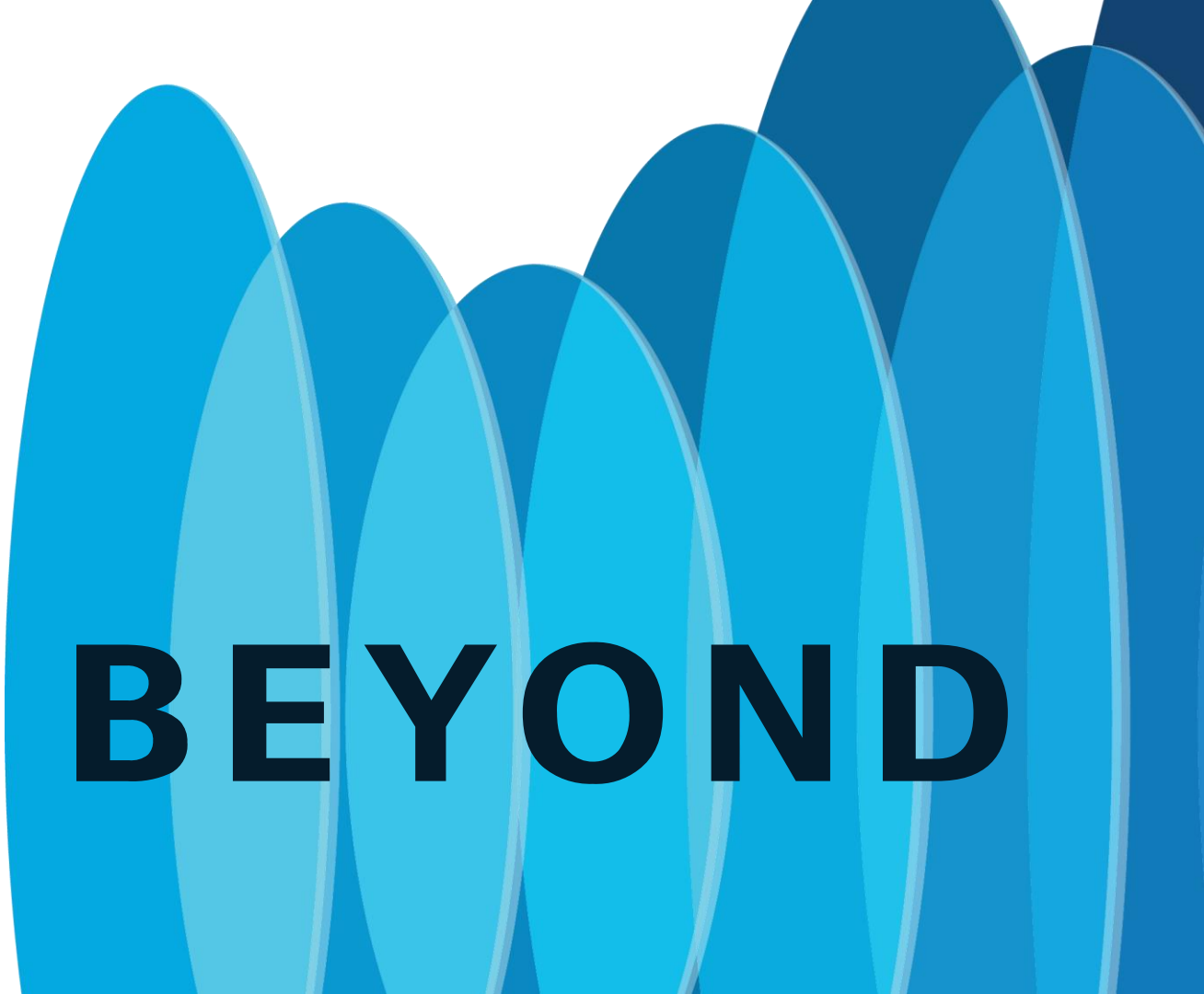
- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at ciscolive.com/on-demand. Sessions from this event will be available from March 3.

Contact me at: Insert preferred comms method

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco Public


cisco
Thank you
cisco Live!

cisco Live!



GO BEYOND