Microsoft AI Ready WAN Infrastructure with Cisco 8000

Rúben Fonte
Customer Delivery Architect - Cisco

Rustam Gaisin
Senior Cloud Network Engineer – Microsoft

Sarav Subramanian
Principal Cloud Network Engineer – Microsoft

CSSENT-2100

Cisco Webex App

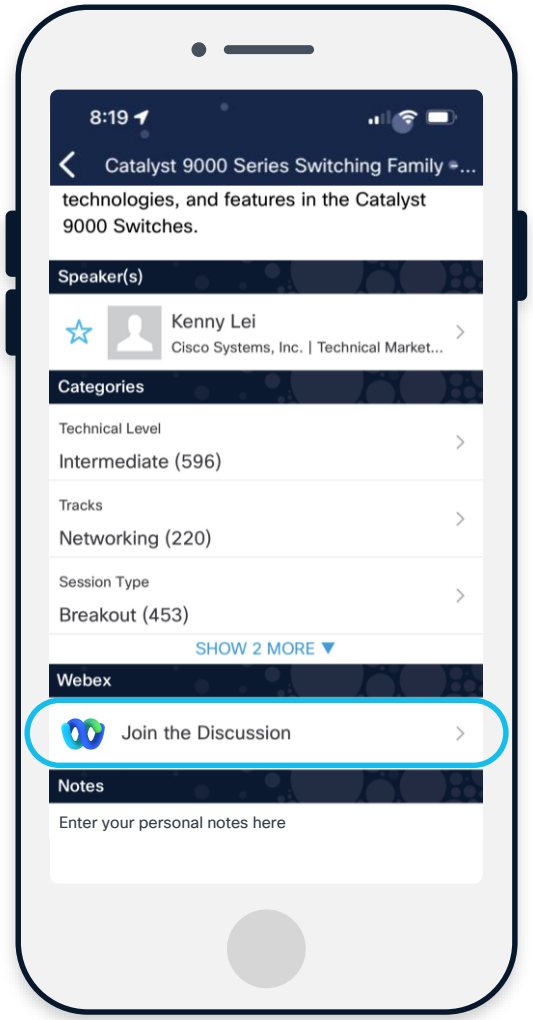
Questions?

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.



DO NOT post any of this content to any blogs or external websites

DO NOT take photos or video of sessions or slides

Speakers

Rúben Fonte

CISCO

Rustam Gaisin

Microsoft Azure

Sarav Subramanian

Microsoft Azure

Agenda

01 Introduction

02 AI Requirements and IP Backbone

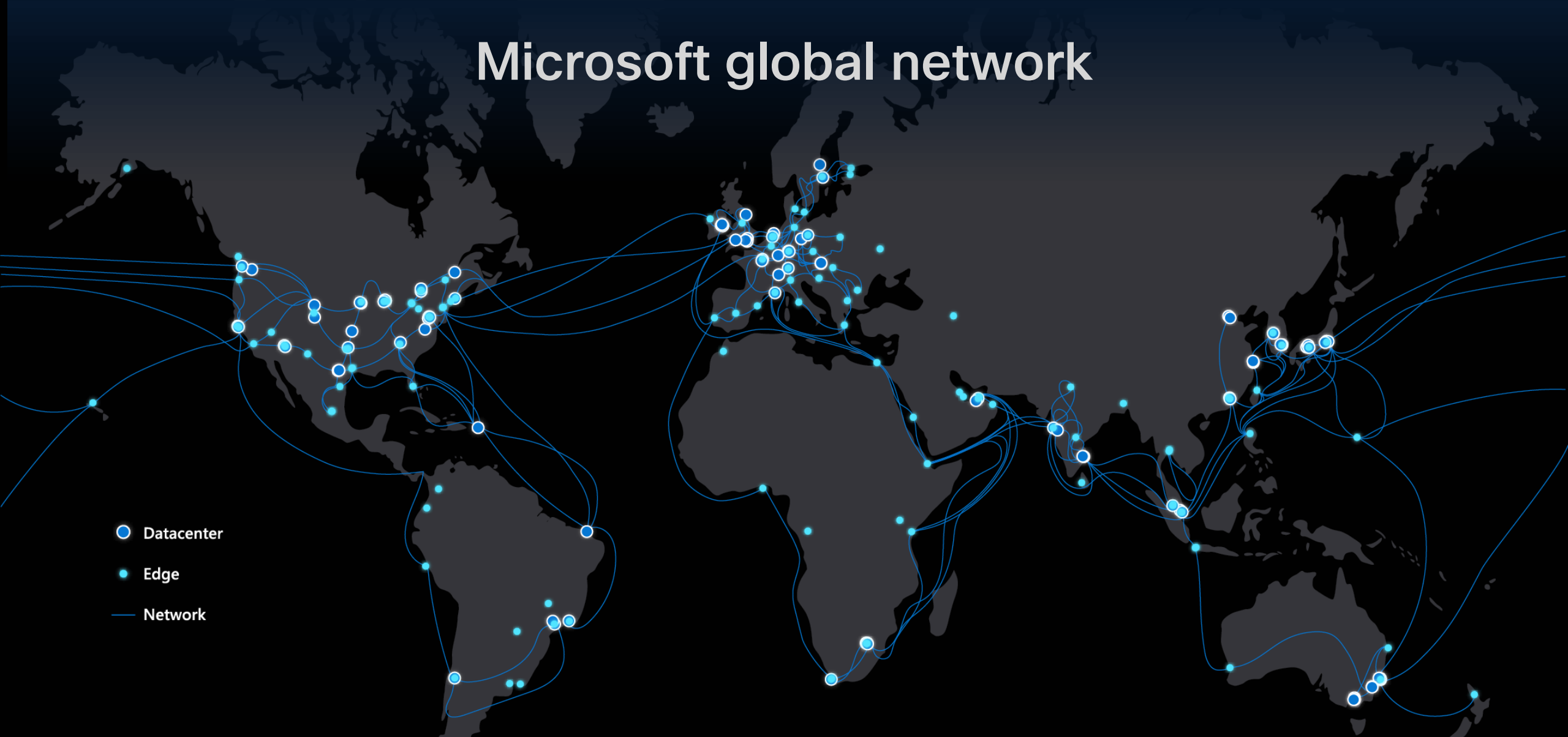
03 Solution

04 Challenges

05 Conclusion

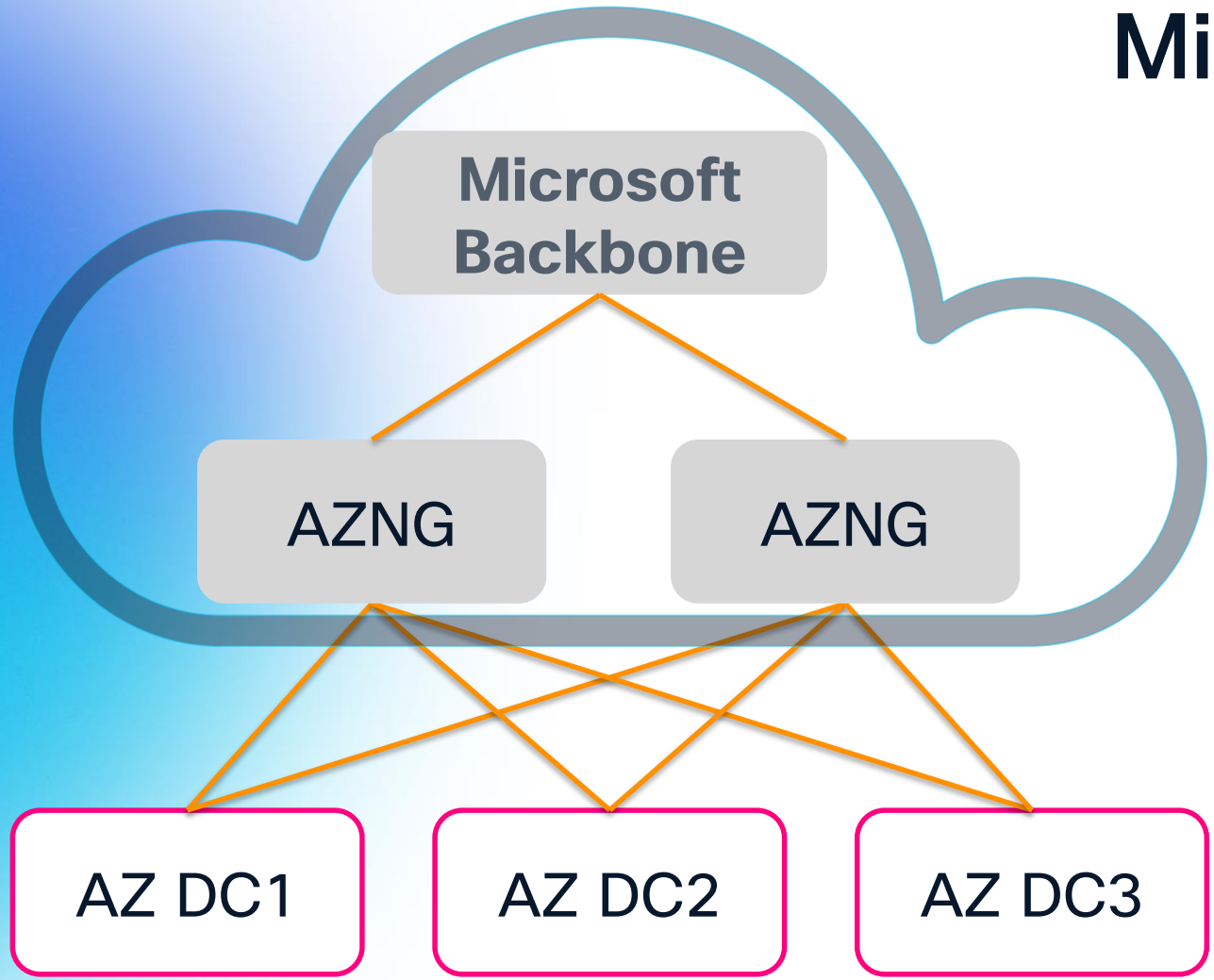
Introduction

Microsoft global network



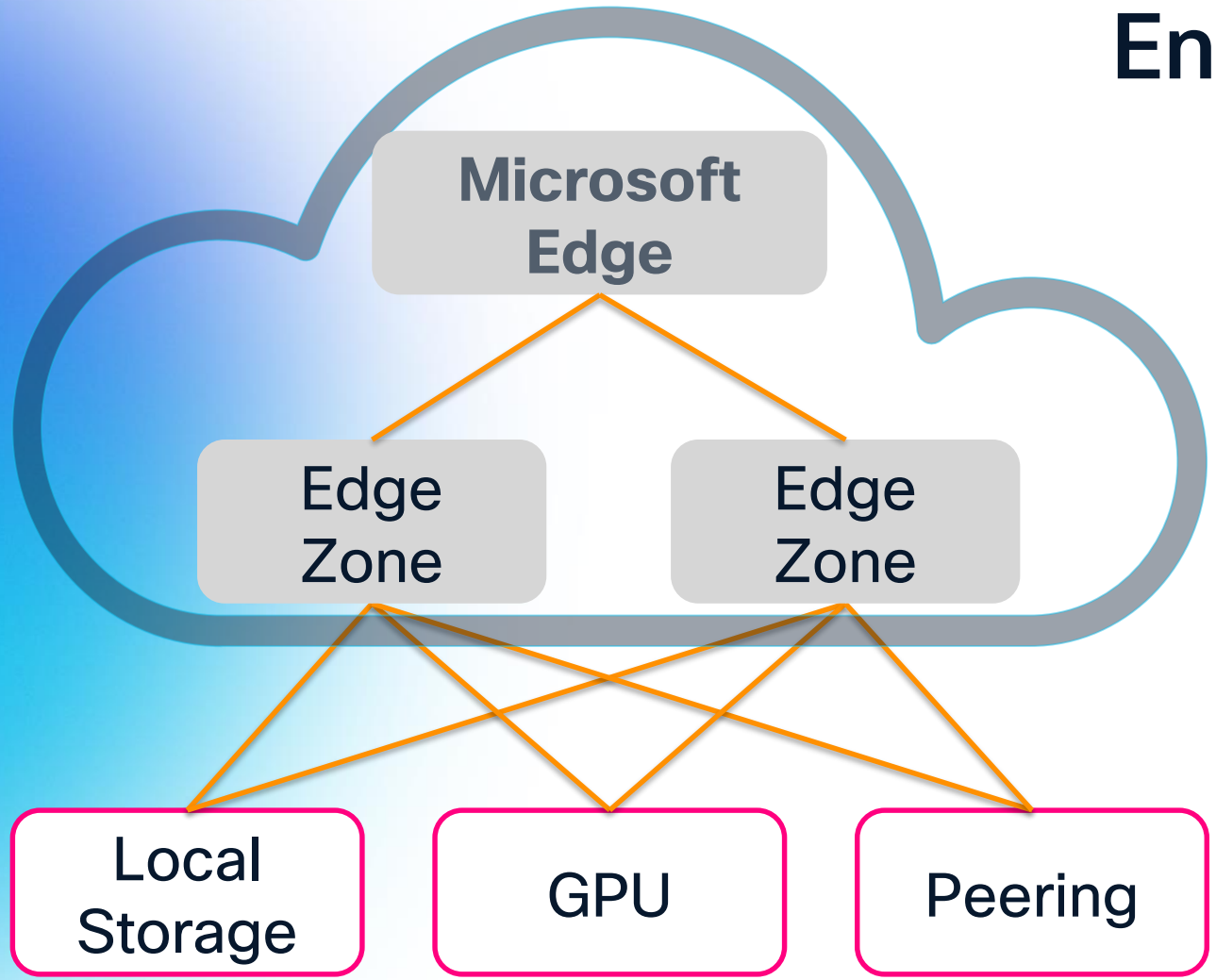
60+ Azure regions | 225k+ miles of fiber | 2Pbps+ WAN Capacity | 185T Peering Capacity | 20k+ Peering Connections

Microsoft WAN



- RNG's connecting over MSFT BB
- DC connecting to multiple RNG for redundancy

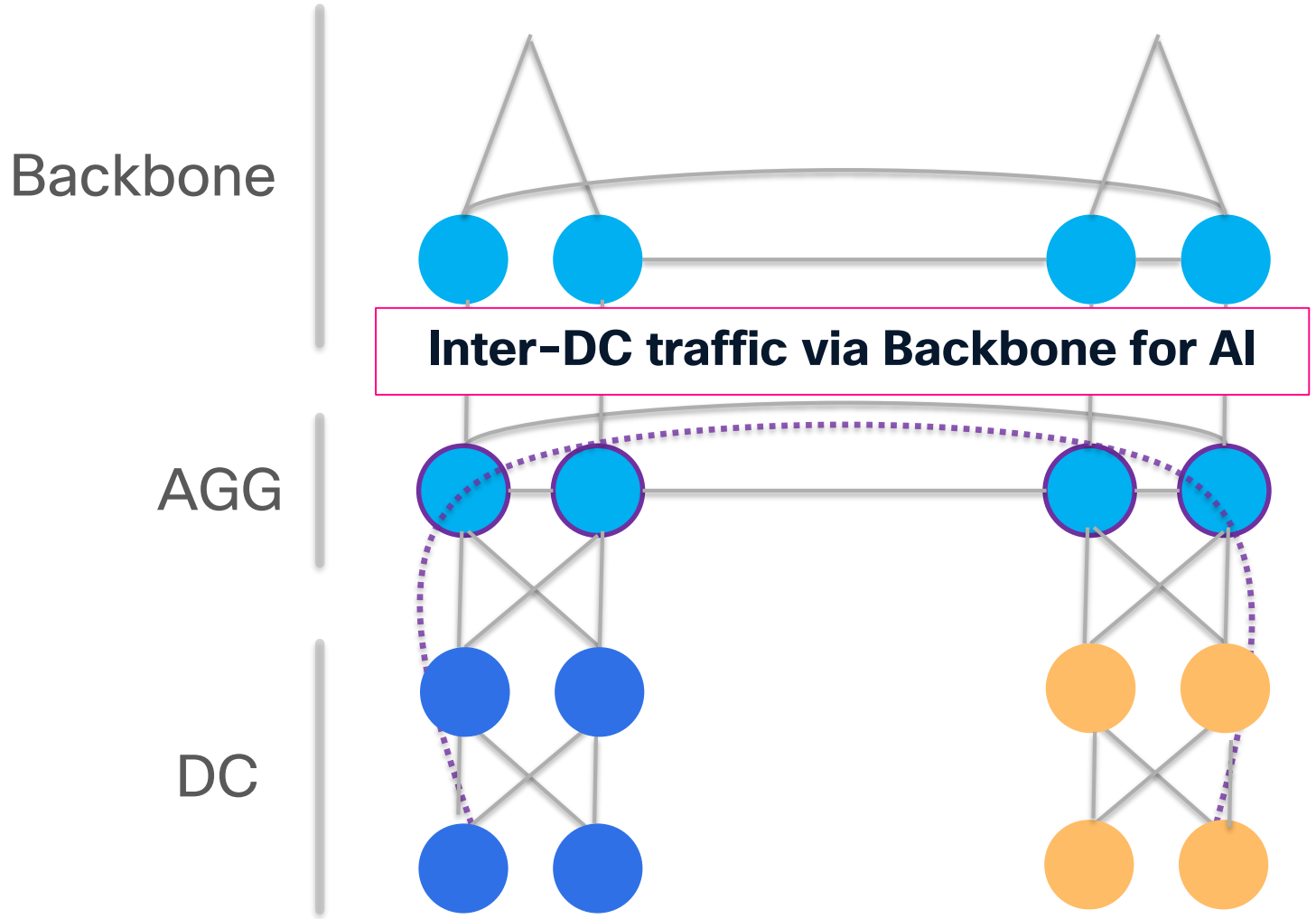
Environment



- Edge Zone connect to parent WAN region
- Transmit Tbps throughput to local Edge Zone

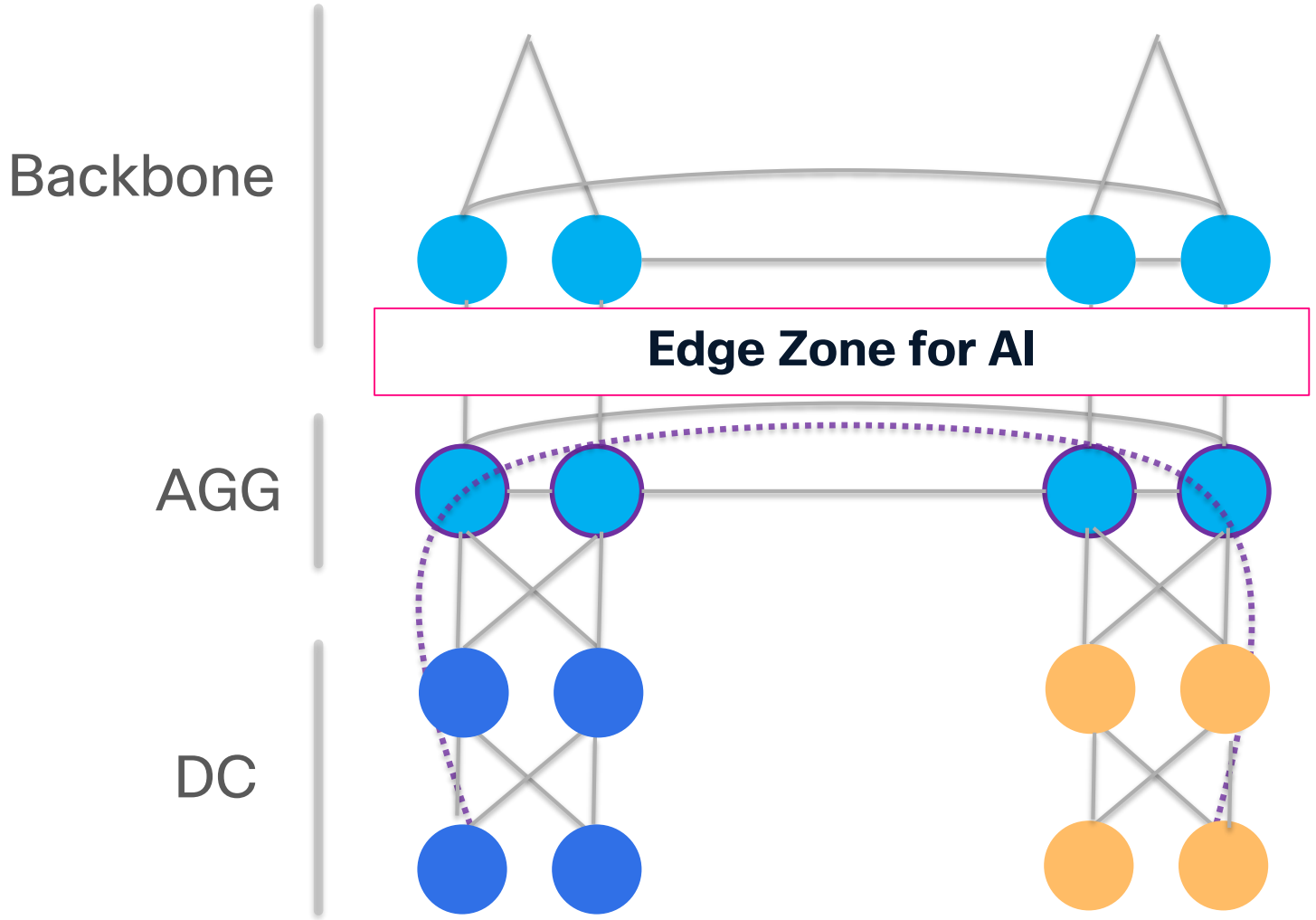
AI Requirements and IP Backbone

Architecture



- High Throughput
- Low Latency
  Milliseconds(*ms*)
- Lossless
  No drops

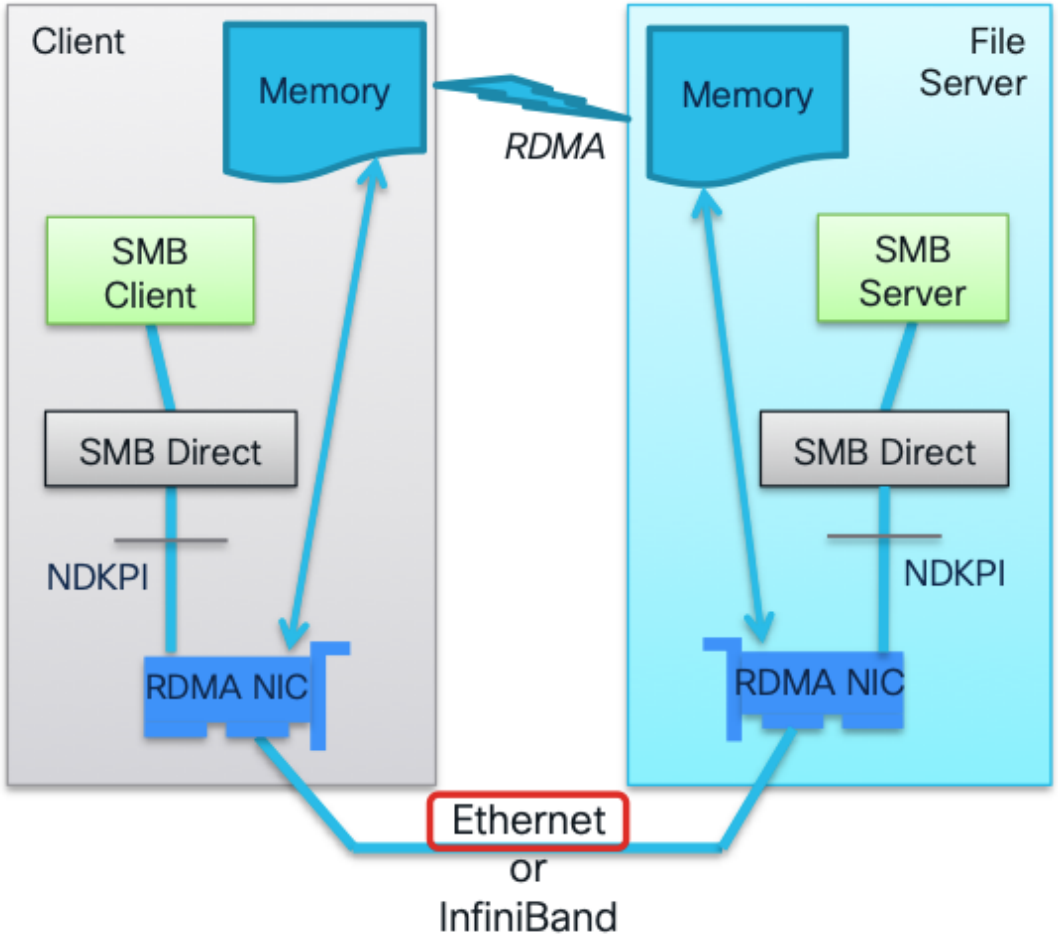
Architecture



- Local BE Traffic
- RoCEv2, RDMA and TCP
  - Buffer Sharing

How to get High Throughput and Low Latency?

Why RDMA?



Remote direct memory access (RDMA) is a well-known technology used for high performance computing (HPC).

Advantages of RDMA:

1. High throughput(>40Gbps)
2. Low latency transfer memory-to-memory
3. Do not burden the CPU.

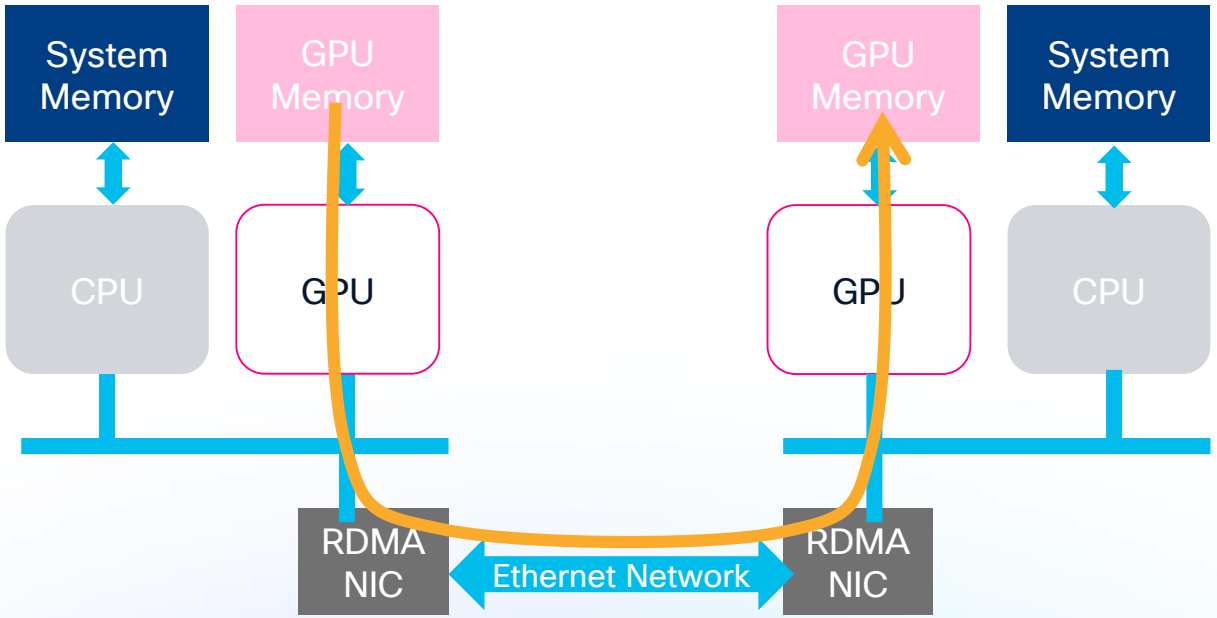
RDMA Hardware Technologies:

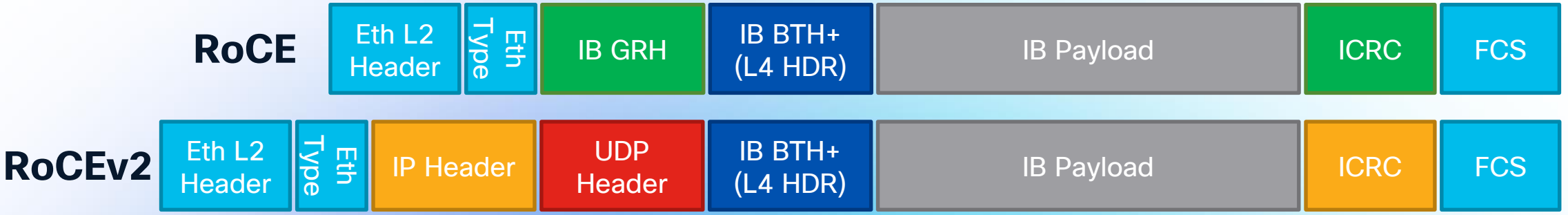
InfiniBand dedicated high-performance solution

RoCEv2 enables RDMA over standard Ethernet networks

iWARP: RDMA over TCP/IP

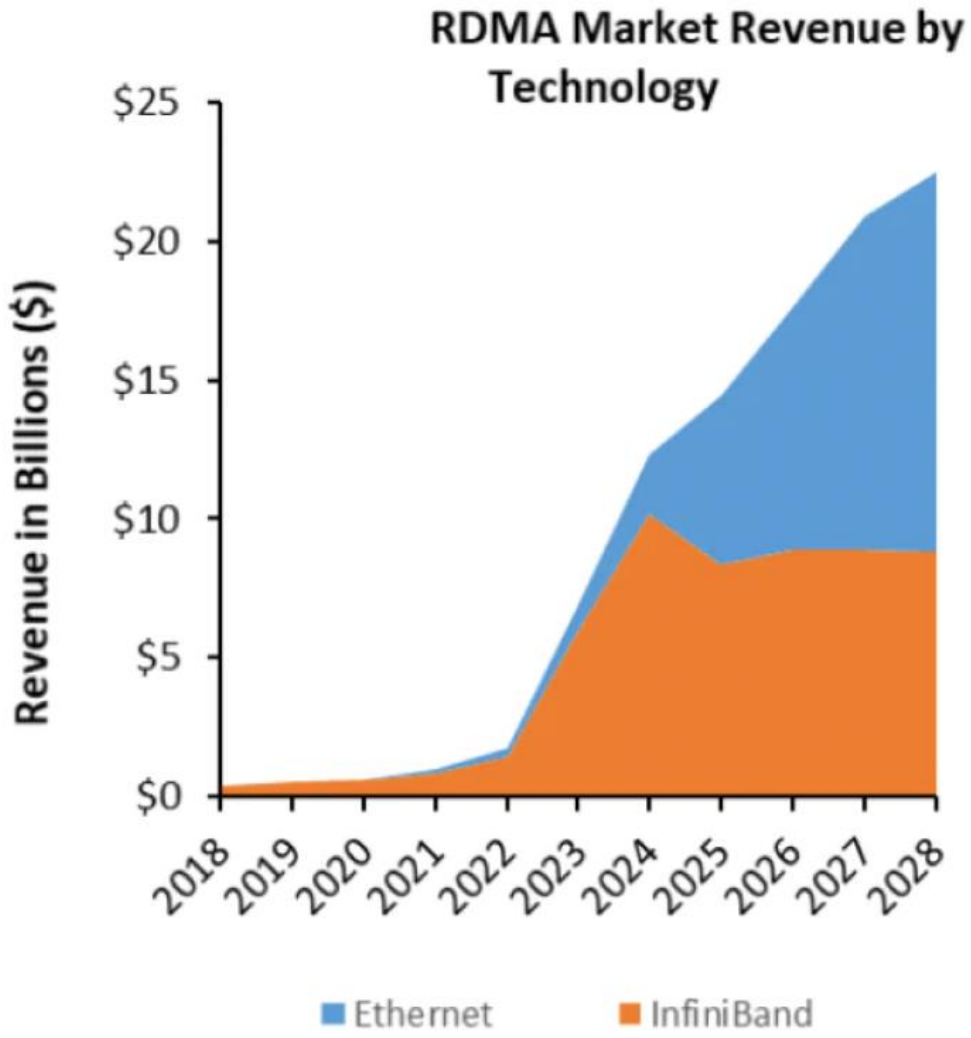
RDMA Over Converged Ethernet – RoCE





https://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet

RDMA Trends



RDMA is predominantly deployed with InfiniBand, but RDMA over Ethernet is getting more traction as expected as adoption increases.

Why?

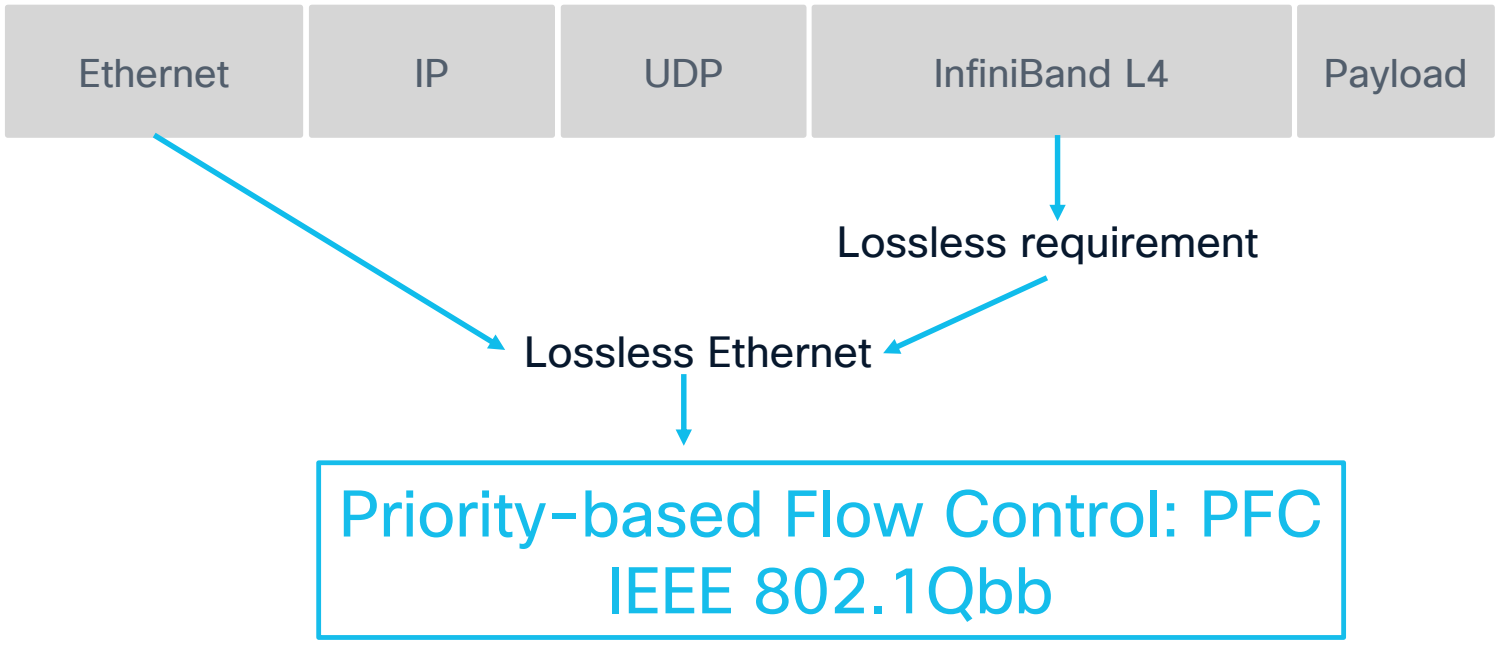
- Widely supported a vast number of vendors
- Cost efficiency

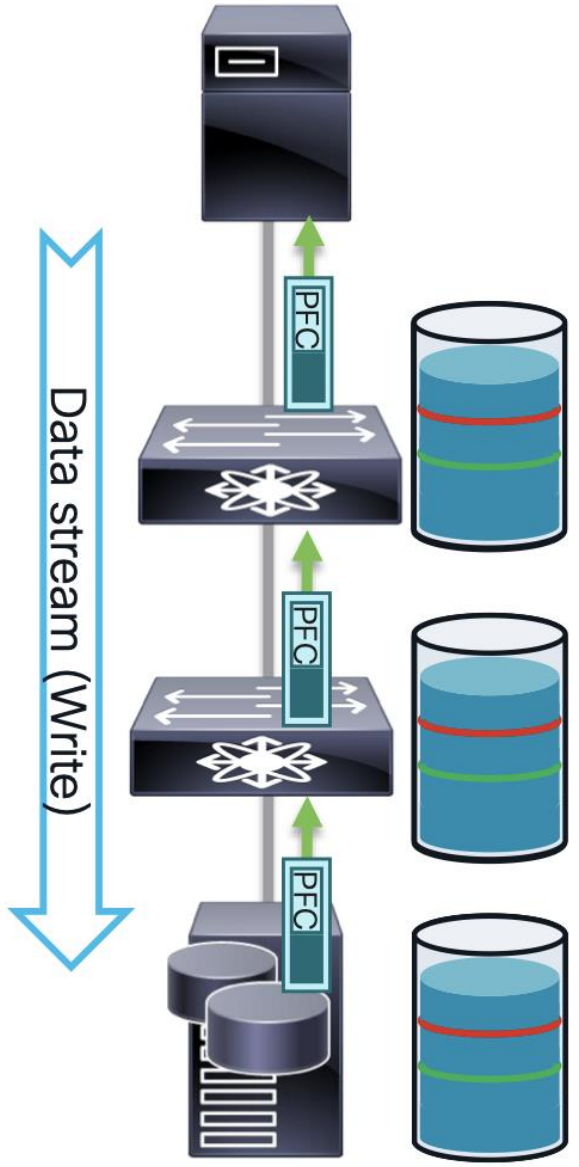
https://www.naddod.com/blog/why-ai-ml-networks-rely-on-rdma

How to achieve lossless?

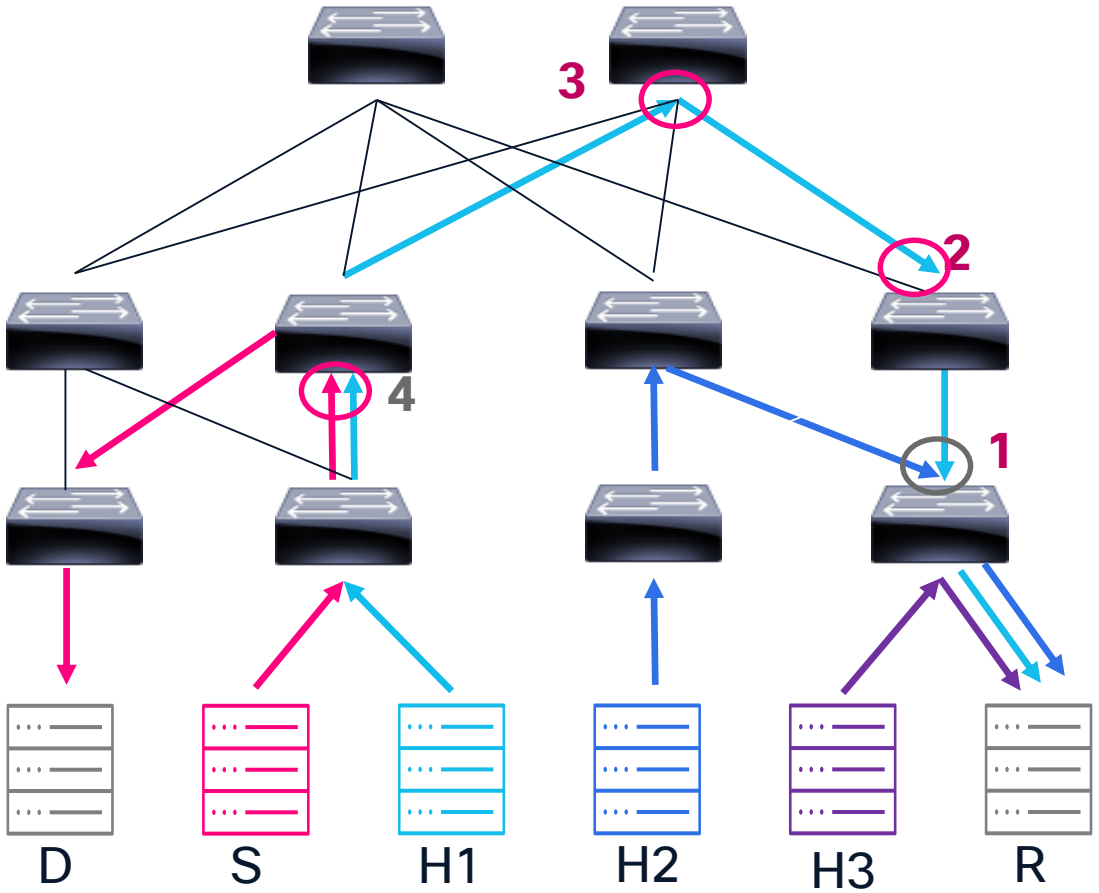
Priority-base Flow Control (PFC)

- PFC frames are sent from node that has experienced congestion toward sender
- PFC is propagated on a hop-by-hop basis





Priority-base Flow Control (PFC)



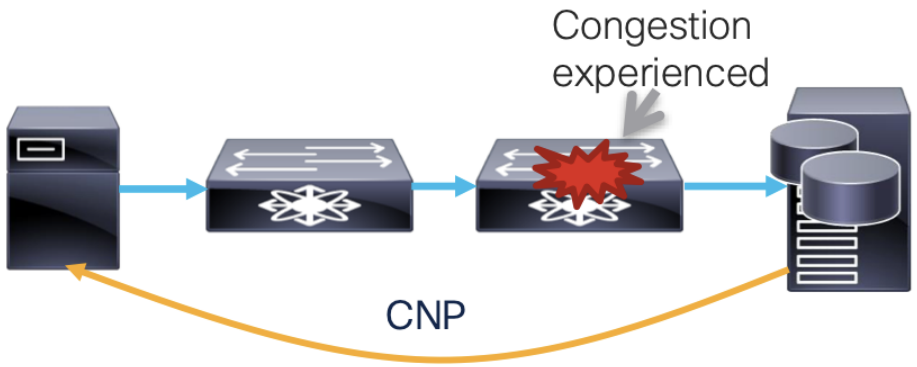
PFC operates on port-level, instead of flow-level

Solution?

DCQCN with ECN for flow control and PFC as last resort

DCQCN with ECN

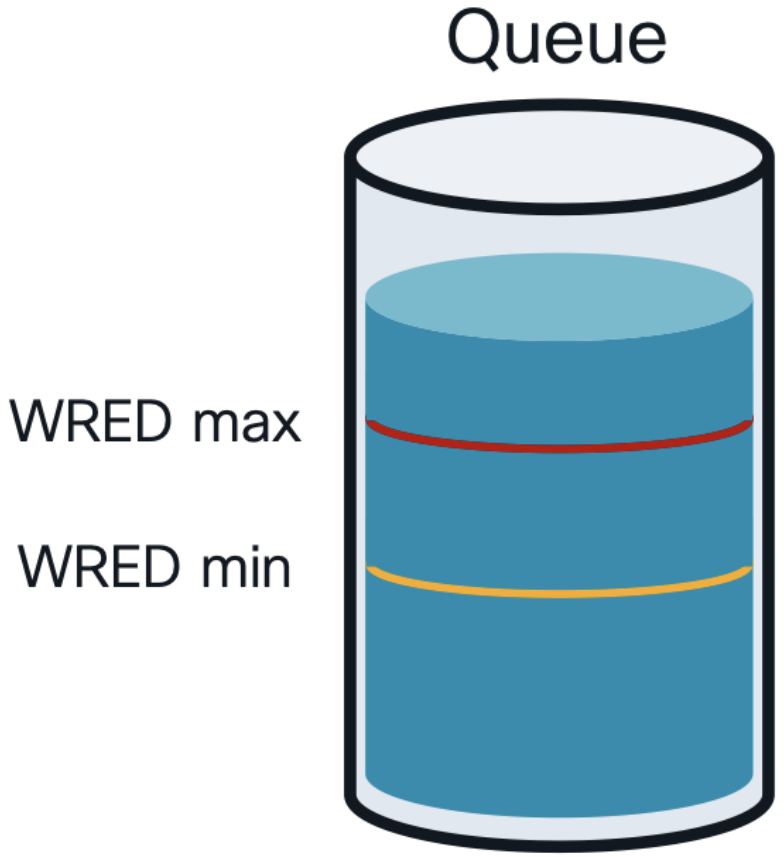
- Data Center Quantized Congestion Notification (DCQCN) is a per-flow congestion control using ECN
- IP Explicit Congestion Notification is used for congestion notification.
- ECN enables end-to-end congestion notification between two endpoints on IP network.
- In case of congestion, ECN gets transmitting device to reduce transmission rate using Congestion Notification Packet (CNP) without pausing traffic.



ECN	ECN Behavior
0x00	Non ECN Capable
0x10	ECN Capable Transport (0)
0x01	ECN Capable Transport (1)
0x11	Congestion Encountered

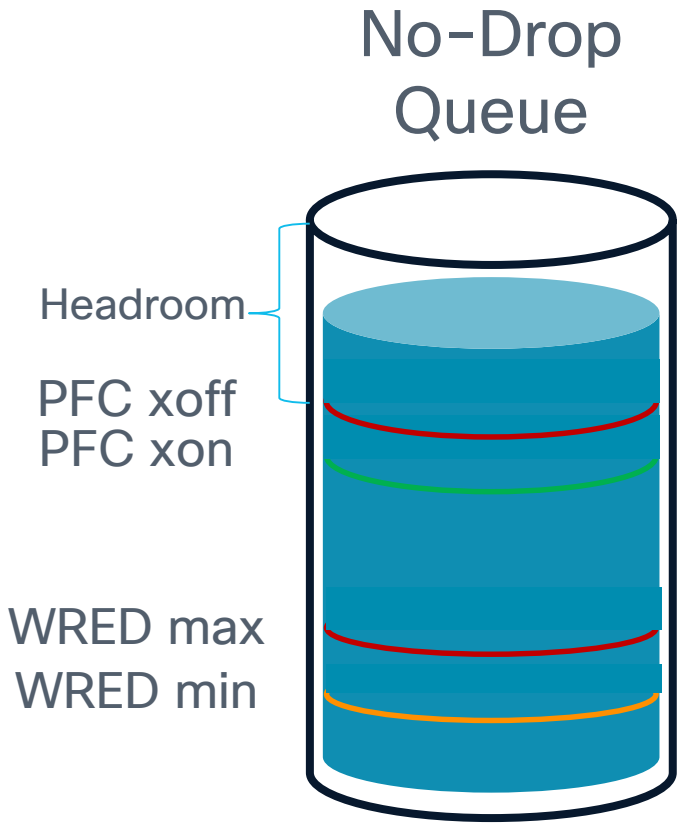
DCQCN with WRED

- WRED (Weighted Random Early Detection) is used to signalize severity of congestion
- ECN is not marked when buffer usage is below WRED min threshold
- When buffer usage is minimal threshold, Congestion Encountered will be marked on N number of randomly selected packets (probability parameter)
- After buffer usage crosses MAX threshold, every ECN capable packet will be marked with Congestion Encountered

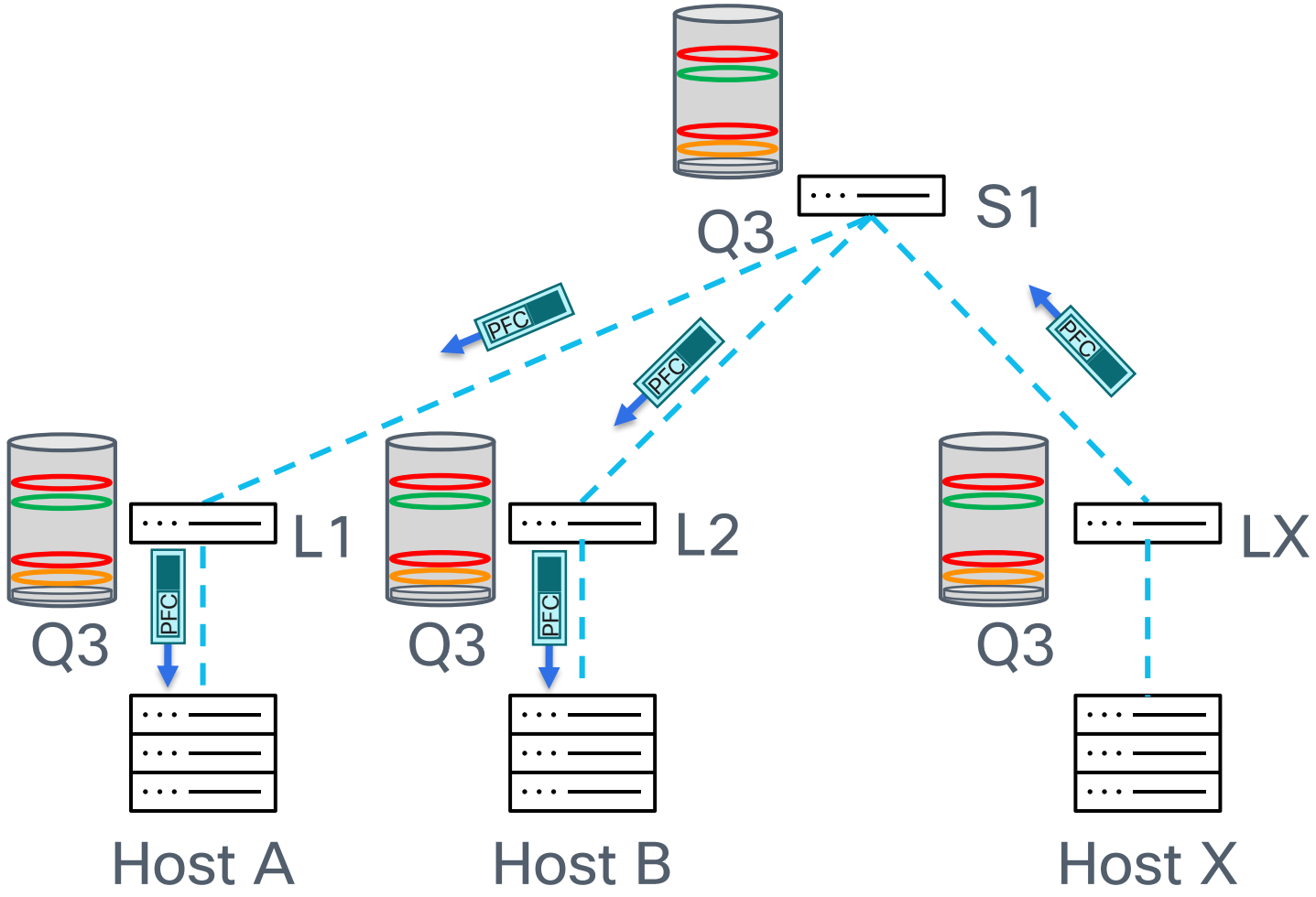


PFC and ECN together

- WRED threshold are set low in no-drop queue
  - Signalize early for congestion, give enough time for end points to react
- PFC threshold are set higher than ECN
  - In case oversubscription buffers can be filled quickly without giving time to ECN to react
  - PFC will react and mitigate congestion



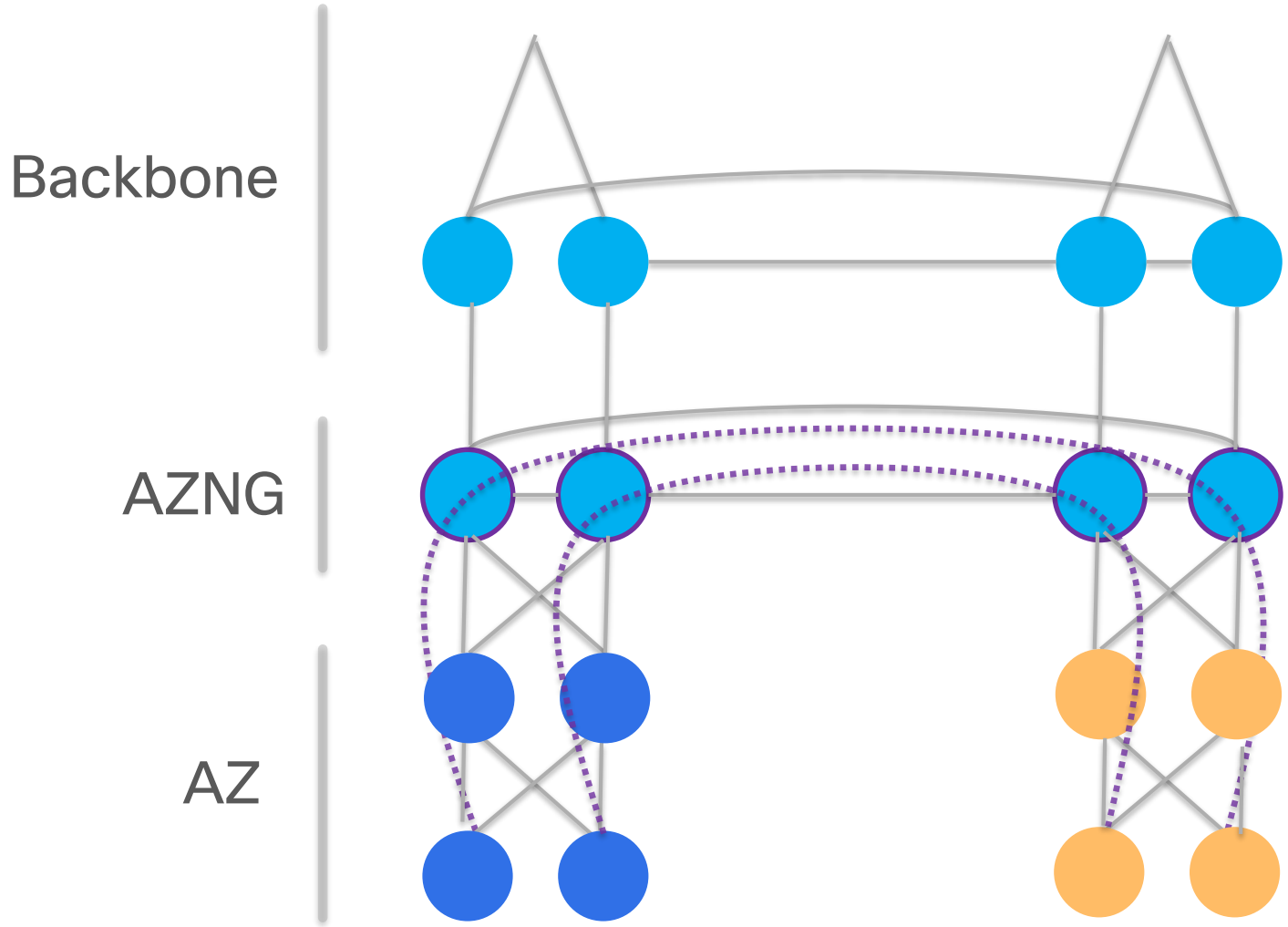
ECN and PFC Together to Build Lossless Ethernet Networks

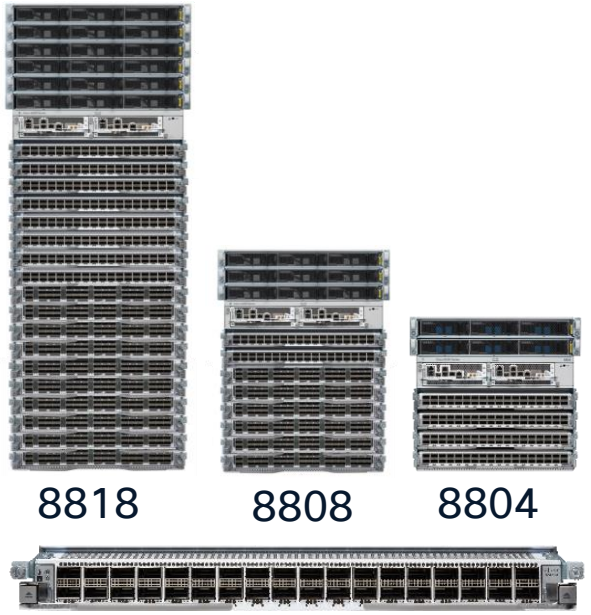


- Buffer occupancy crosses WRED minimum threshold, and Leaf X marks ECN in IP header
- Host X informs hosts A and B about network congestion by sending them CNP packets
- PFC signaled from Leaf X to Spine 1, pausing traffic from the spine switch to the leaf switch
- PFC signaled from Spine 1 to leaf switches 1 and 2, pausing traffic from Leaf 1 and Leaf 2 to the spine switch
- Leaf 1 and Leaf 2 send PFC frames down to hosts A and B, which mitigates congestion

Solution

Architecture



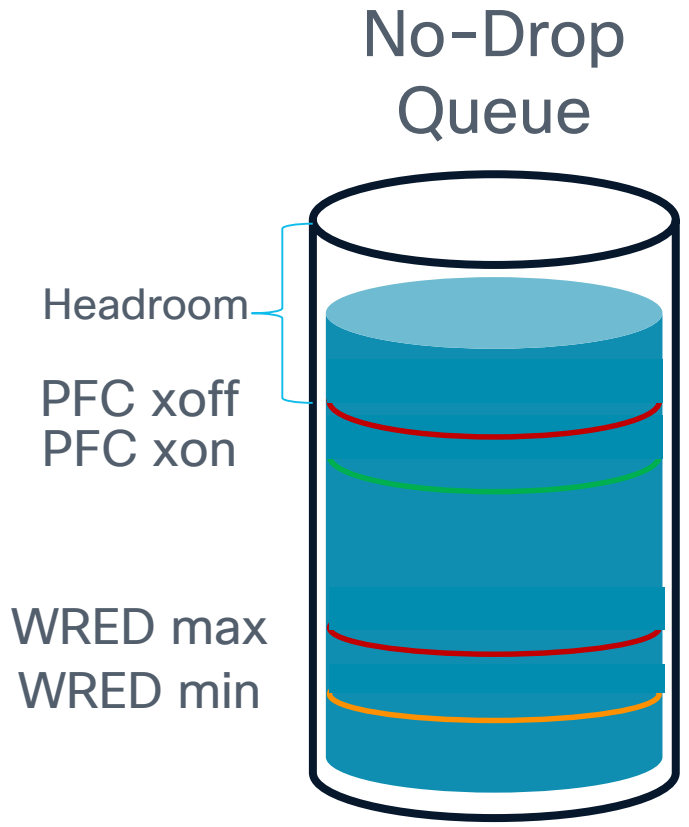


LSR/Core
AZ1
AZ2
AGG/RDMA enable

Router Buffer tunning

PFC, operates on port-level, instead of flow-level

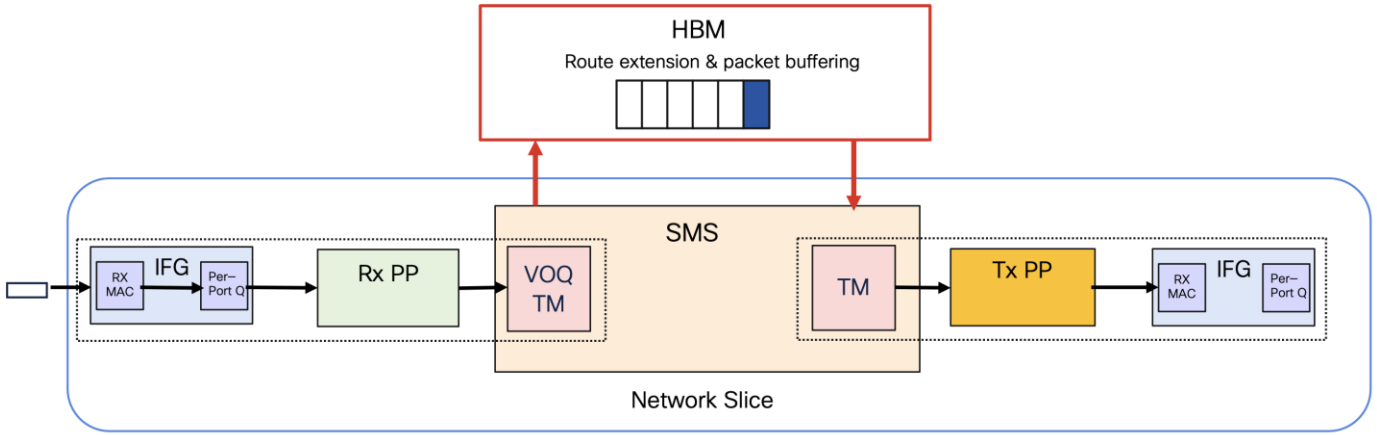
Router should mark ECN before firing PFC

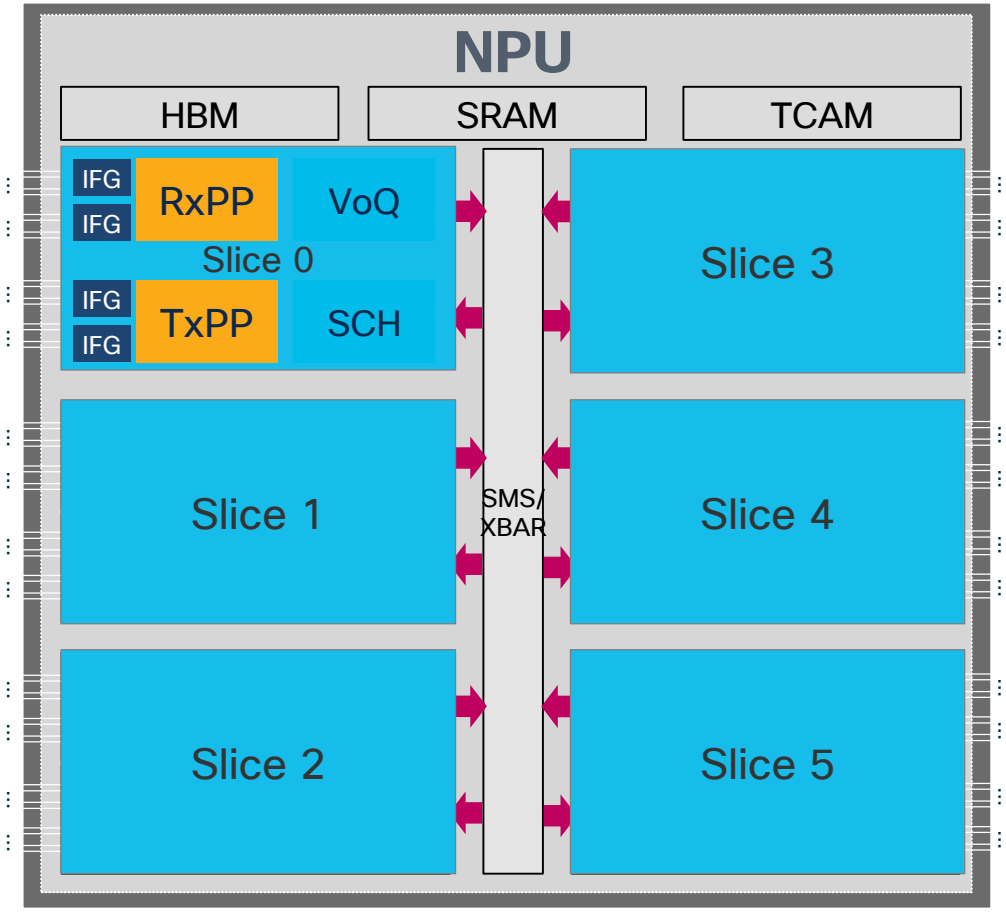


Challenges

PFC and non-PFC memory

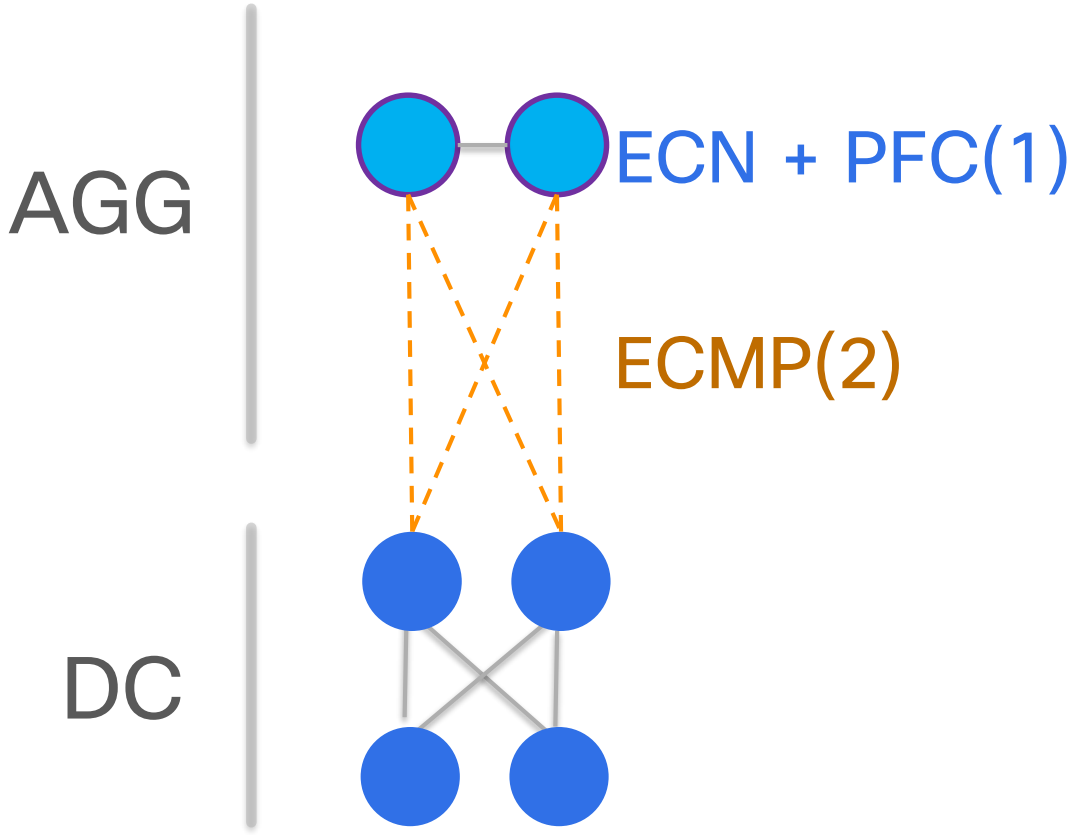
- Expandable packet buffer to HBM (8GB)
- PFC and non-PFC will share memory from HBM





Cisco Silicon One Q200

Challenges



1- ECN + PFC buffers has been improved

2- Load Balance improvements

Conclusion

Take Aways

- AI ready on WAN brought confidence to inter-DC traffic, allowing strong resilience and scale capability
- Microsoft and Cisco will continue evolving on scale AI network with new requirements and AI workloads

Complete your session evaluations

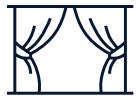
Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.

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

Thank you

cisco Live !


CISCO