Cisco Silicon One & Ultra Ethernet for AI Infrastructure

Ramesh Sivakolundu
Director, Technical Marketing

BRKNWT-2508

Cisco Webex App

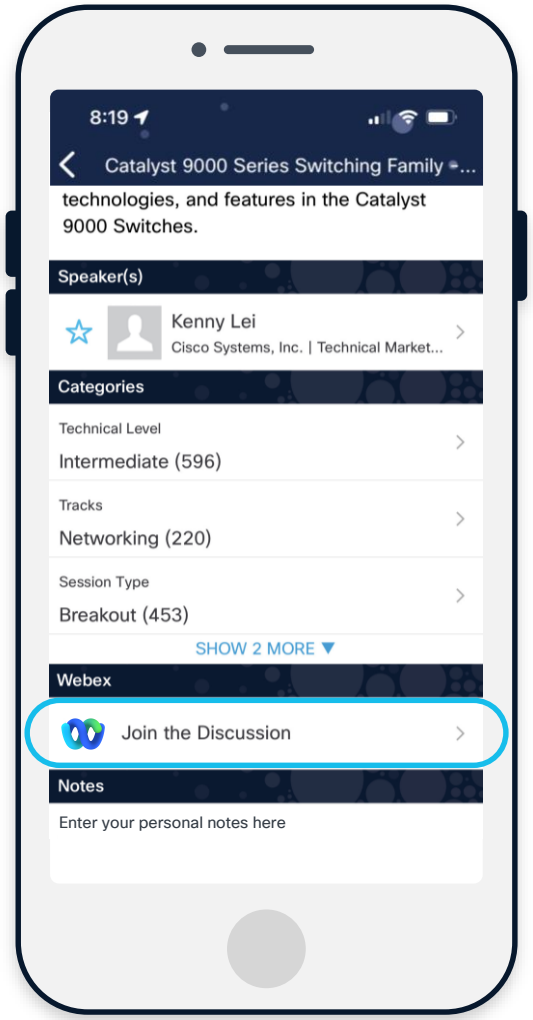
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live 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 June 13, 2025.



Agenda

Brief State of AI Networking

AI Workflows and the Training Network Bottleneck

Today’s AI infrastructure Options

Addressing Trends - Ultra Ethernet

Building AI Infrastructure with Silicon One

Wrap-up and Questions

Brief State of AI Networking

Cisco has solid products that can address the full range of data center switching use cases, from midmarket to large enterprise data centers, including open networking and AI Ethernet fabric.

Gartner, 2025

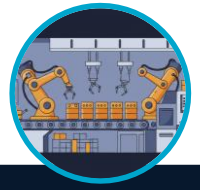
“As AI becomes more pervasive, we are well positioned to help our customers scale their network infrastructure, increase their data capacity requirements and adopt best-in-class AI security,”

Cisco CEO Chuck Robbins, 2025

Artificial Intelligence outcomes span every industry

Government

- Deliver enhanced citizen services
- Data-driven policy decisions and creation
- Modernizing and streamlining operations
- Optimizing infrastructure management
- AI-powered traffic design and public safety



Manufacturing

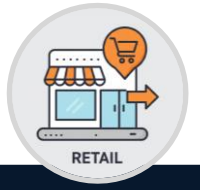
- Intelligent quality control
- Proactive machine maintenance
- Digital twin creation
- Supply chain optimization and tracking
- Optimizing production processes

Finance

- Predictive trading algorithms
- Fraud detection and prevention
- Personalized financial advice
- Investment portfolio optimization
- Virtual assistants and seamless transaction experiences

Healthcare

- Medical imaging analysis
- Enhance diagnosis and treatment
- Patient management through predictive analytics
- Improved access to healthcare with remote monitoring tools
- Drug research and development



Retail

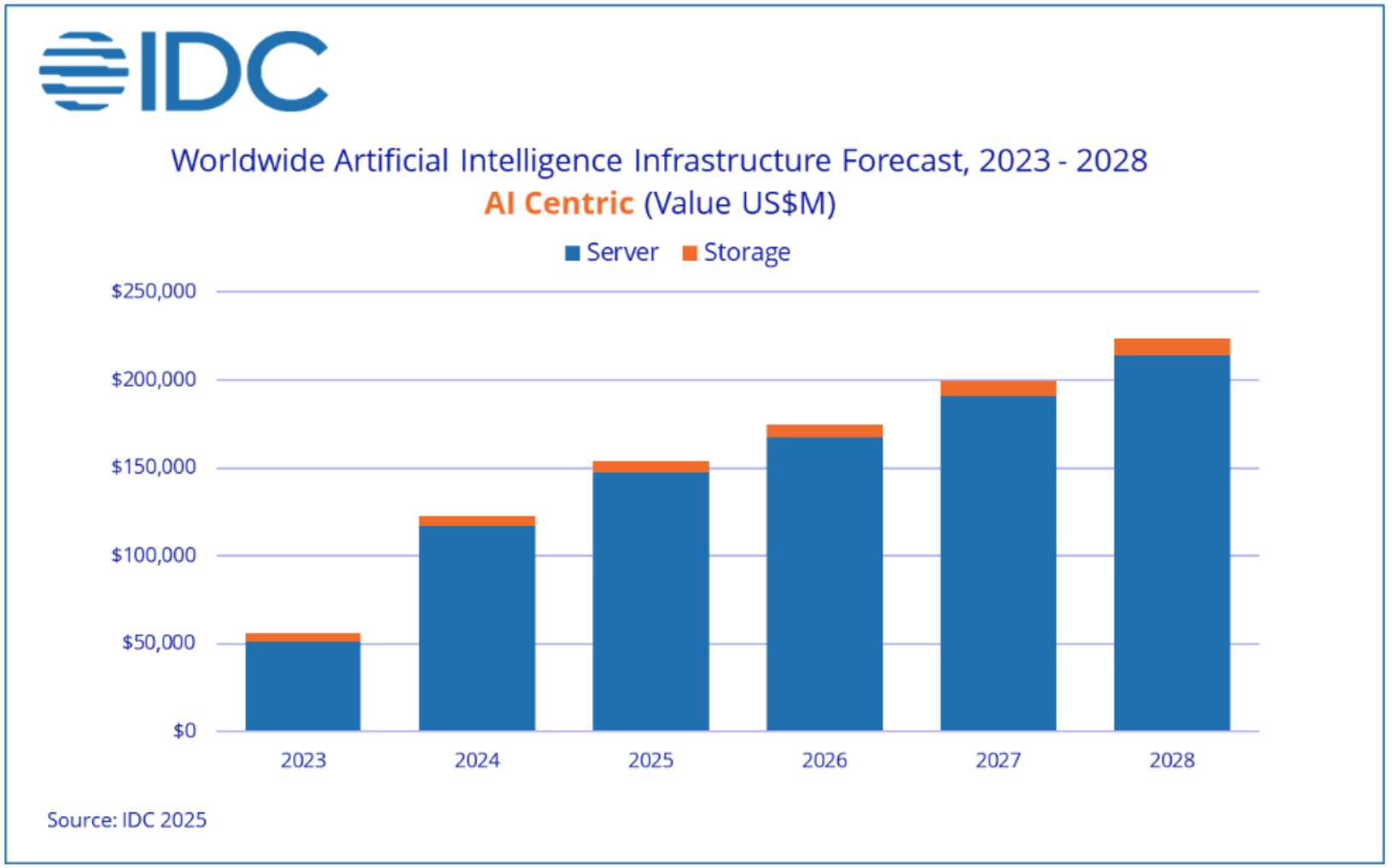
- Consumer behavior analytics
- Enhanced customer experiences
- Personalized product recommendations
- Dynamic virtual shopping experience
- Demand analysis and prediction

Build the model | **Training**

Optimize the model | **Fine-tuning and RAG**

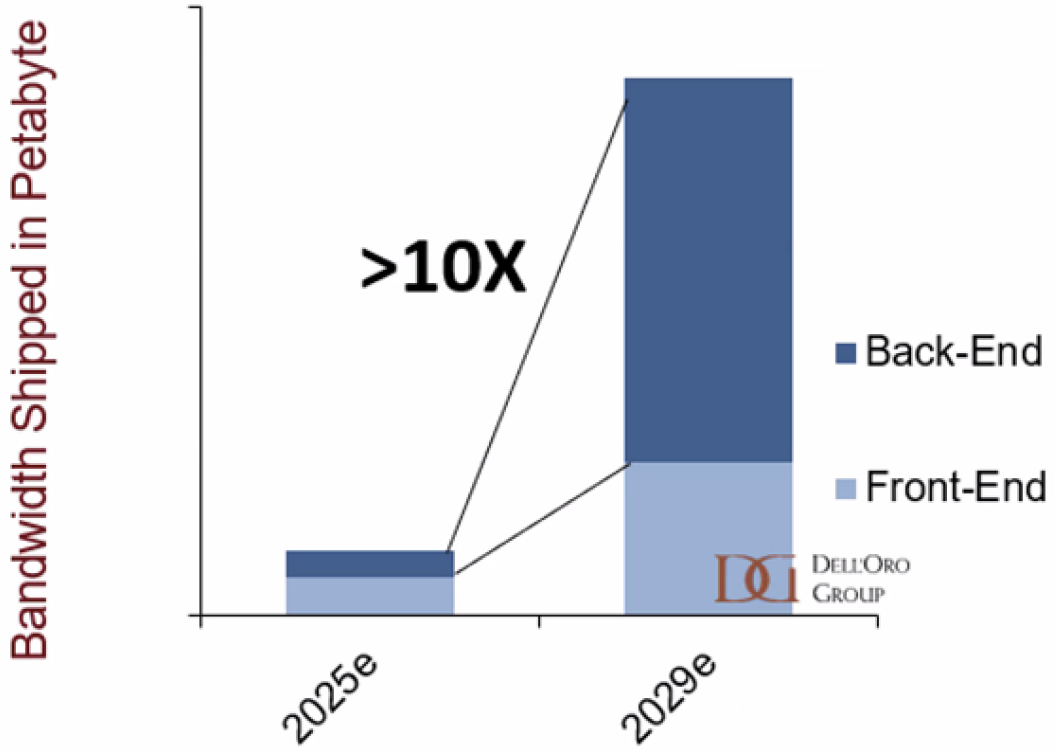
Use the model | **Inferencing**

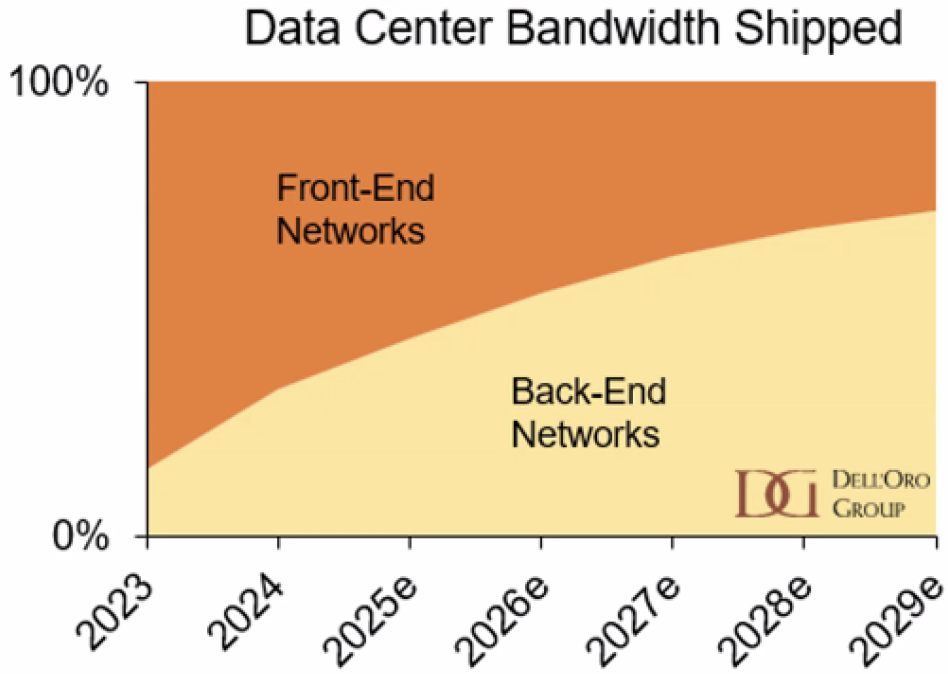
AI Infrastructure: Spend Forecast



Source: IDC Worldwide Semiannual Artificial Intelligence Infrastructure Tracker

AI Infrastructure: Bandwidth Requirements





Source: Dell'Oro Group ADVANCED RESEARCH REPORT AI NETWORKS FOR AI WORKLOADS MARKET FORECAST (January 2025)

The Elephant in the Room

Performance Considerations

System Radix
What Network type will allow me to scale out more efficiently?

Network Performance & Scale
Will the network be performant at scale?

Multi-Tenancy and Data Security
Can I keep my customer's training data sovereign and protected?

Multi-Job Performance
How will the network handle simultaneous jobs?

Operational Considerations

Multi-Vendor Support
How many vendors support the technology?

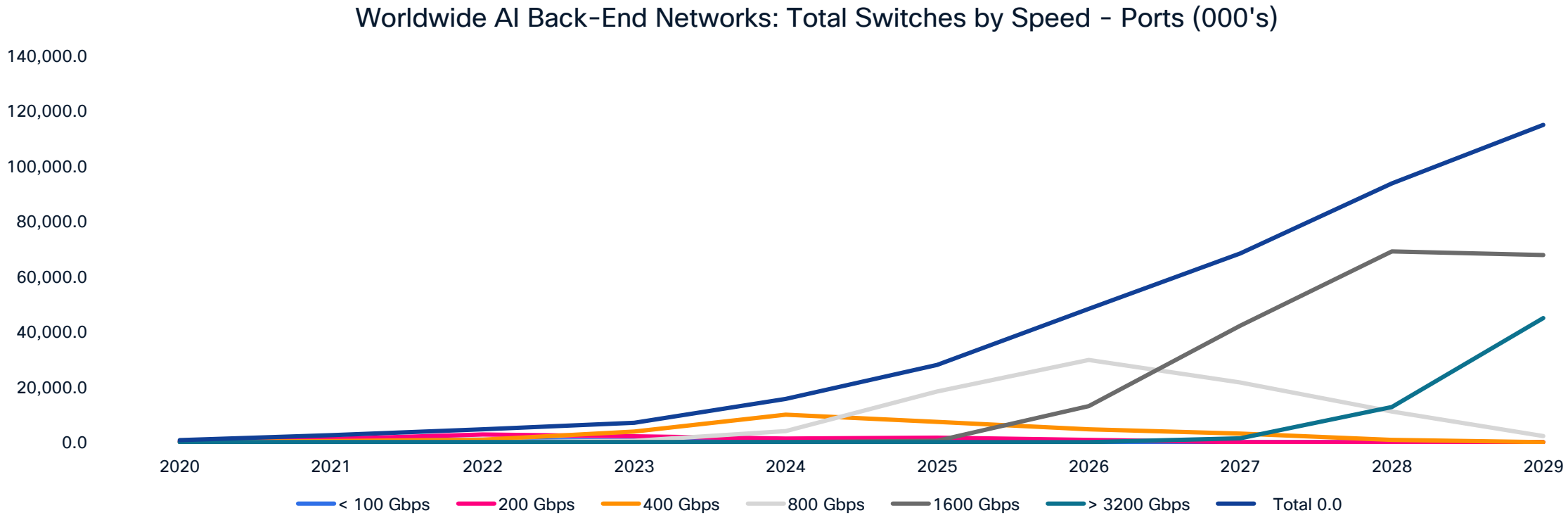
Support for customer-built AI Machines
Is the network flexible to support multiple GPU types?

Fault Tolerance for Optics Failures
Can my network handle a failed link mid-job

Talent availability
Can I hire experts to run my AI clusters?

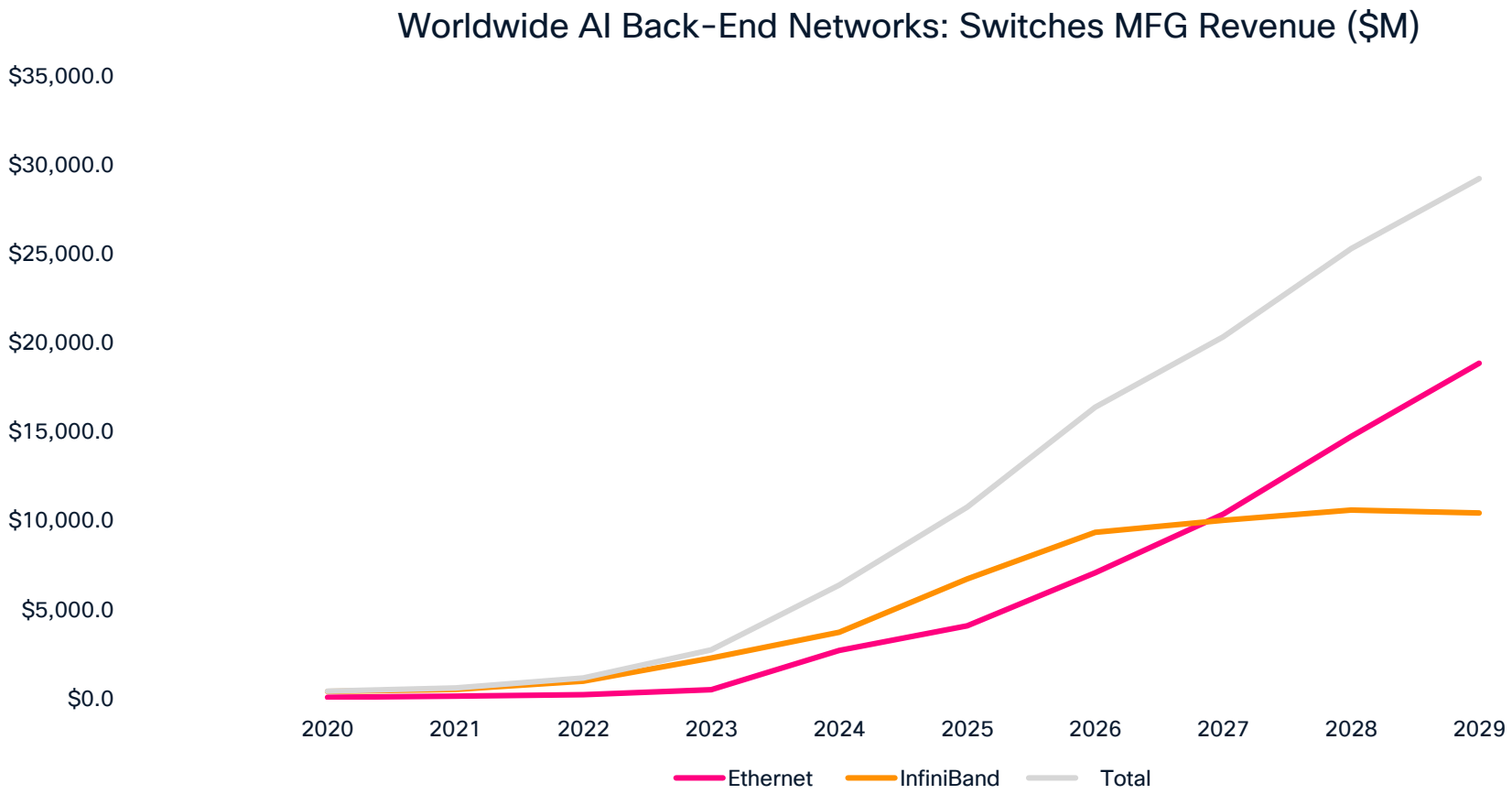
Let's ask ourselves some questions about InfiniBand

AI back-end network switch port speed



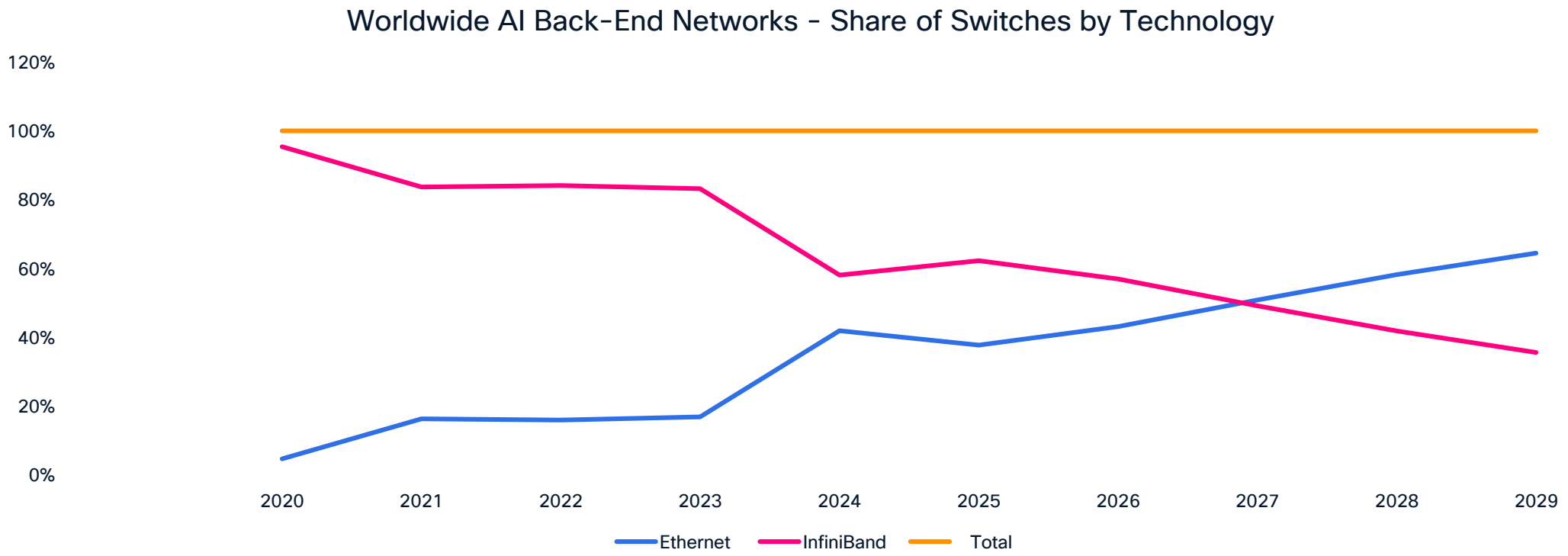
Source: Dell'Oro Group ADVANCED RESEARCH REPORT AI NETWORKS FOR AI WORKLOADS MARKET FORECAST (January 2025)

AI back-end network switch TAM



Source: Dell'Oro Group ADVANCED RESEARCH REPORT AI NETWORKS FOR AI WORKLOADS MARKET FORECAST (January 2025)

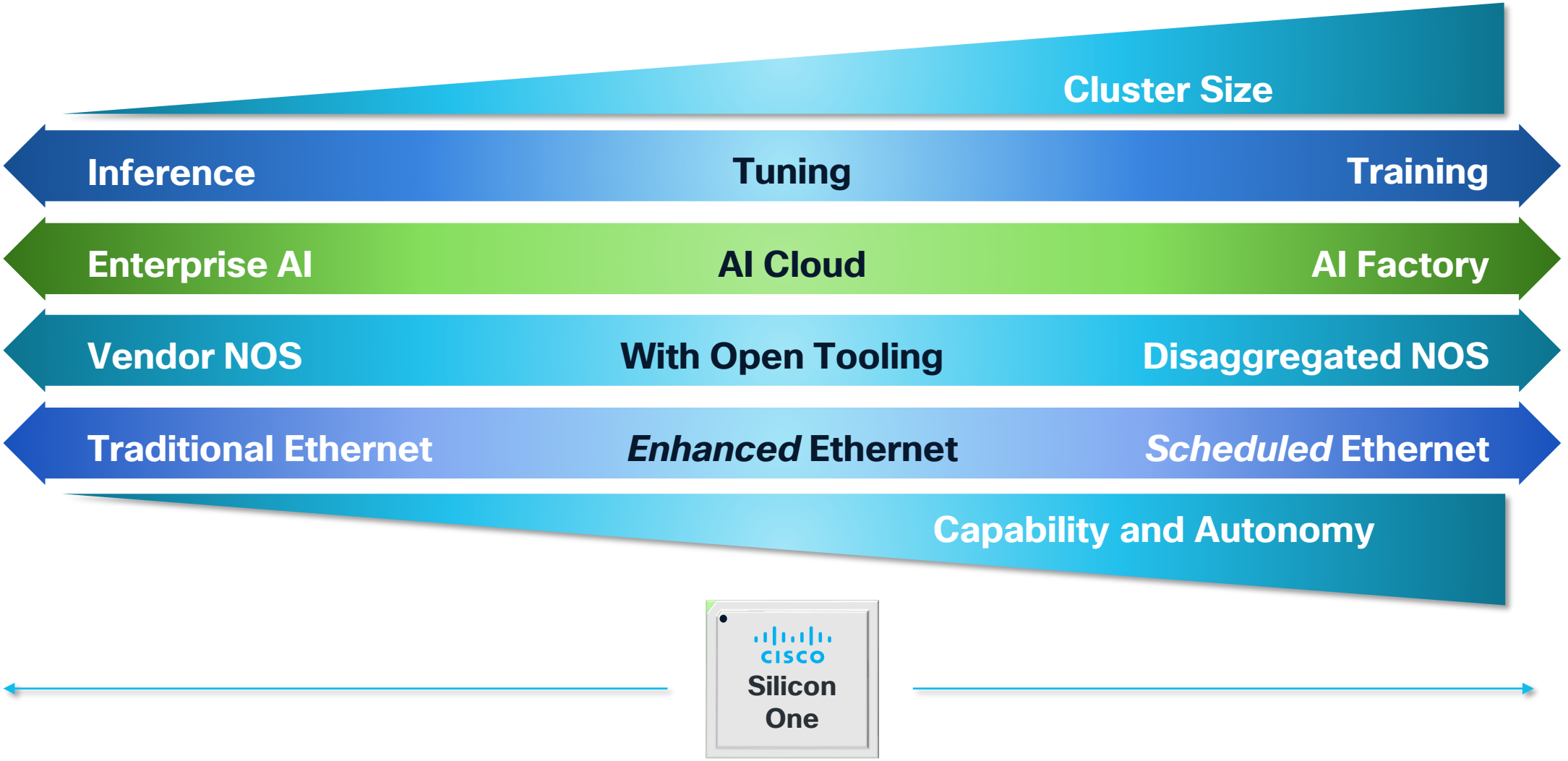
AI back-end network switch market share



Source: Dell'Oro Group ADVANCED RESEARCH REPORT AI NETWORKS FOR AI WORKLOADS MARKET FORECAST (January 2025)

Where does Ethernet fit across AI Landscape?

An Oversimplified Clustering of AI Technology Requirements



AI Workflows and the Training Network Bottleneck

AI Models & Training

LLMs are orders of magnitude more intensive than DLRM

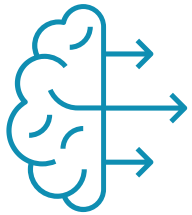
Deep Learning Recommendation Models

Search, Feed ranking. Ads & content recommendation

Inference needs a few Gigaflops for 100ms TTFT

Narrower scope, domain specific

Training: ~100 Gigaflop/ sentence



Large Language Models

Intricacies of human language

Inference needs 10s of Petaflops for 1 sec TTFT

Generate intelligent, creative responses

Training : ~1 Petaflop/ sentence

An Improved user experience means *a faster time to first token*

AI Workloads & Use Cases

Generative AI

Trained AI model enables users to quickly generate new content based on a variety of inputs

Identify the patterns and structures within existing data to generate new and original content.

Inference

Trained AI model makes predictions or conclusions based on new, unseen data.

AI inference is crucial for putting trained models into action and enabling them to perform useful tasks in various applications.

Agentic AI

Autonomously plan, reason, and take actions.

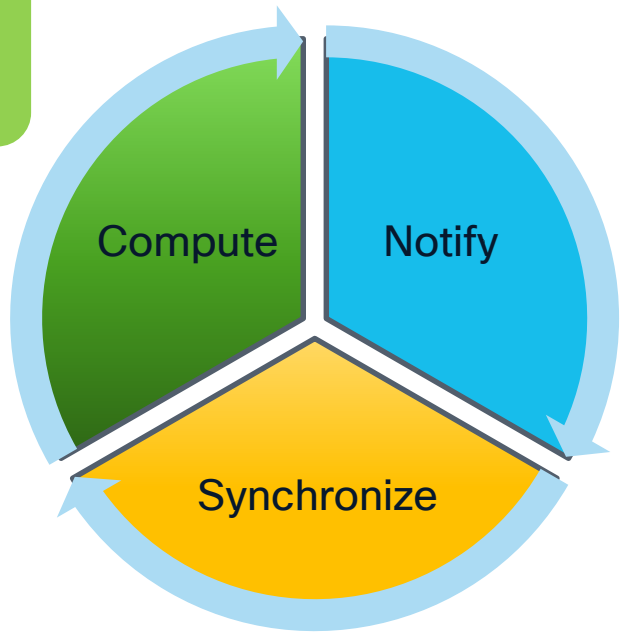
Interacting with tools and working across systems without human prompts

An improved user experience requires AI workload *performance that is both dependable and flexible*

The AI/ML Workload Cycle

GPU Execute Instructions

High Bandwidth capable GPUs can saturate network links



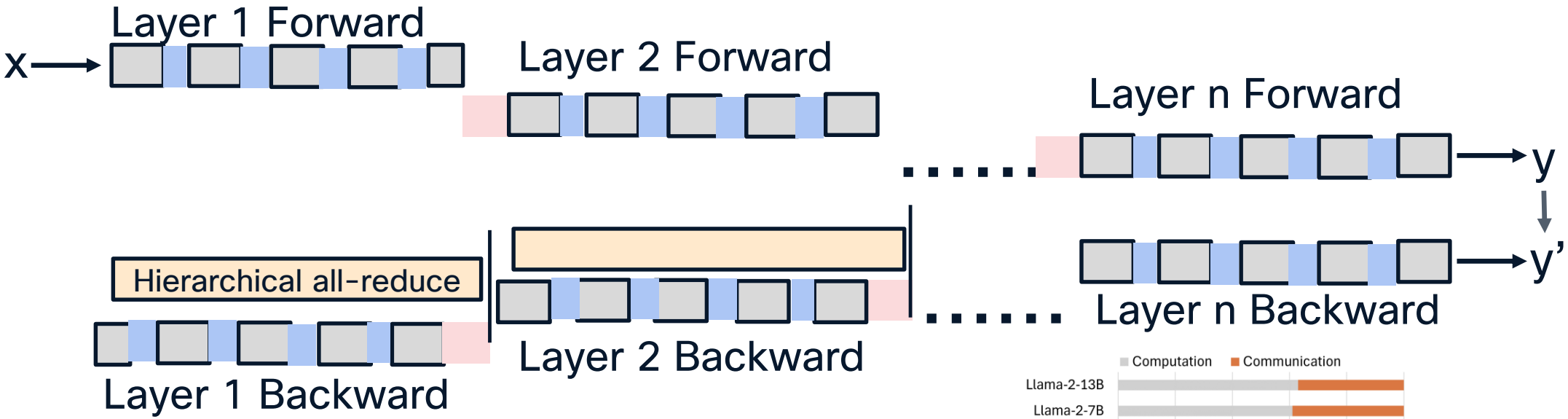
Send results of computation

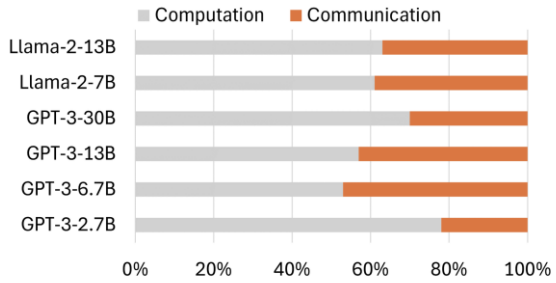
Different collective communication patterns
All Reduce (Aggregate/reduce everyone's data and send to everyone

Wait for all GPUs to complete

Synchronizes all GPUs
Compute stalls, waiting for the slowest path
Job Completion Time (JCT) influenced by the worst-case tail latency

Training Computation/Communication





Source: https://arxiv.org/html/2409.15241v1

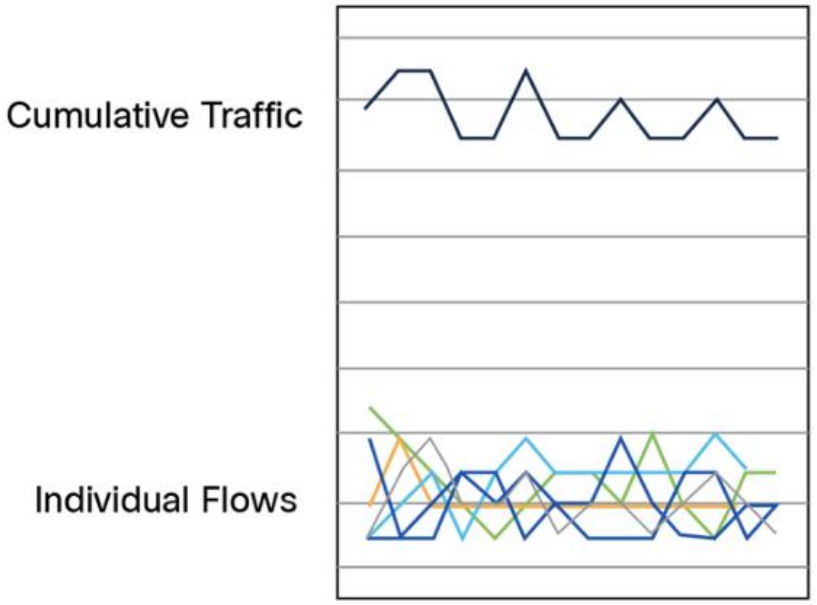
Bottomline: We need to keep communication latency low

Compute | Scale-up | Scale-out | Scale-up + Scale-out + intra-DC

Your AI/ML Training is only as fast as the slowest GPU

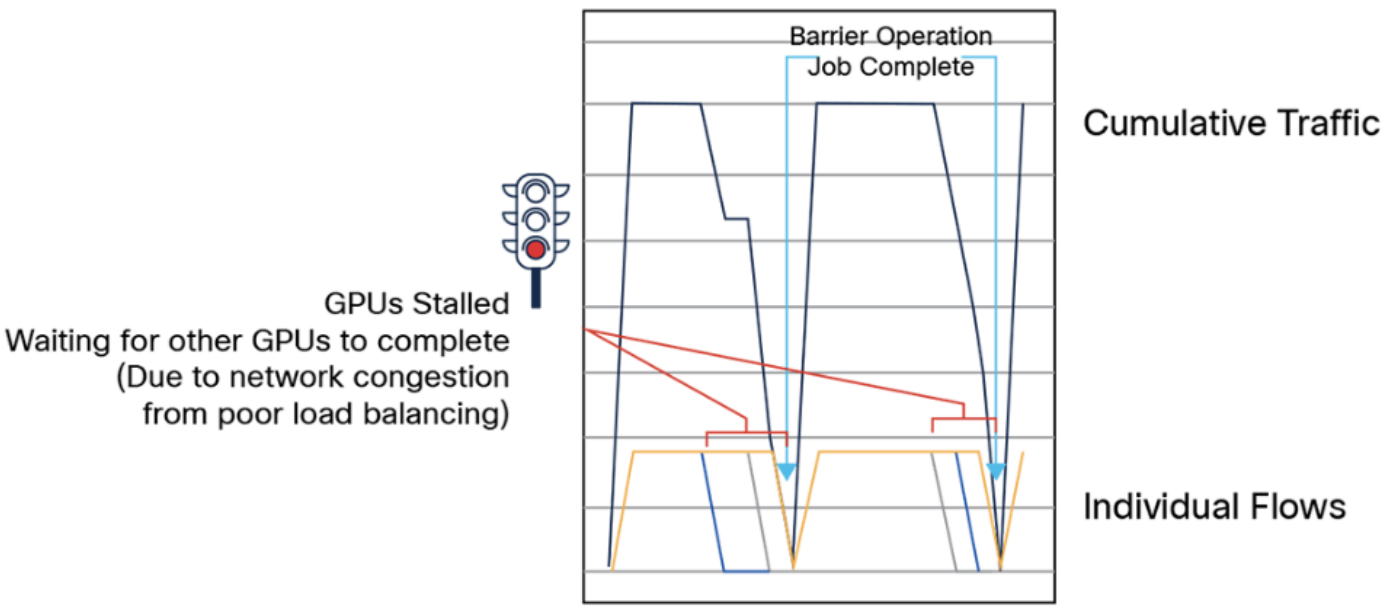
The network can become the bottleneck

Traditional DC Traffic Pattern



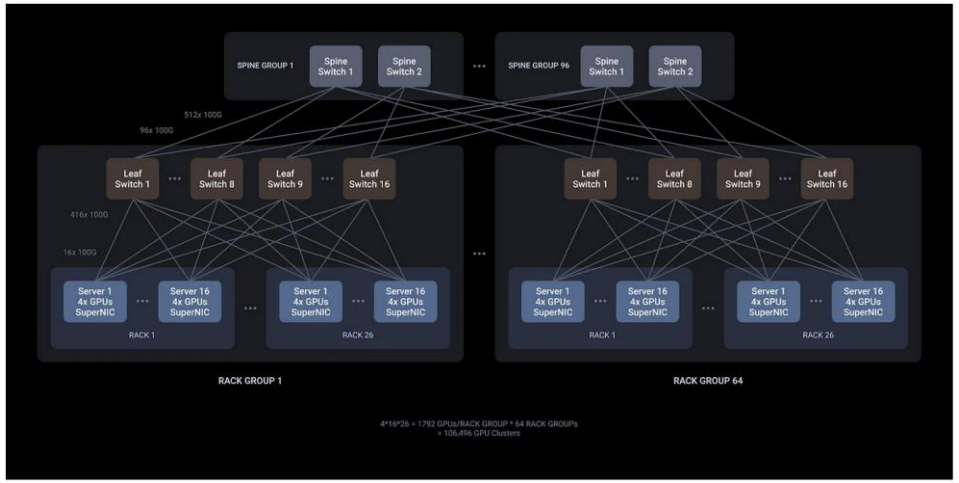
Many asynchronous small BW flows
Chaotic pattern averages out
to consistent load

AI (All-to-all Collective) Traffic Pattern

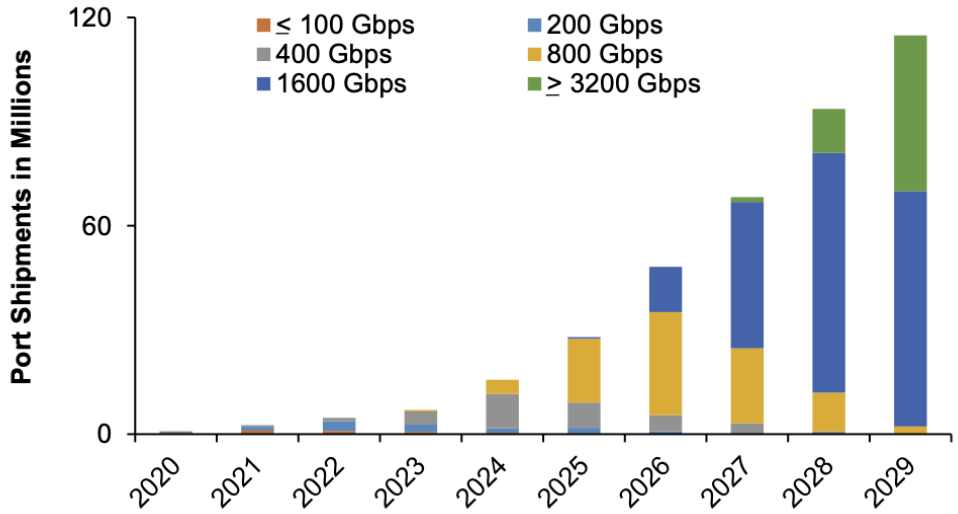


Few synchronous high BW flows
Synchronization magnifies long tail
latency and bad load balancing decisions

Critical Requirements for AI Back-end Networks



100K+ GPUs are reachable using 2 network tiers or less



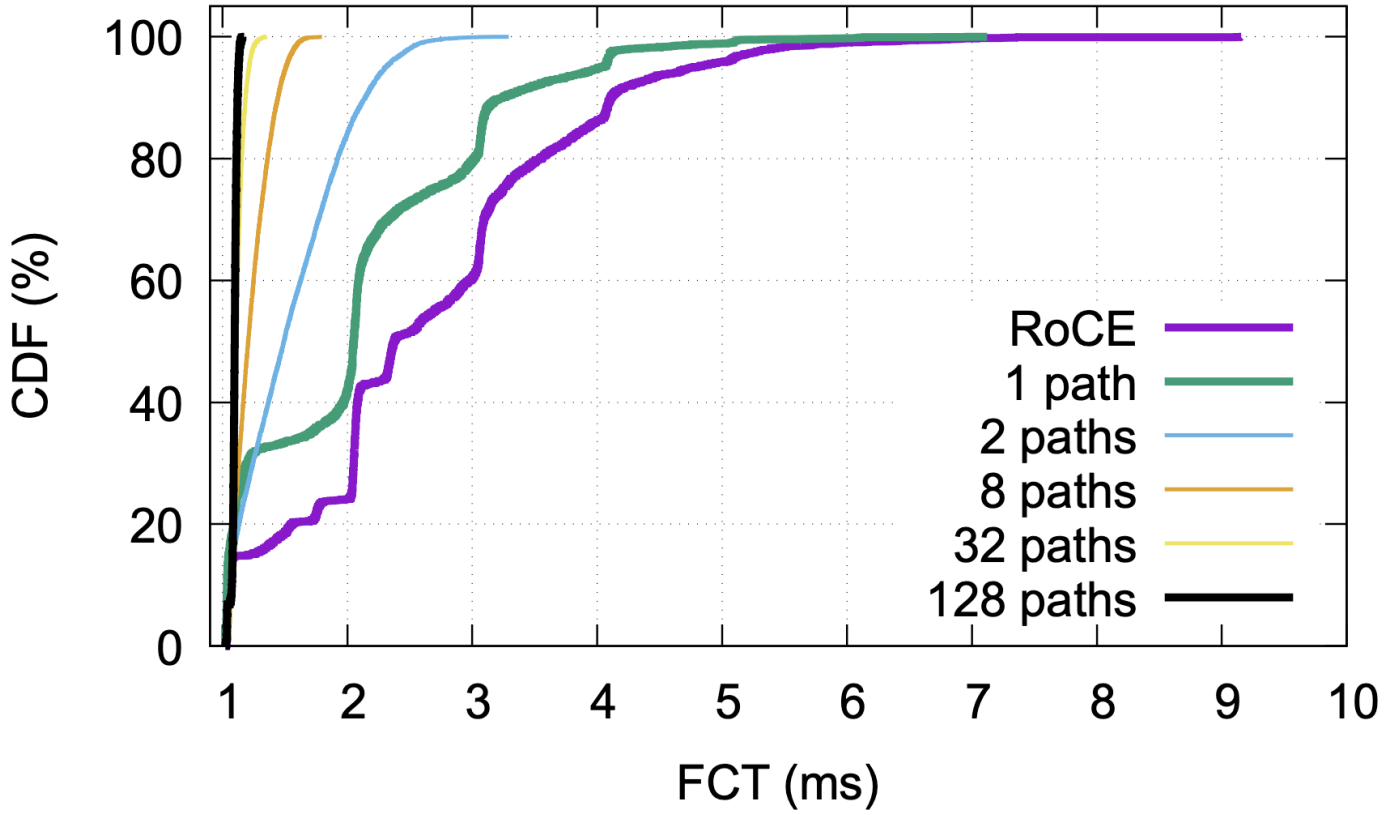
Scalable

High Speed

Tail Latency

Source: Dell'Oro Group ADVANCED RESEARCH REPORT AI NETWORKS FOR AI WORKLOADS MARKET FORECAST (January 2025)

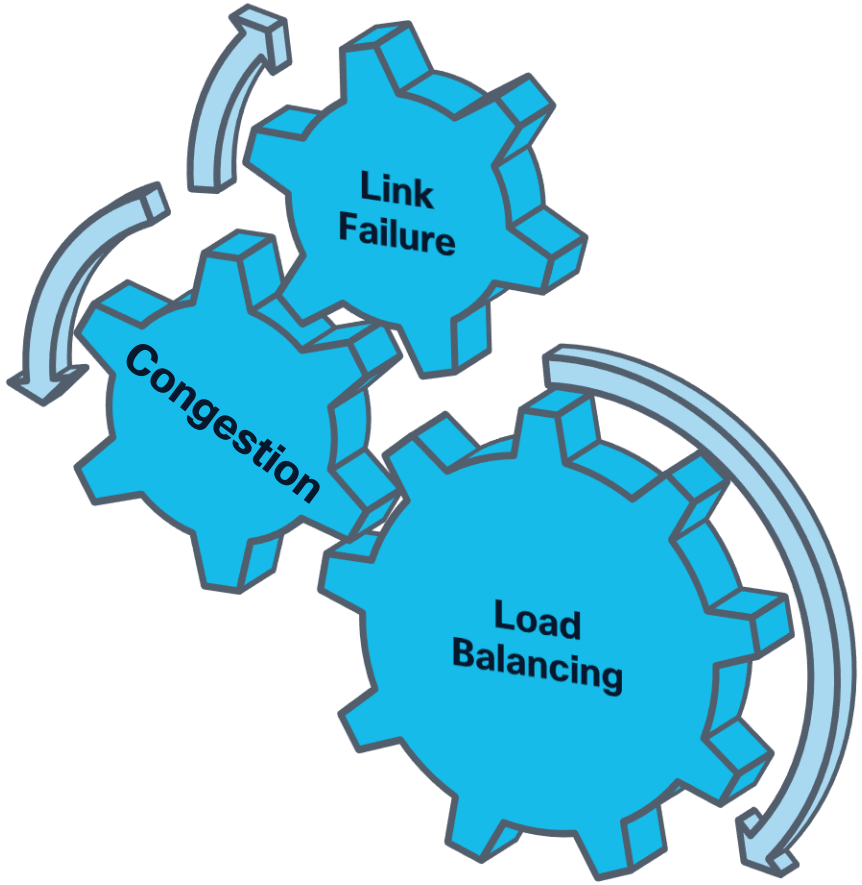
Multi-pathing for AI Backend Networks



Multipath is key to reduce tail latency and JCT

Source: STrack: A Reliable Multipath Transport for AI/ML Clusters

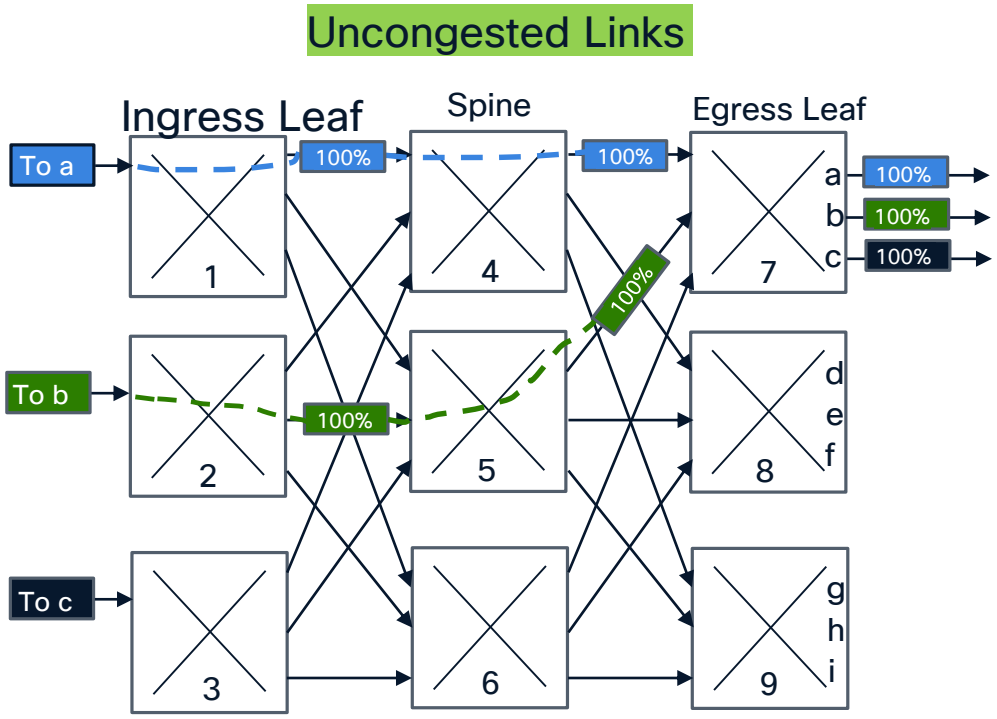
Minimizing Job Completion Time is *the AI Challenge*

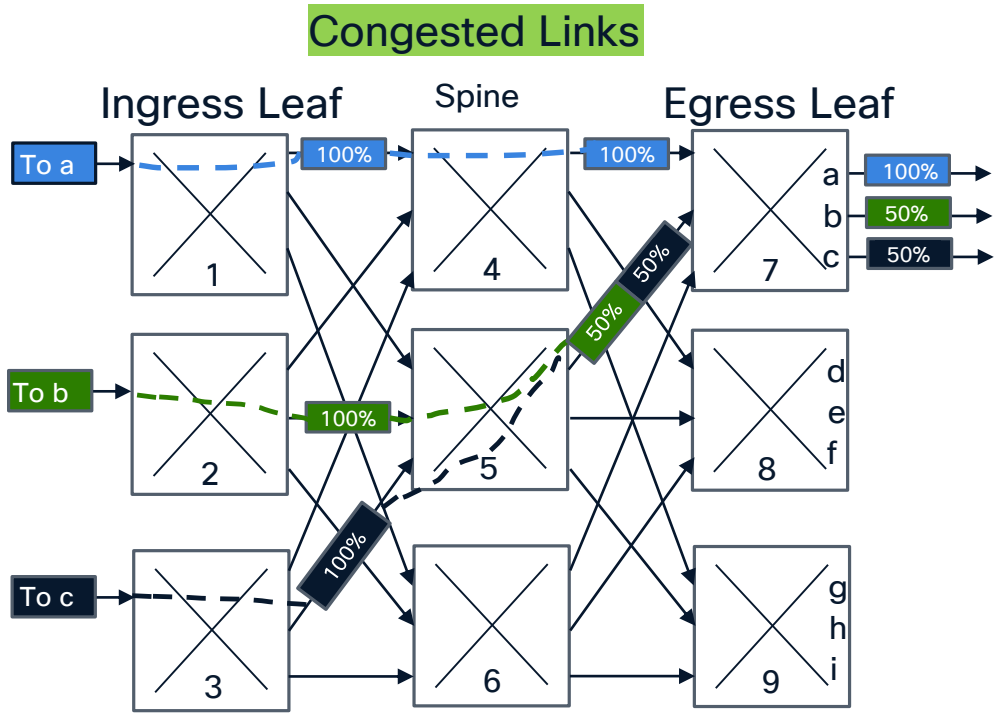


Wrenches in the works

- Underutilized fabric links
- Head of Line blocking
- Incast Congestion
- Link failures and black holing

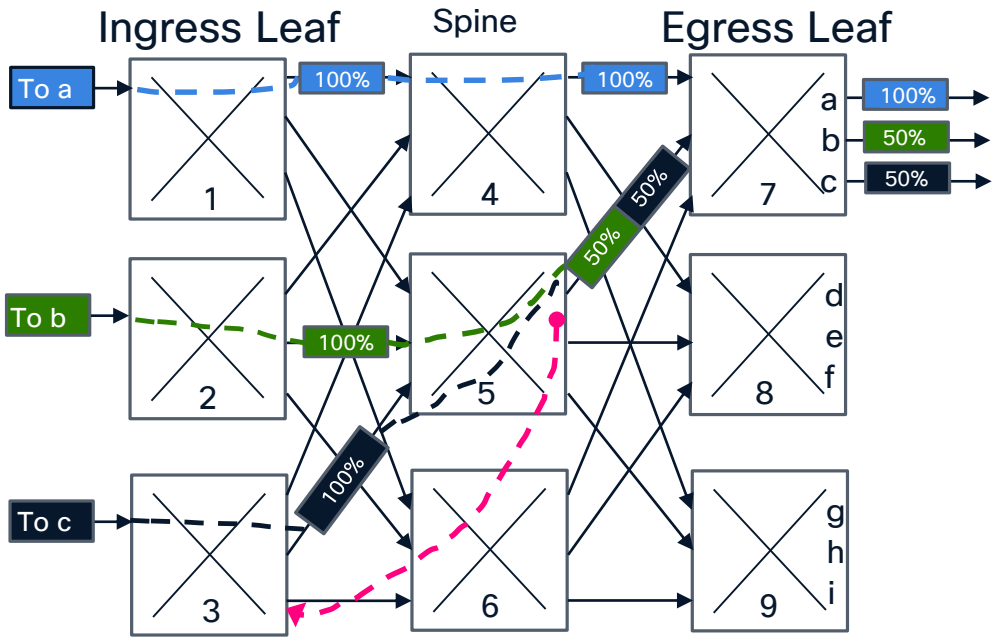
Load Balancing Basics – ECMP



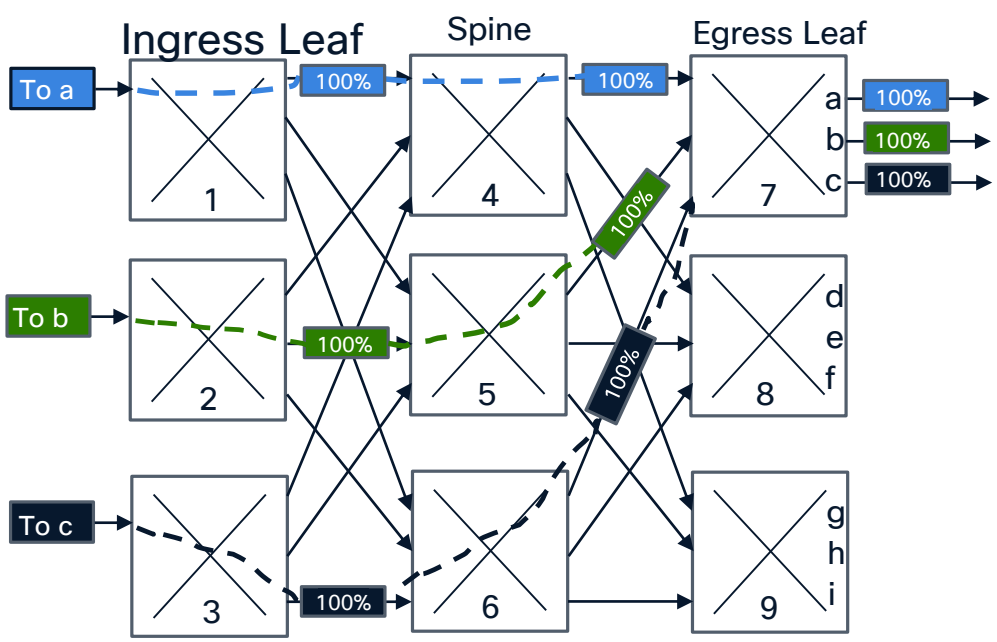


Load Balancing Basics – Enhanced Ethernet

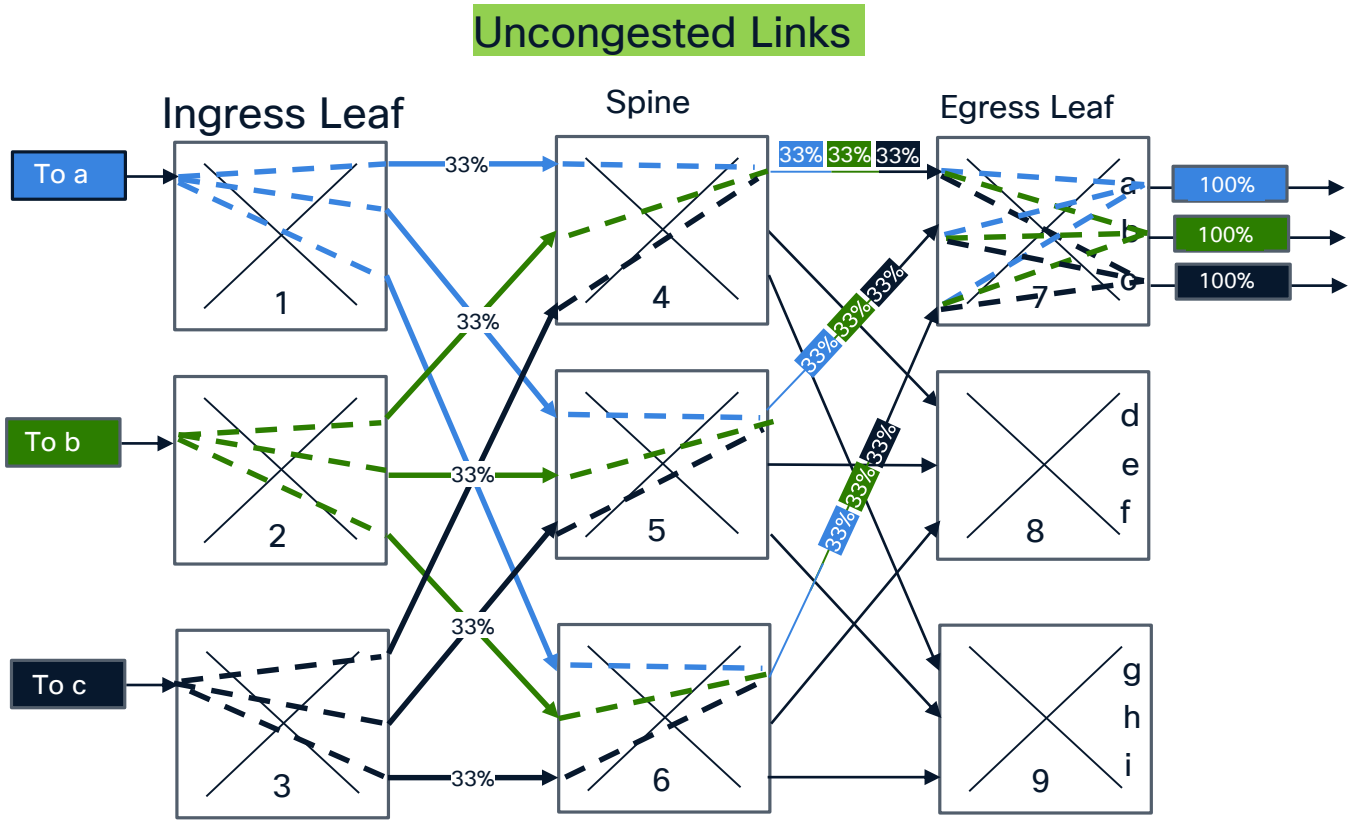
Congested Links



Uncongested Links



Load Balancing Basics – Scheduled Ethernet



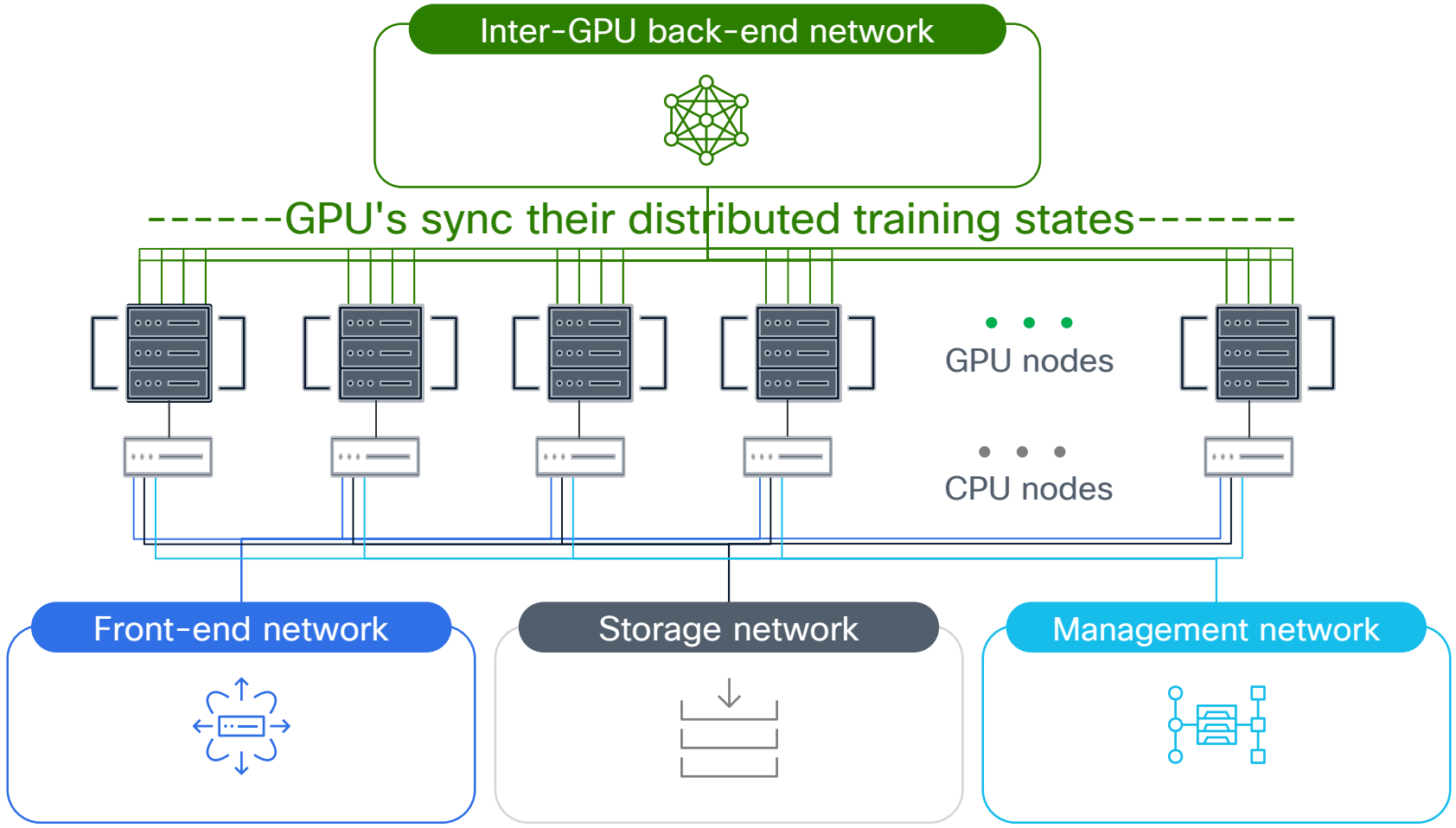
AI Ethernet Fabric Options

	1 Ethernet	2 *Enhanced* Ethernet		3 *Scheduled* Ethernet
Load Balance	Stateless ECMP	Stateful Flow/Flowlet	Spray & Re-order in SmartNIC	Spray & Re-order in leaf
Fabric Congestion Management	Congestion Reaction with ECN/PFC	Congestion Reaction with congestion score to adjust distribution		Congestion Avoidance
Link Failure	Software	Hardware		
Job Completion Time	Good	Better		*Best*
Coupling between NIC and Fabric	No		Yes	No
Place in Network	Frontend & Backend, Training & Inference			
Fabric Infrastructure	Leaf/Spine or Modular Chassis			Modular Chassis

Todays AI Infrastructure Options

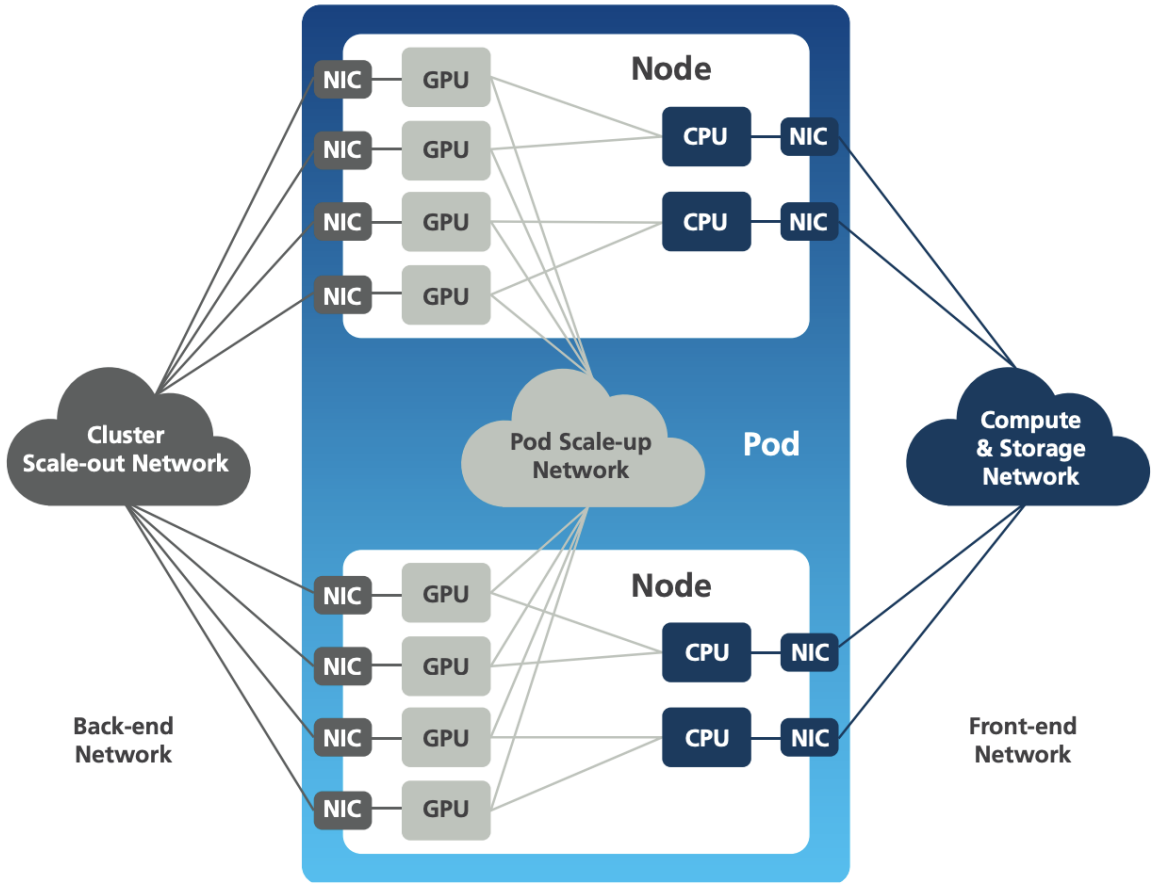
AI Network Type Fundamentals

Multiple Networks for AI infrastructure



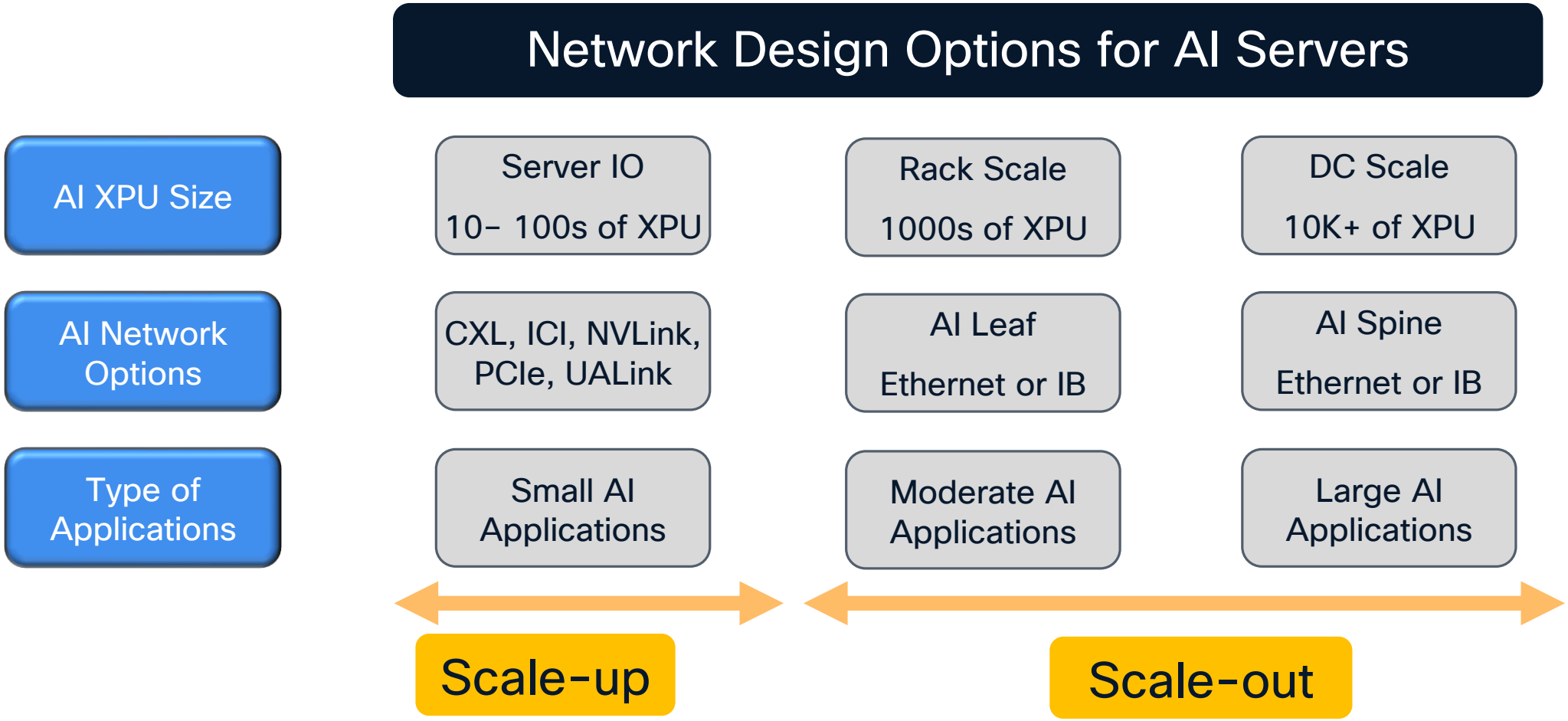
AI Infrastructure Scaling

- Scaling AI involves two primary approaches:
  - Scaling up
  - Scaling out.
- Goal:
  - Optimize performance
  - Manage resources efficiently
  - Meet the growing computational demands of AI applications



Scale-up (Vertical) vs. Scale-out (Horizontal)

Strategies to increase capacity and performance



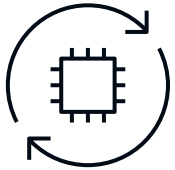
*XPU: could be GPU, TPU, or any other type of Accelerator

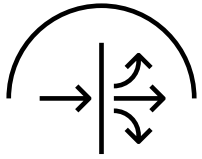
Addressing Trends – Ultra Ethernet

RoCEv2 as a Scale-out Transport Protocol

Good, but can be improved upon

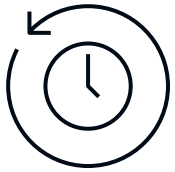
PFC Requires Intense Buffering





Victim flows, congestion trees, PFC storms and deadlocks

Go-back-N retransmission



Congestion control relies on pause and timeout

ULTRA ETHERNET VISION

Deliver an Ethernet based open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI & HPC at scale

THE NEW ERA NEEDS A NEW NETWORK

Ultra Ethernet

As **performant** as a supercomputing interconnect

As **ubiquitous** and **cost-effective** as Ethernet

As **scalable** as a cloud data center

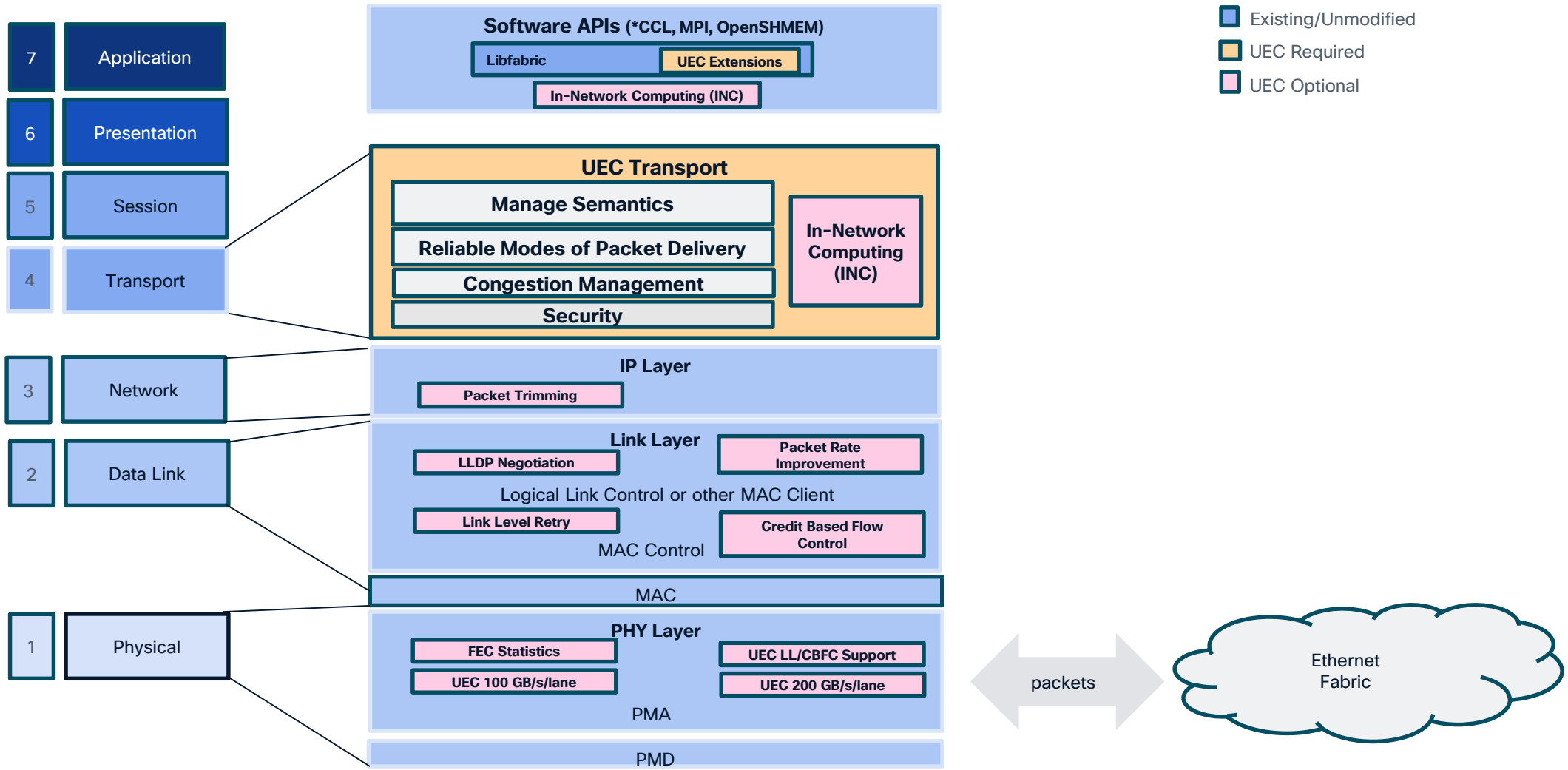
UEC Seeks to Bring Open Standards to AI Networks

Open specifications, APIs, source code for optimal performance of AI and HPC

AMD ARISTA BROADCOM CISCO EVIDEN an atos business enfabrica

Hewlett Packard Enterprise intel Meta Microsoft ORACLE MARVELL Alibaba Cloud

UE – Overview



UE Physical Layer

Signaling

Subset of Ethernet PHYs 100G per lane signaling, defined in

IEEE Std 802.3™-2022
IEEE Std 802.3db™-2022
IEEE Std 802.3ck™-2022

Media

Backplane (KR, clause 163)
Copper cable (CR, clause 162)
MMF up to 50 m (VR, clause 167)
MMF up to 100 m (SR, clause 167)
Parallel SMF up to 500 m (DR, clauses 124, 140)
WDM SMF up to 2 km (FR, clauses 140, 151)

Optional Support

Control Ordered Set (CtlOS)

Message mechanism utilized by the UE Link Layer features

Link-Level Retry (LLR)

Credit-Based Flow Control (CBFC)

FEC Statistics for Link Quality Prediction

UE Link Layer

Optional Enhancements at Link Layer

Link Layer Reliability (LLR)

Extra reliability for links

Link level ACKs and retries

Credit Based Flow Control (CBFC)

Lossless Link

Replace Priority Flow Control (PFC)

Packet rate improvements

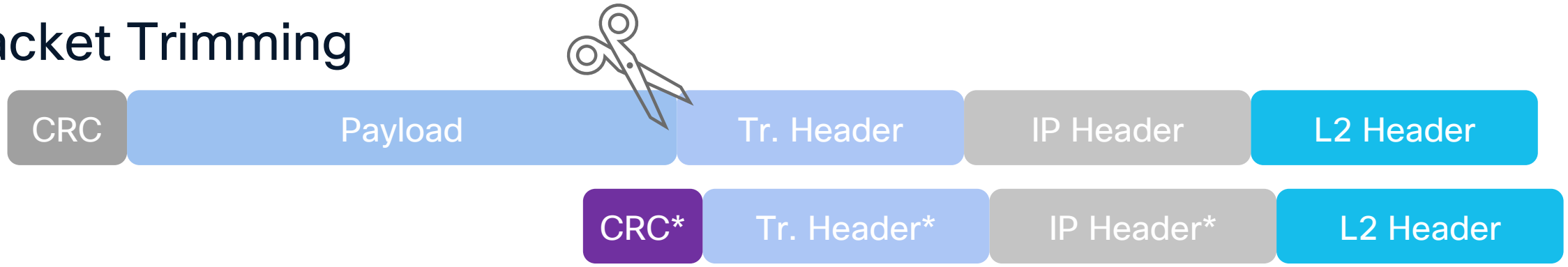
Extra performance/minimal overhead

Reduced IPG, L2/L3 header compression

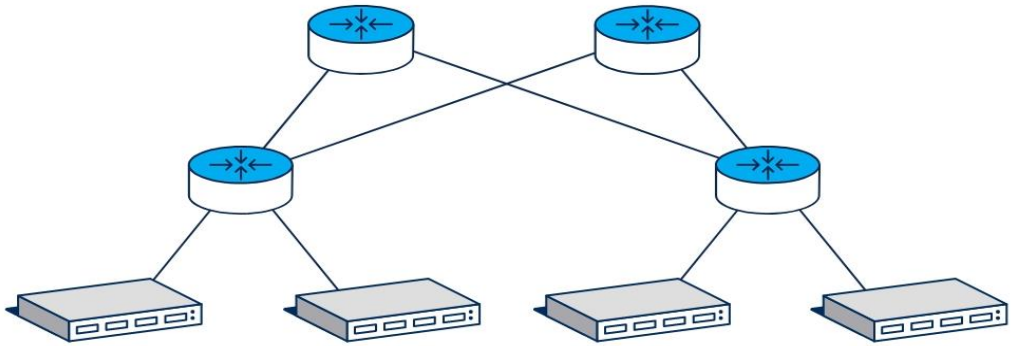
Capability Negotiation - LLDP

UE – IP Layer

Packet Trimming

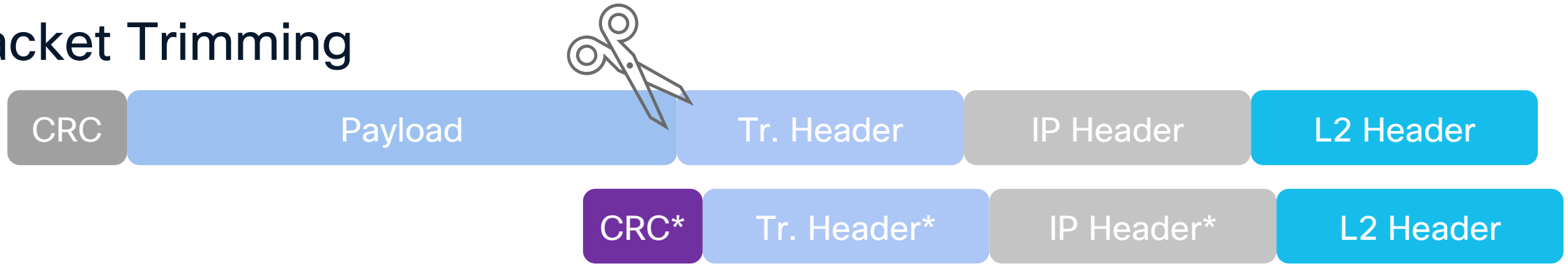


*Modified/updated



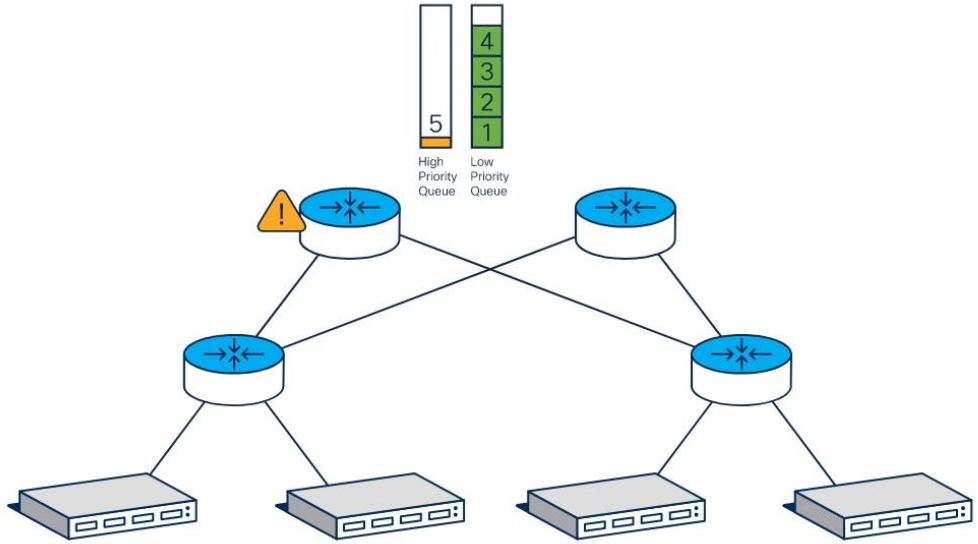
UE – IP Layer

Packet Trimming

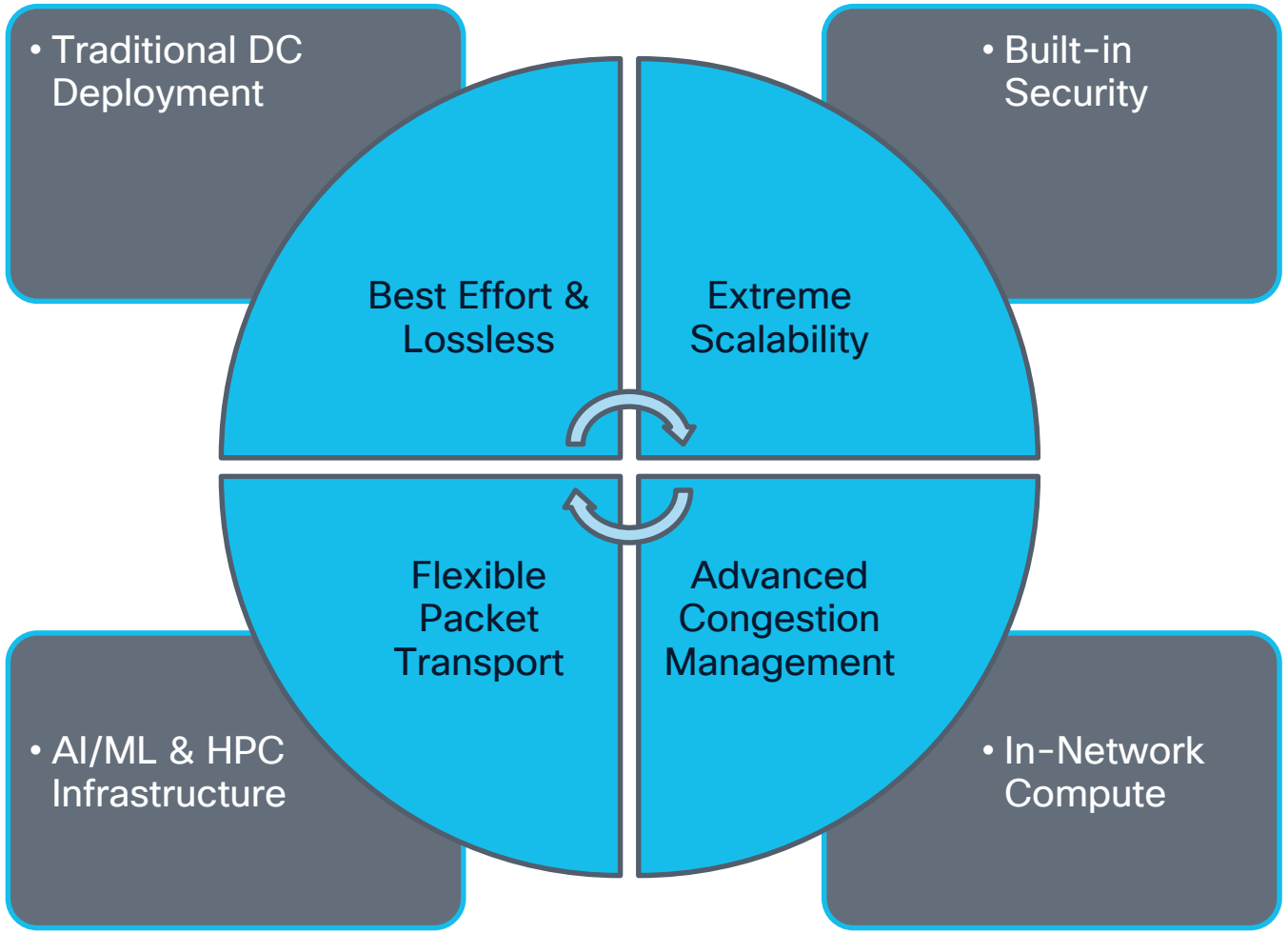


*Modified/updated

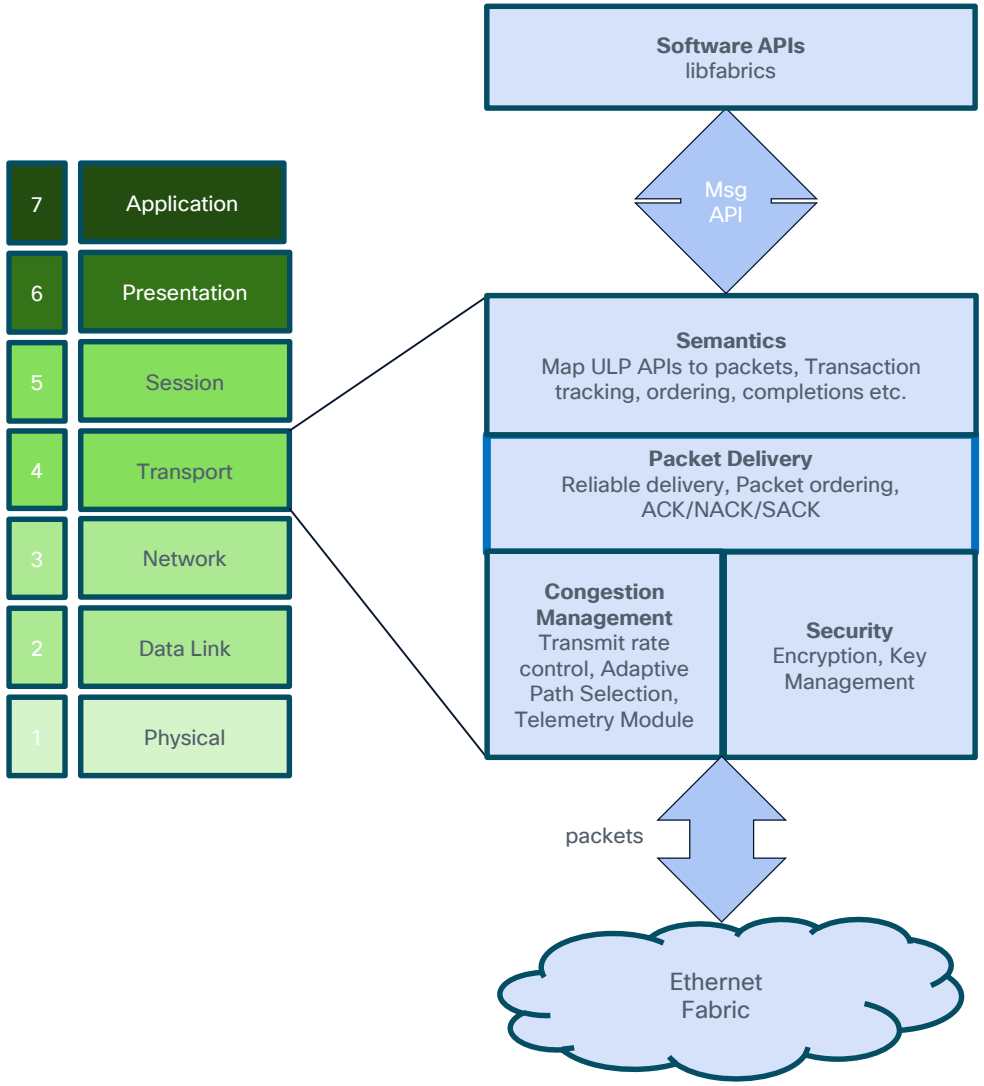
- Enables 1RTT loss detection & trigger fast retransmission.
- Only packets belonging to the TRIMMABLE category can be trimmed.



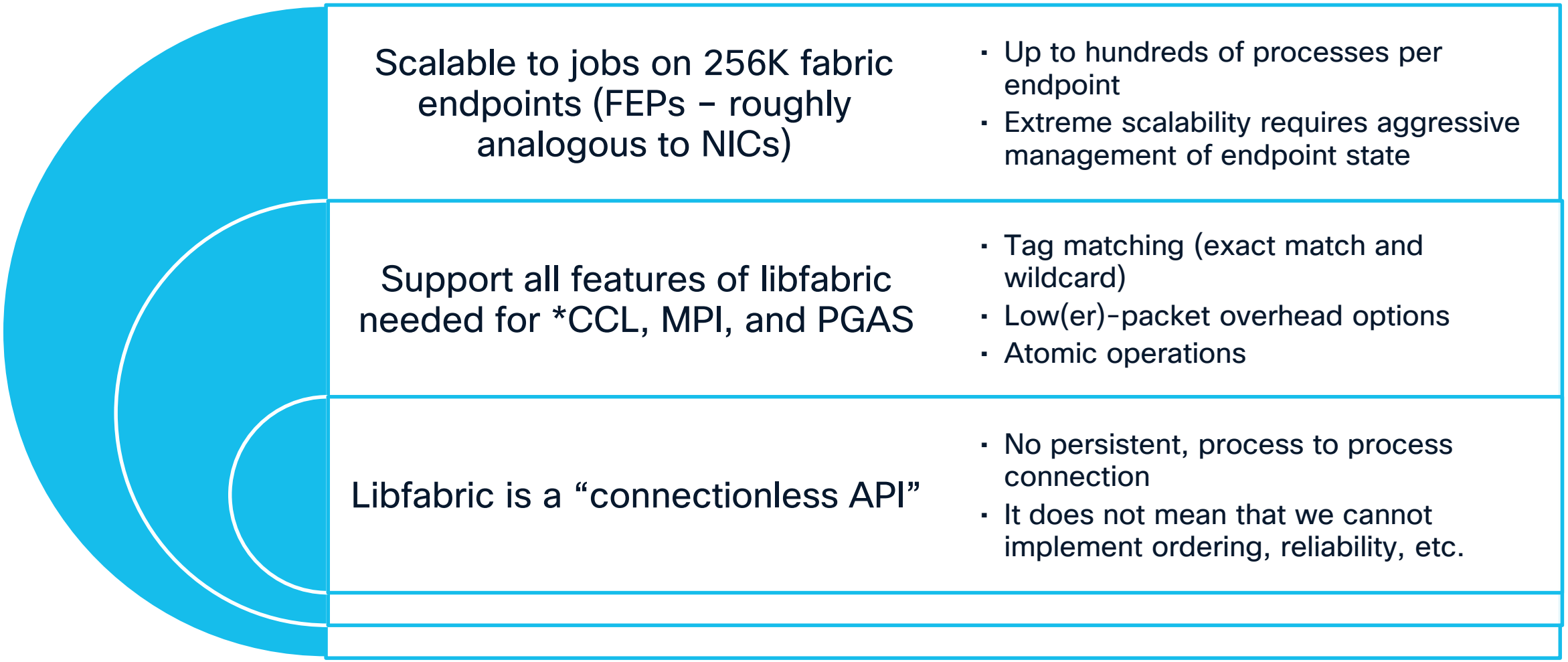
UE – Transport Key Features + Design Goals



UE – Transport Layer



UET – Semantics – Objectives

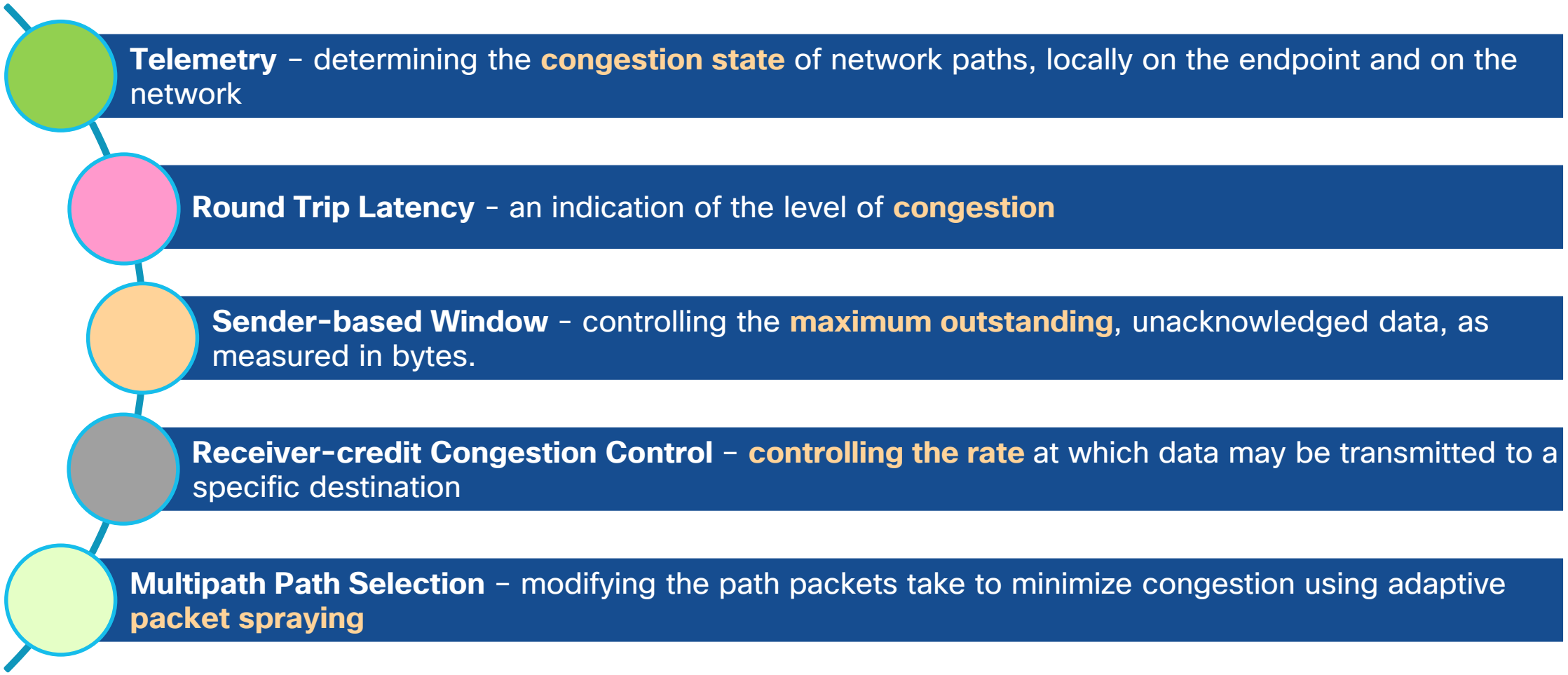


UET – Packet Delivery – Modes

- Reliability and Ordering modes
- Selective acknowledgement (SACK)
  - Loss detection and retransmission
- Best effort and lossless networks

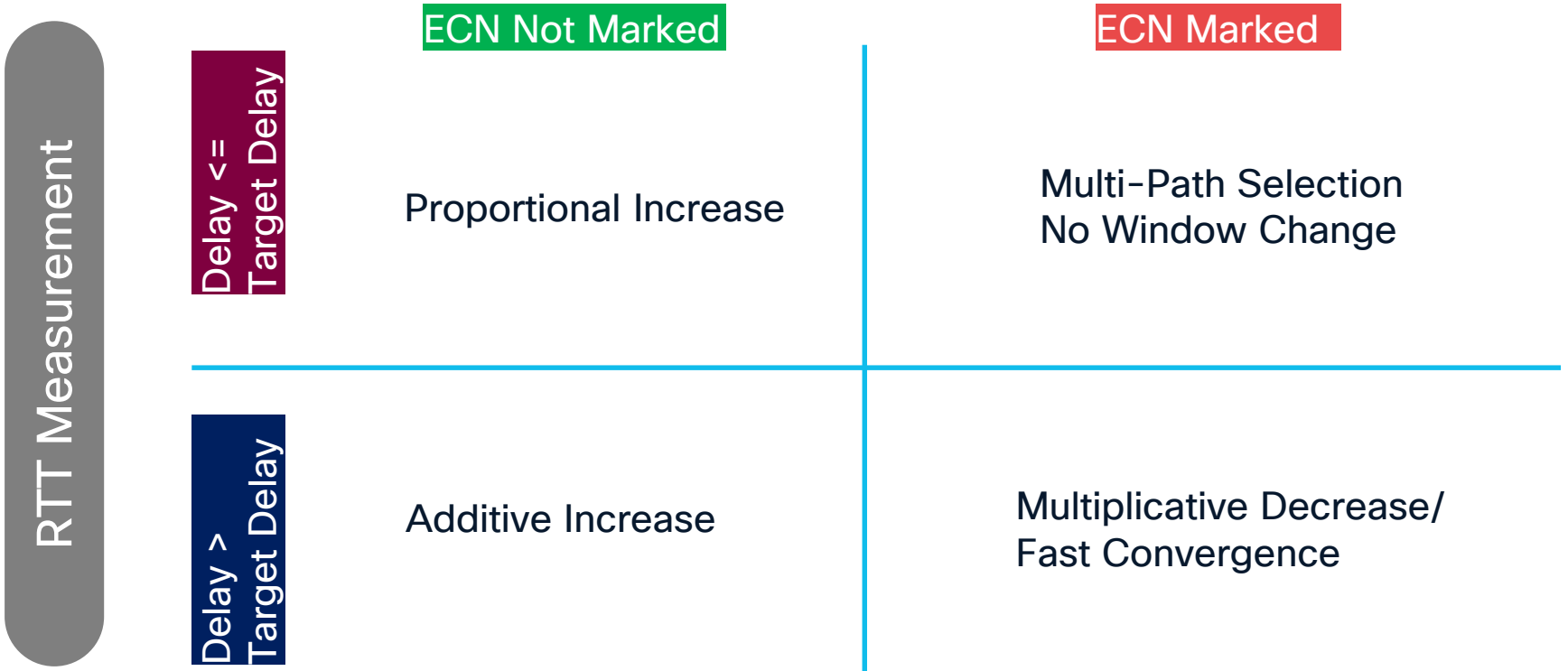
PDS Mode	Overview Description
RUD – Reliable, Unordered	• Establish dynamic, ephemeral connections (PDC) • Packet Sequence Numbers (PSN) increment by one, Multipath, SACK
ROD – Reliable, Ordered	• Establish dynamic, ephemeral connections (PDC), PSN increment by one • Single path with GoBackN – OR – Multipath with re-order buffer & SACK
RUDI – RUD for Idempotent	• Optimized for Idempotent operations – RMA Write & Read • Unique packet IDs, multipath, ACK per packet, Selective retransmit
UUD – Unreliable, Unordered	• Generic service

UET – Congestion Management Tools



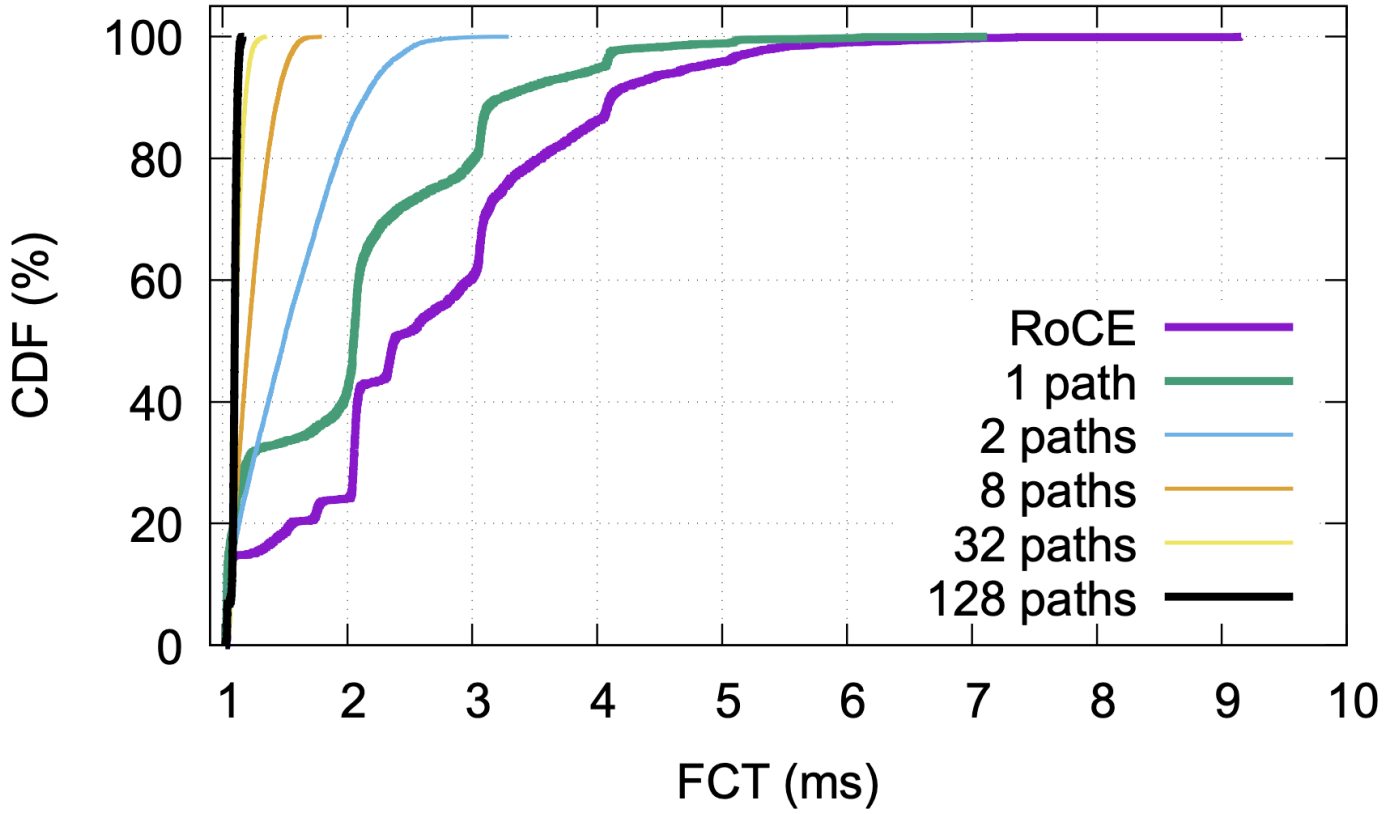
UET – Congestion Management

Transmit Window Control



Window adjustment is a quantitative decision

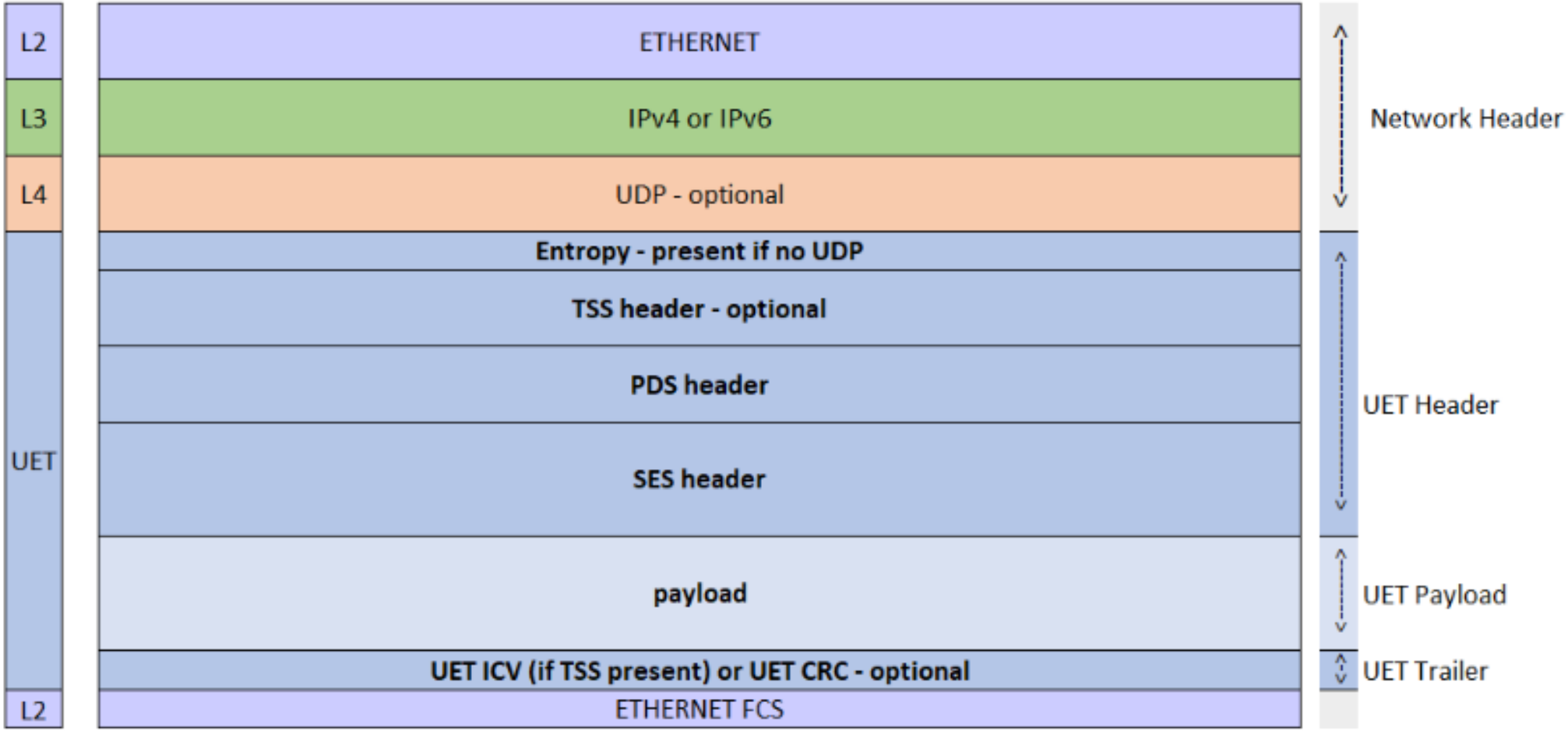
Multi-pathing for AI Backend Networks



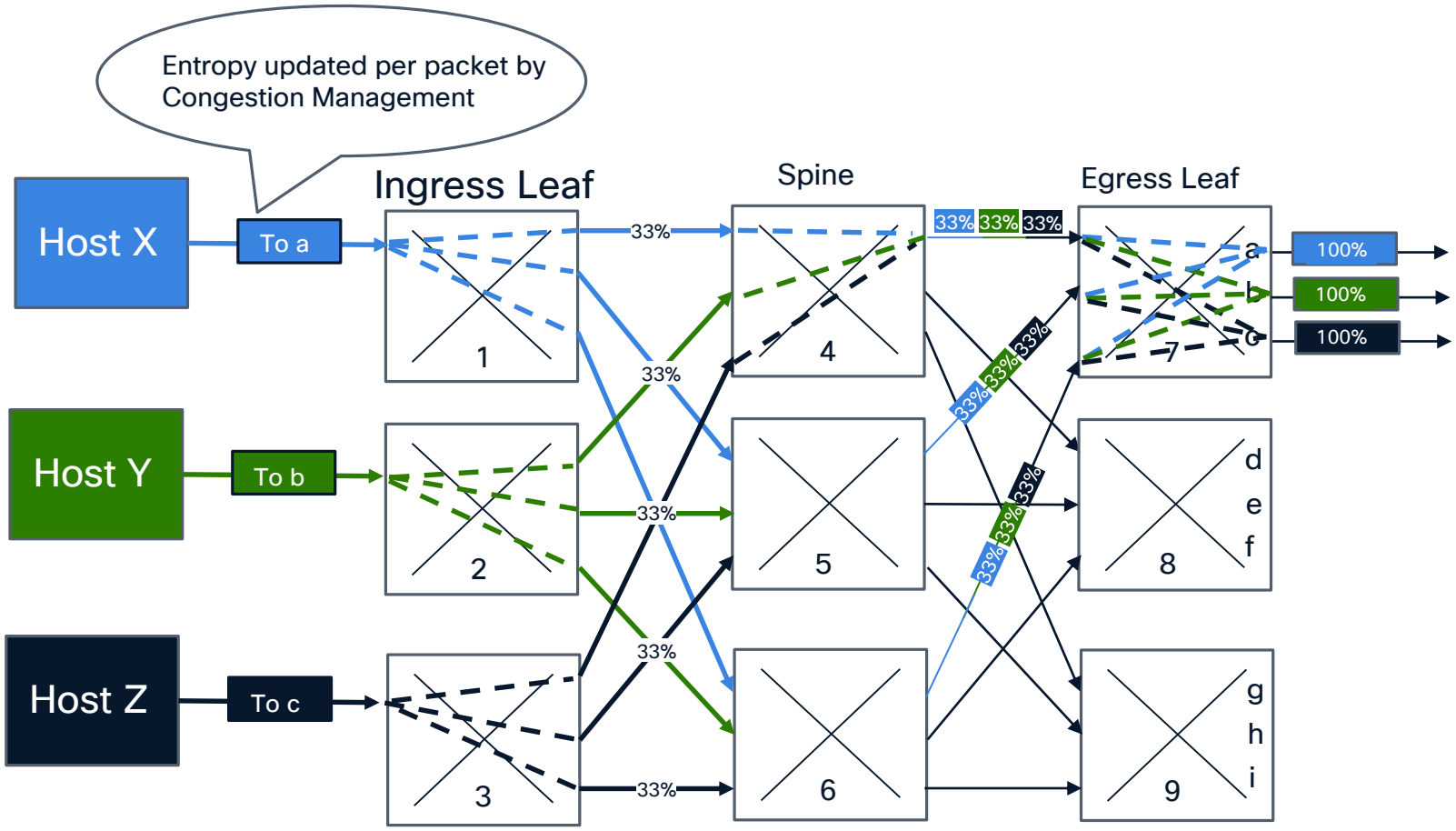
Multipath is key to reduce tail latency and FCT

Source: STrack: A Reliable Multipath Transport for AI/ML Clusters

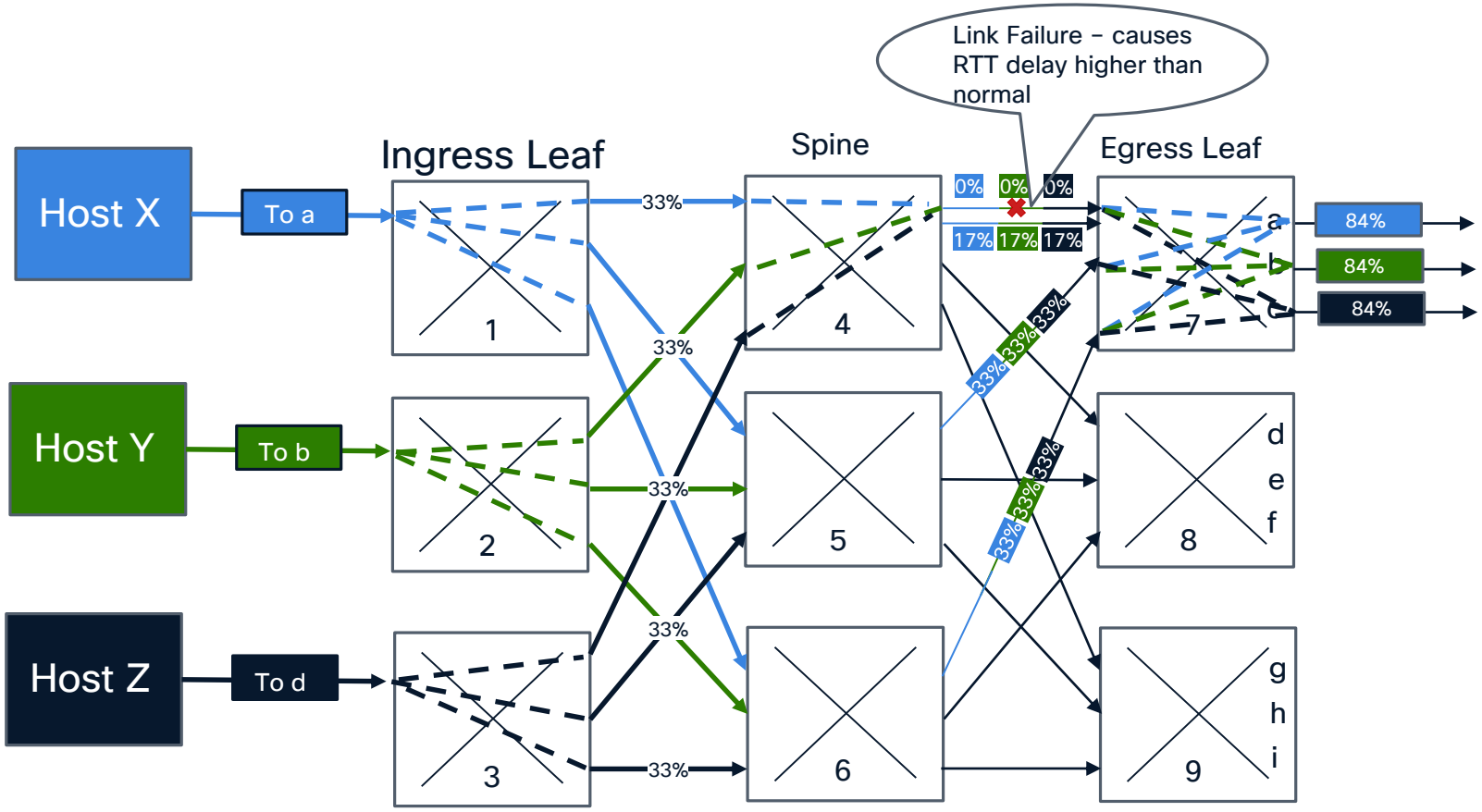
UET Packet Structure



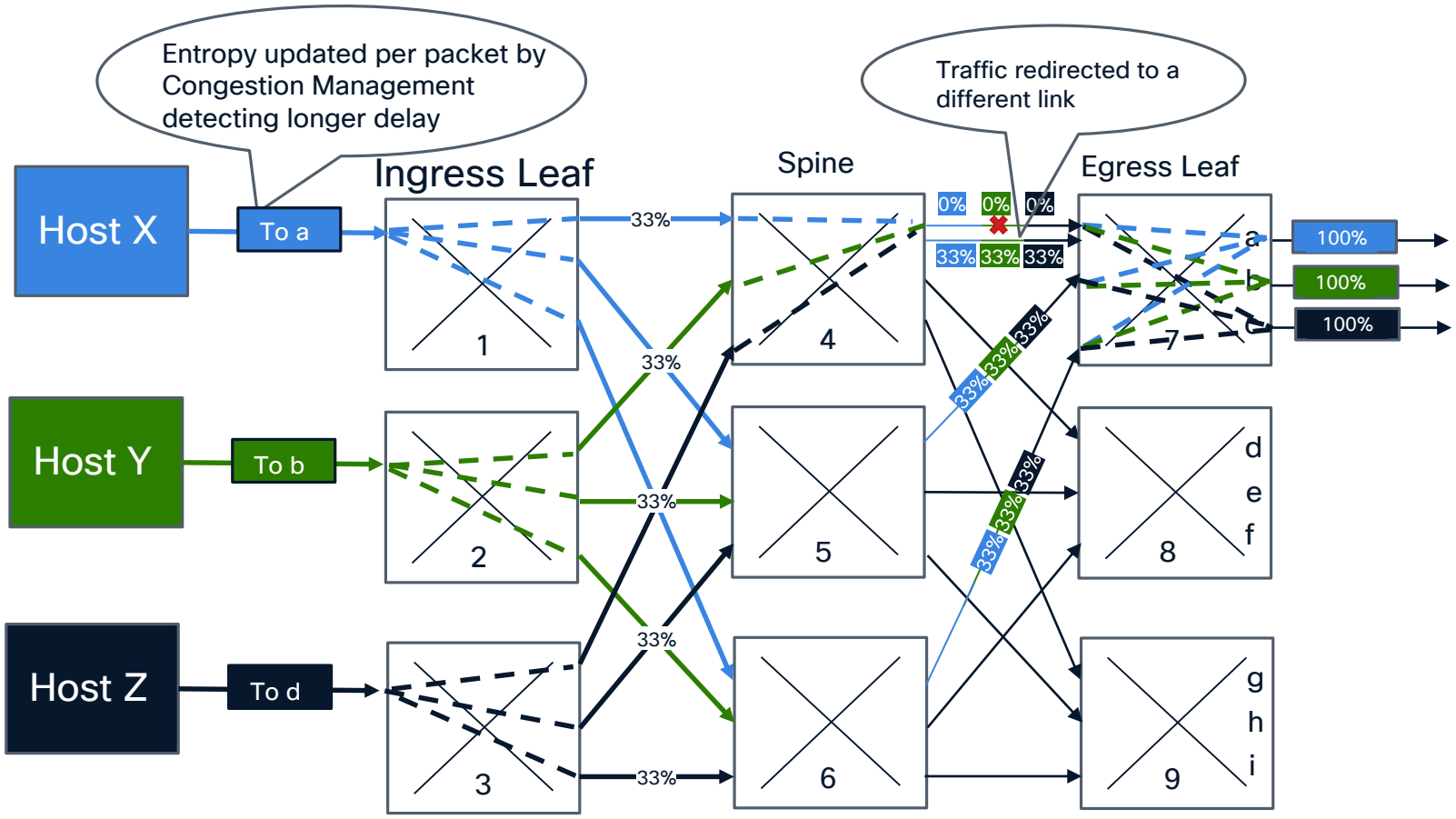
UE –Multipath Selection (Packet Spray)



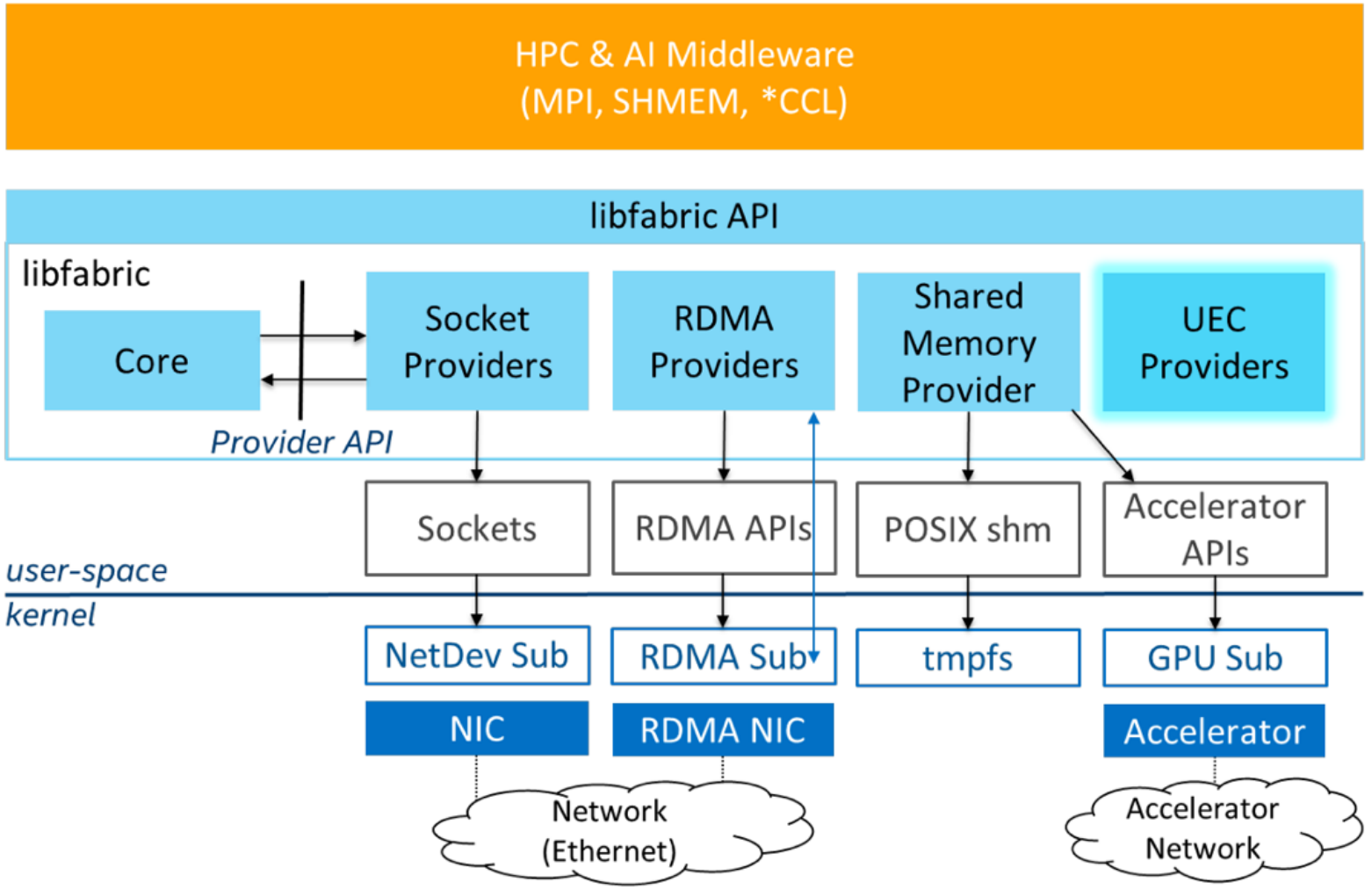
UE – Multipath Selection – Network Aware



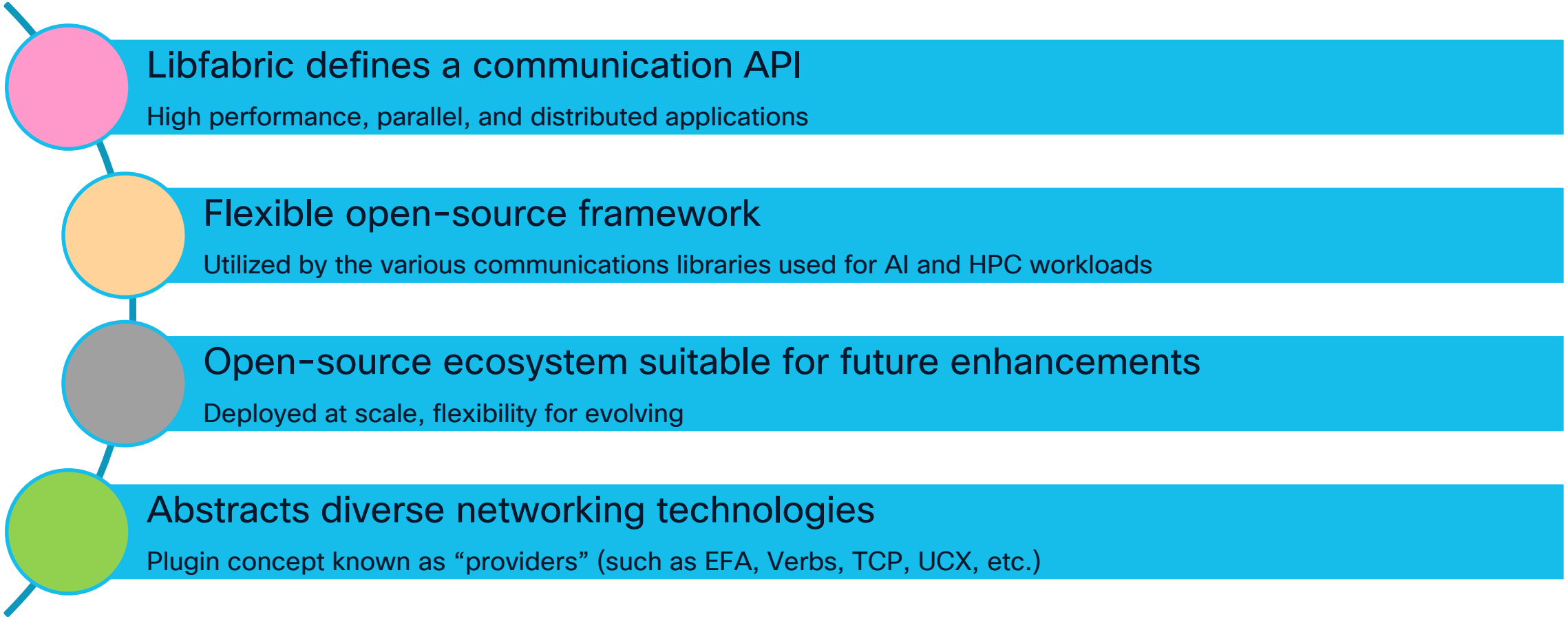
UE – Multipath Selection – Network Aware



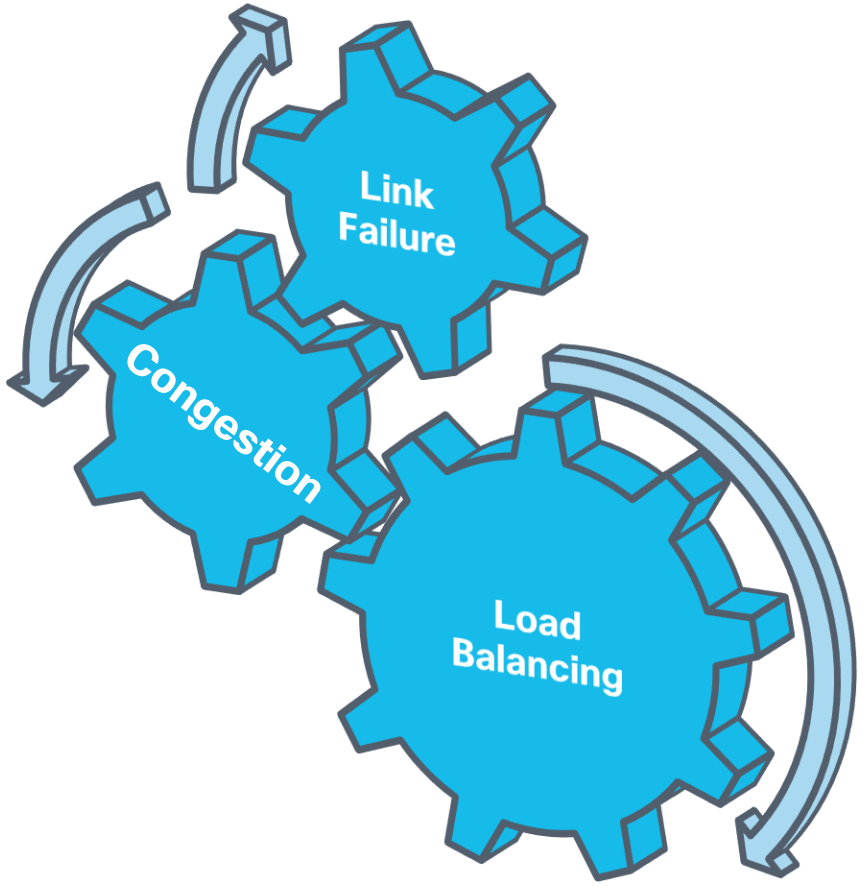
UE – Software – Libfabric



UE – Software – Libfabric Architecture



Revisiting *the AI Infrastructure Challenge*



Wrenches in the works

- Underutilized fabric links
- Head of Line blocking
- Incast Congestion
- Link failures and black holing

Greasing the skids

- Improved LB & Adaptive Path Selection
- Network influenced Congestion Management

AI Ethernet Fabric Options

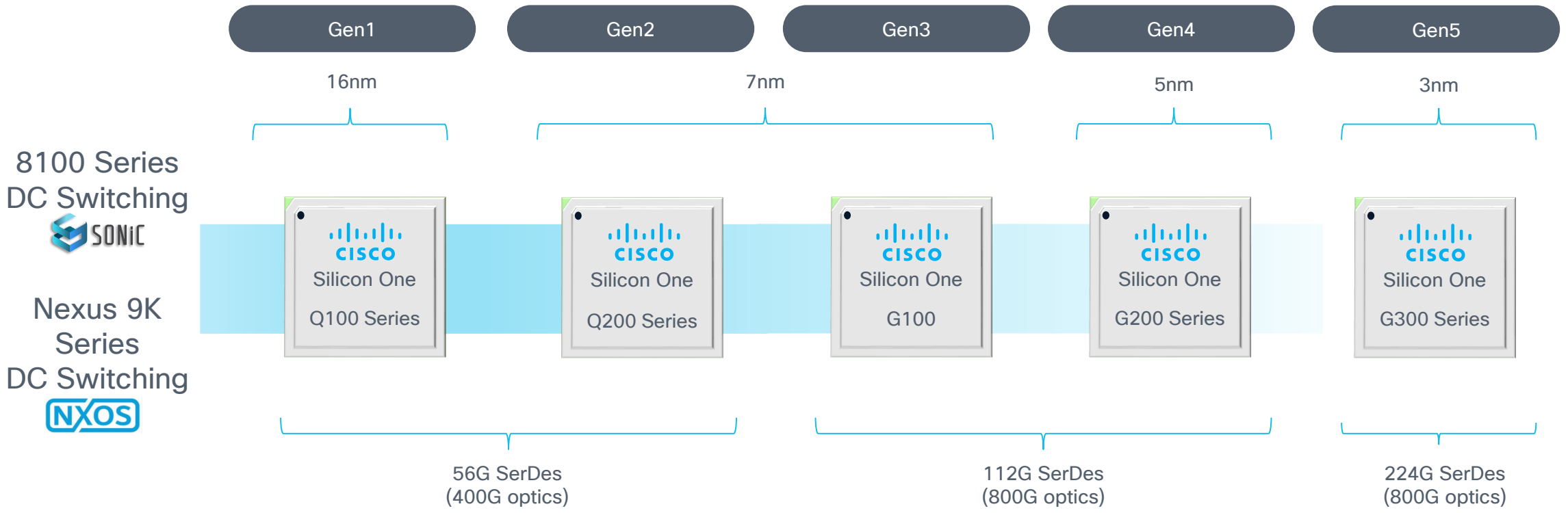
	1 Ethernet	2 Enhanced Ethernet		2U Ultra Ethernet	3 Scheduled Ethernet
Load Balance	Stateless ECMP	Stateful Flow/Flowlet	Spray & Re-order in SmartNIC	Endpoint Controlled adaptive packet spraying	Spray & Re-order in leaf
Fabric Congestion Management	Congestion Reaction with ECN/PFC	Congestion Reaction with congestion score to adjust distribution		Network influenced Congestion Management	Congestion Avoidance
Link Failure	Software	Hardware			
Job Completion Time	Good	Better		Even Better	*Best*
Coupling between NIC and Fabric	No		Yes		No
Place in Network	Frontend, Backend			Backend	Frontend, Backend
Fabric Architecture	Leaf/Spine or Modular Chassis				Modular Chassis

Effectiveness IS dependent on Traffic Characteristics

Effectiveness IS NOT dependent on Traffic Characteristics

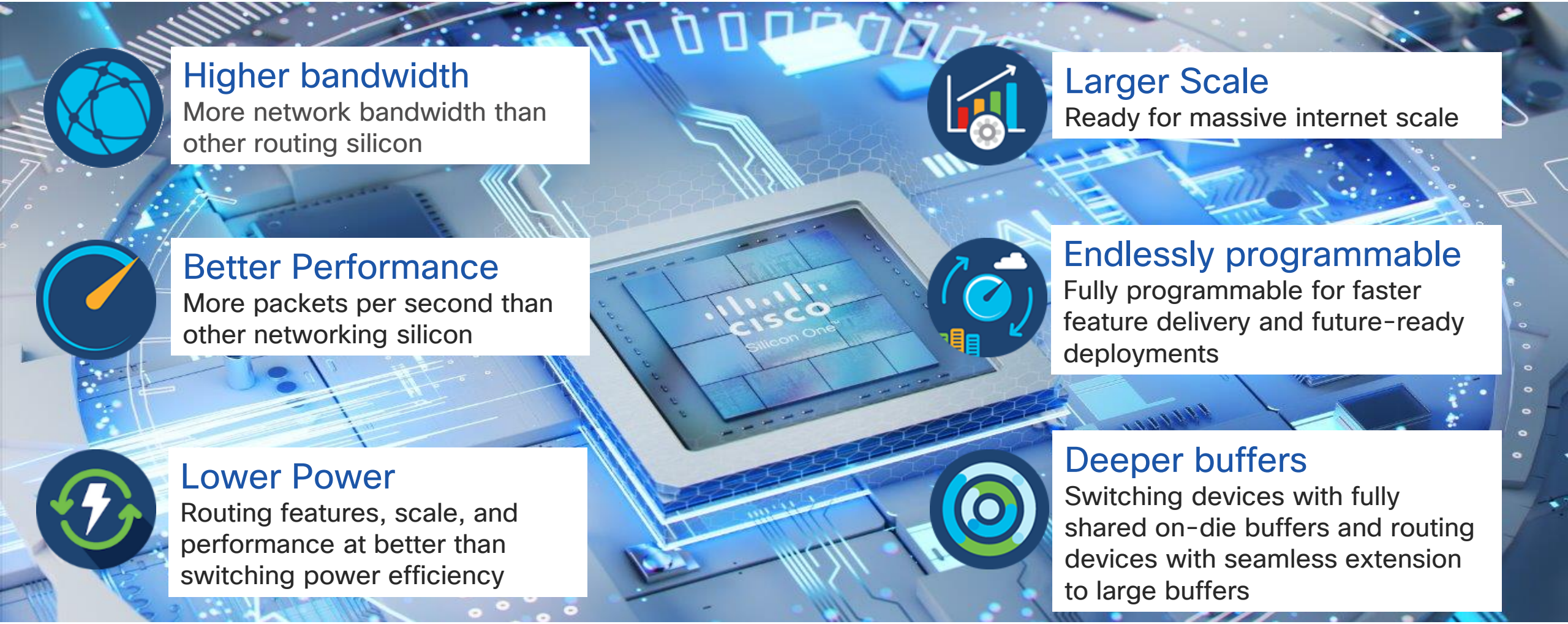
Building AI Infrastructure with Silicon One (Cisco 8000 & Nexus 9K)

Cisco Silicon One



One architecture. Unmatched capabilities

Unmatched programmability, performance, flexibility, and efficiency



Cisco Silicon One G200

Uniquely efficient and Optimized for AI/ML

One architecture
A simpler and easier network to maintain

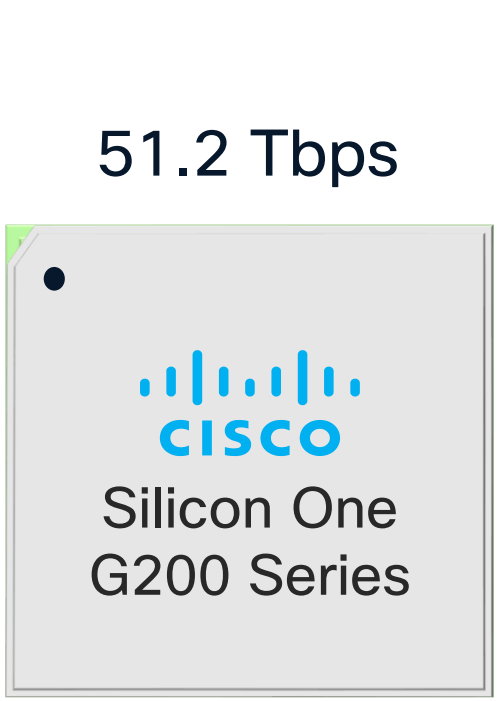
High Performance
2x higher performance than G100

Sustainability via technology
2x more power efficient than G100

Ultra-low latency*
2x Lower Latency than G100

Optimal network design
512-wide radix enables flatter, more efficient networks

Fully shared packet buffer
Optimal fairness, burst performance, JCT



Advanced 112 Gbps SerDes
Cisco designed next-generation ADC SerDes
Support for Std. Optics, 4-meter DAC and linear optics

Advanced load balancing*
Non-correlated WECMP avoids hash polarization
Congestion-aware stateful load balancing
Congestion-aware packet spraying

Link failure avoidance*
HW based traffic link failure redistribution optimizes real-world large-scale deployments

Programmable processor
Deterministic ultra-low latency processor with run to completion for ultimate flexibility

435B+ Lookups per second
Enables advanced features like SRv6 uSID

Deep visibility & Analytics*
In-band telemetry including emerging protocols
Hardware analyzers enable post event debuggability

cisco

Cisco's AI networking portfolio strategy

Cisco 8000

Tier 2 web
AI aaS

Hyperscalers

Customizable solution
BYO management, SONiC, or BYO-NOS

Shipping

Nexus 9000 with Nexus Dashboard

Enterprise
Public sector
Commercial

Service
providers

Tier 2 web
AI aaS

Network solution
Private cloud managed, interoperable

Shipping

Nexus Hyperfabric for AI

Enterprise
Public sector
Commercial

SaaS solution
Public cloud managed, full stack

** H1 CY25*

Register Right now for BRKNWT-2504 to hear about Ethernet Evolution and Challenges in the World of AI

Cisco 8100 Fixed SONiC Portfolio Roadmap

Datacenter Switching

8101-32FH-O
32x 400G (Q200L)

8102-28FH-DPU-O
28x 400G, 4x DPU SLEDs (Q200L)

8101-32FH-O-C01
32x 400G (Q200L)

8101-64H-O
64x 100G (Q200L)

8122-64EH-O
64x 800G QSFP-DD800 (G200)

8122-64EF-O
64x 800G OSFP800 (G200)

8122X-64EF-O
64x 800G OSFP800 (G200x)

8121-32EF-O
32x 800G OSFP800 (G202x)

In Production	2025	TBD

8122-64EHF-O

'Superbolt'

AI Primed

SONiC

64 x 800G
Shipping

Cisco Silicon One G200

Hardware Summary
Single 51.2T G200 ASIC (5nm) 256 MB SRAM packet buffer
Eight Core x86 CPU 64 GB DRAM
RS-232, and 10 GbE control plane expansion. 1 GbE Management Ports 1x QSFP PIE/Telemetry, 1xUSB 2.0
4 Fans, 1+1 PSU Redundancy Port Side Intake airflow
3kW AC & 3kW DC (ORv3) PSUs
(H) 3.45 x (W) 17.3 x (D) 24.7 in. (H) 8.76 x (W) 43.95 x (D) 67.4 cm 37 lbs - 54.5 lbs (16.7 - 24.78 kg)

- 51.2T G200 Optimized for high-radix DC and AI applications
- 64x800G **OSFP** (IHS)
- ETC 800 GbE support
- 512 x 100G – full 512 interface network radix
- Exceptional Serdes performance for powering LPO optics

Ultra Ethernet Consortium READY

8122-64EH – O
'Lightning'

AI Primed

SONiC

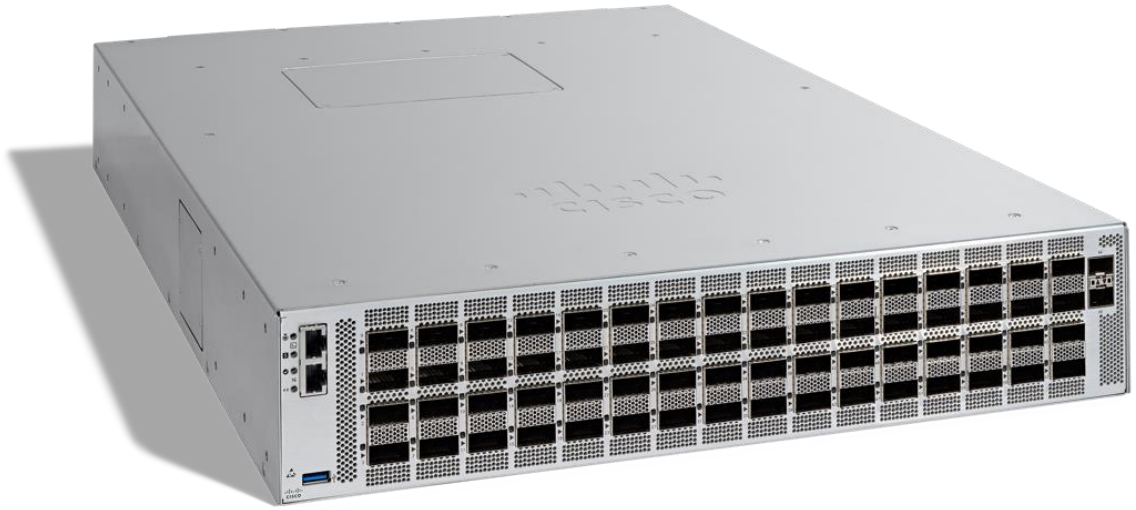
64 x 800G
Shipping

Cisco Silicon One G200

Hardware Summary
Single 51.2T G200 ASIC (5nm) 256 MB SRAM packet buffer
4-Core Core x86 CPU 64 GB DRAM
RS-232 and1 GbE Management Ports 2xSFP25G PIE Telemetry Ports, 1xUSB 2.0
4 Fans, 1+1 PSU Redundancy Port Side Intake airflow
3kW AC & 3kW DC (ORv3) DC PSUs
(H) 3.45 x (W) 17.3 x (D) 24.7 in. (H) 8.76 x (W) 43.95 x (D) 67.4 cm 37 lbs - 54.5 lbs (16.7 - 24.78 kg)

- 51.2T G200 Optimized for high-radix DC and AI applications
- 64x800G **QSFP-DD**
- ETC 800 GbE support
- 512 x 100G – full 512 interface network radix
- Exceptional Serdes performance for powering LPO optics

Ultra Ethernet Consortium READY



Cisco Nexus 9000 Series Portfolio Roadmap

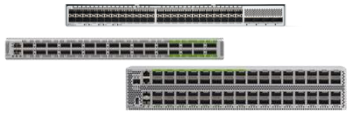
Datacenter Switching

Shipping | *New + Roadmap*

Nexus Fixed

Nexus 9300
48p 25G + 6p 100G
36p & 64p 100G

Nexus 9300
48p 50G + 4p 400G
32p & 64p 400G



Cisco N9300 Smart Switch
24p 100G &
48p 25G + 2p 100G + 6p 400G

Nexus 9300 SG2
64p 800G (QDD & OSFP)

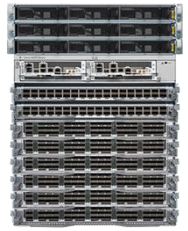
Nexus 9300 SP2R
64p 800G (QDD & OSFP)

Nexus Modular

Nexus 9500
Up to 576
100G ports

Nexus 9400
64p 400G (or)
128p 200G

Nexus 9800
Up to 288
400G ports



Nexus 9800
Up to 288
800G ports

Port Speed

100G | 400G | 800G

Cisco Silicon

<= 3.6T
25G Serdes

4.8 – 25.6T
50G Serdes

>= 51.2T
100G Serdes

N9364E-SG2-Q

'Bud-Lightning'

AI Primed

Ultra Ethernet Consortium READY

NXOS

64 x 800G
Shipping

cisco Silicon One G200

Hardware Summary
Single 51.2T G200 ASIC (5nm) 256 MB SRAM packet buffer
4-Core Core x86 CPU 32 GB DRAM
RS-232 and1 GbE Management Ports 2xSFP25G PIE Telemetry Ports, 1xUSB 2.0
4 Fans, 1+1 PSU Redundancy Port Side Intake airflow
3kW AC & 3kW DC (ORv3) PSUs
(H) 3.45 x (W) 17.3 x (D) 24.7 in. (H) 8.76 x (W) 43.95 x (D) 67.4 cm 37 lbs - 54.5 lbs (16.7 - 24.78 kg)

QSFP-DD

N9364E-SG2-O

'Optimator'

AI Primed

Ultra Ethernet Consortium READY

NXOS

64 x 800G
Shipping

cisco Silicon One G200

Hardware Summary
Single 51.2T G200 ASIC (5nm) 256 MB SRAM packet buffer
8-Core Core x86 CPU 32 GB DRAM
RS-232 and1 GbE Management Ports 2xSFP25G PIE Telemetry Ports, 1xUSB 2.0
4 Fans, 1+1 PSU Redundancy Port Side Intake airflow
3kW AC & 3kW DC (ORv3) PSUs
(H) 3.45 x (W) 17.3 x (D) 24.7 in. (H) 8.76 x (W) 43.95 x (D) 67.4 cm 37 lbs - 54.5 lbs (16.7 - 24.78 kg)

OSFP

Wrap-up and Questions

Key Takeaways

AI presents a ***new challenge*** to the way networks are built

Scaling-up and Scaling-out with Ethernet ***gets the network out of the way***

Choice of Parallelism has a profound impact on cluster performance

Cisco's Ethernet options are ***open, flexible and ready*** for the AI challenge

One solution with ***Silicon One for any AI Ethernet fabric option.***

Complete your session evaluations

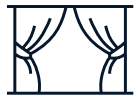
Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live mobile app.

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 www.CiscoLive.com/on-demand

Contact me at: sramesh@cisco.com

Thank you

cisco Live !


CISCO