
cisco Live!
Let's go

The bridge to possible

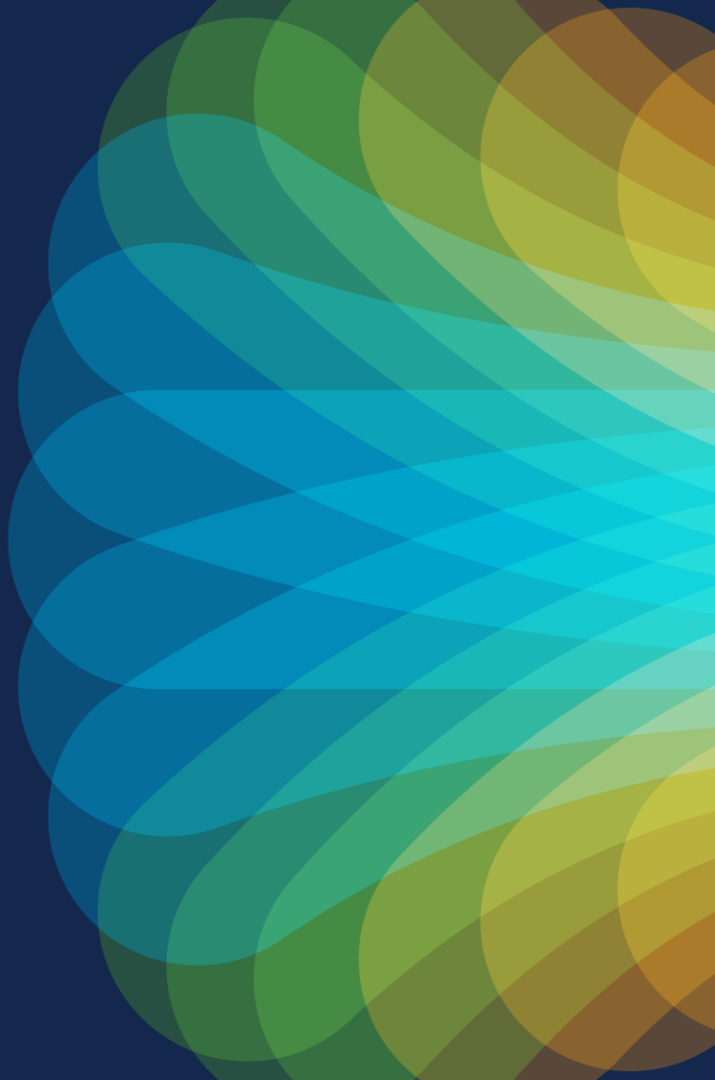
Network Best Practices for Artificial Intelligence Data Centre

Nemanja Kamenica, Technical Marketing Engineer

Agenda

- Why AI is important today and, in the future
- Network For AI Cluster
- Automation and Visibility
- The Blueprint For Today

Why AI is important today and, in the future



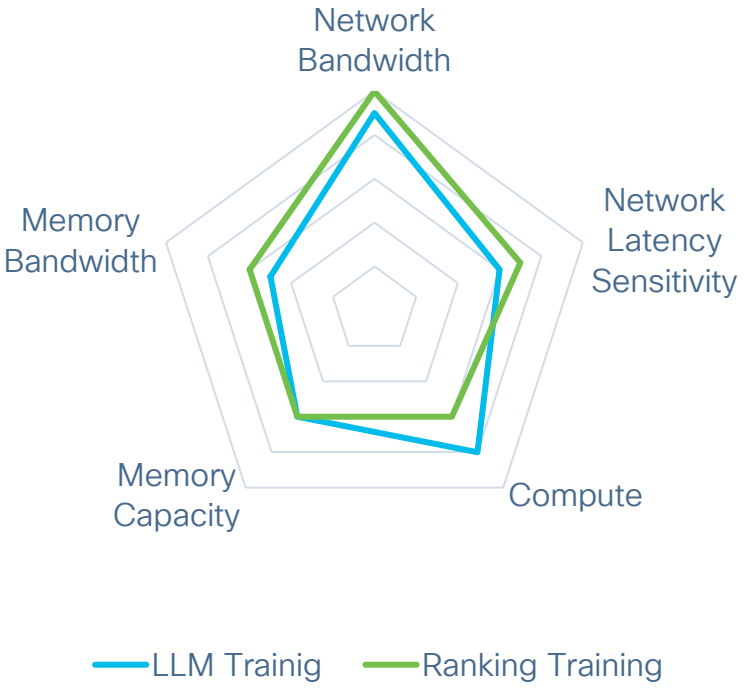
AI/ML Can Help Many Industries

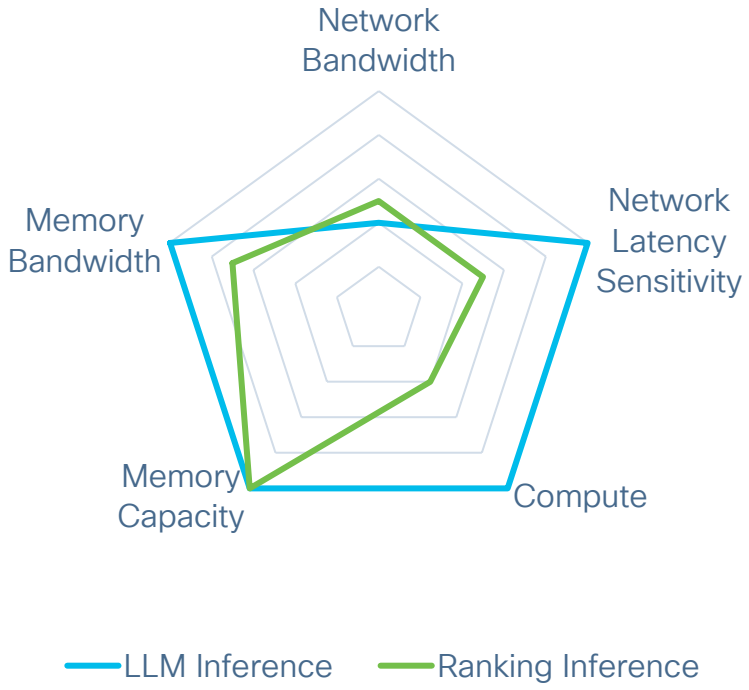
Healthcare	Financial Services	Public Sector	Media and Entertainment	Manufacturing	Retail
Medical Risk Prediction	High Frequency Trading Analysis	Intelligent Public Transport	Speech Recognition	Visual Inspection	Personalized Recommendation
Early diagnostics	Quant Research	Security Log Analytics	Natural Language Processing	Anomaly Detection	Demand Forecasting
Medical Research	Fraud and Risk Analytics	Disaster Recovery Assistance	Content Classification	Asset Management	Visual Search

AI cluster types and interaction with network

	Distributed Training	Production Inference
Node to Node Bandwidth	High	Low
Key Metric	Training time of a model	High Availability and Latency
Operational Mode	Model training is offline	Usually online, requires real time response
Infrastructure requirement	Large network with many GPU/CPU hosts	Smaller network with mid size of CPU/GPU hosts

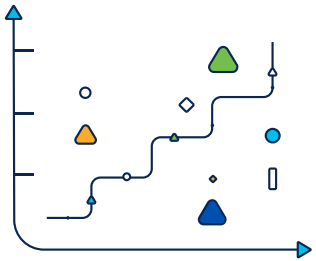
Training vs Inference





Large Scale Distributed Training

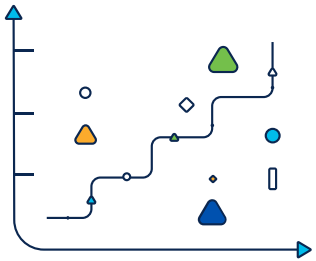
- Key Challenge of Training Cluster
  - Model Doubles every 2 mounts
  - Bigger model, higher accuracy
  - Most common single training runs on 512 GPUs



- Cluster Key Components
  - Compute Nodes
  - Network
  - Distributed File System/Storage
  - Job Scheduling and Orchestration
  - Software Framework for AI model

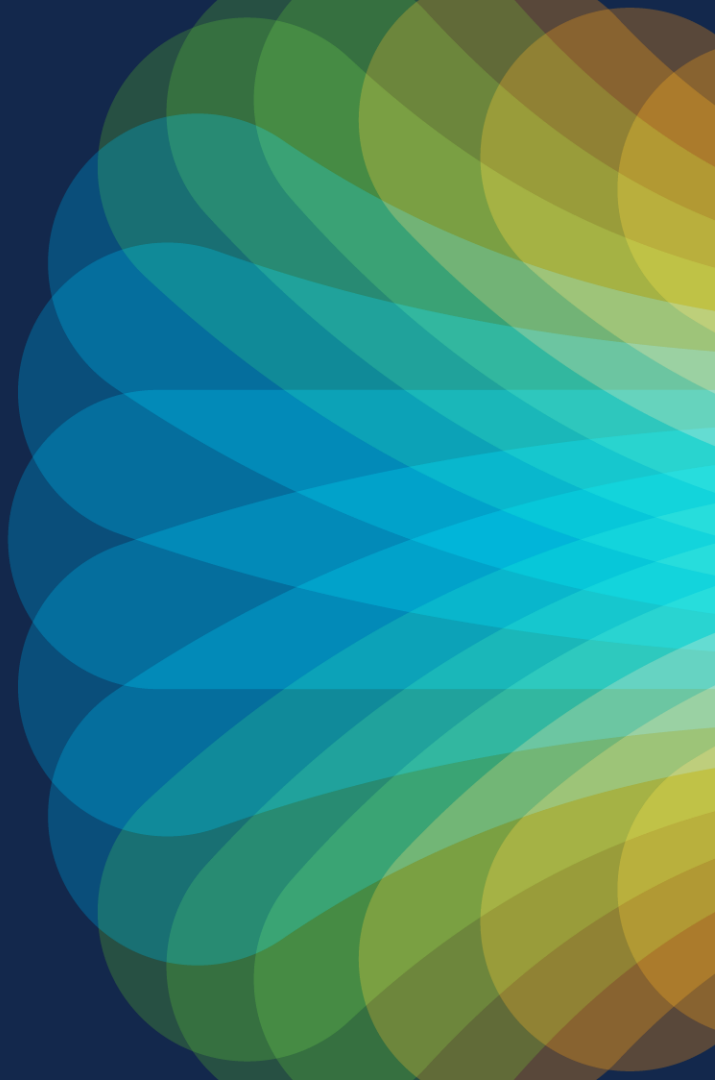
Large Scale Distributed Training

- Key Challenge of Training Cluster
  - Model Doubles every 2 mounts
  - Bigger model, higher accuracy
  - Most common single training runs on 512 GPUs



- Cluster Key Components
  - Compute Nodes
  - Network
  - Distributed File System/Storage
  - Job Scheduling and Orchestration
  - Software Framework for AI model

Network For AI Cluster



AI Training Network

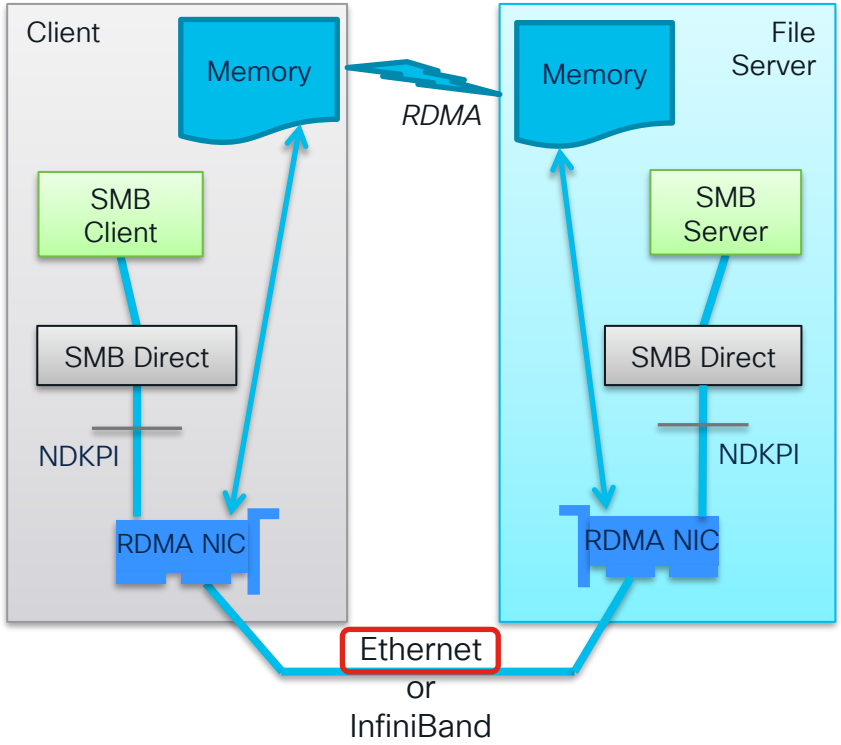
For RoCEv2 transport the network must provide high throughput and low latency by passing CPU/GPU

The Cisco Nexus 9000 switches, provide the required low latency, and with up to 25.6Tbps of bandwidth per ASIC, to satisfy AI/ML clusters requirements

Shipping feature set with many customers in production

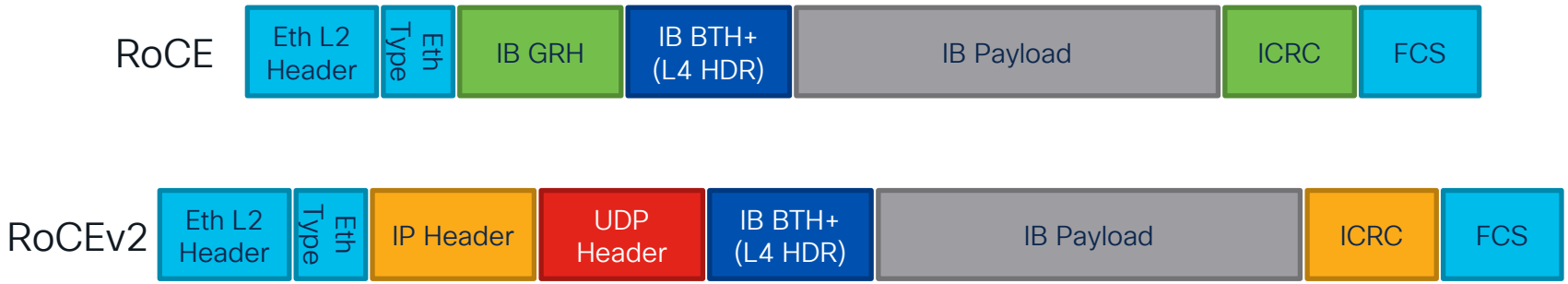
RDMA – Remote Data Memory Access

- Allows application software to communicate directly with the hardware (RDMA NIC)
- Bypasses OS stack
- RDMA delivers, low latency, high throughput, zero copy capabilities
- RDMA Hardware Technologies
  - RoCE: RDMA over Converged Ethernet
  - iWARP: RDMA over TCP/IP
  - Infiniband



RoCEv2 – Basics

- Extension of RoCE protocol that involves a simple modification of the RoCE packet format
- Carry IP header and UDP header that serves as a stateless encapsulation layer for RDMA transport over IP

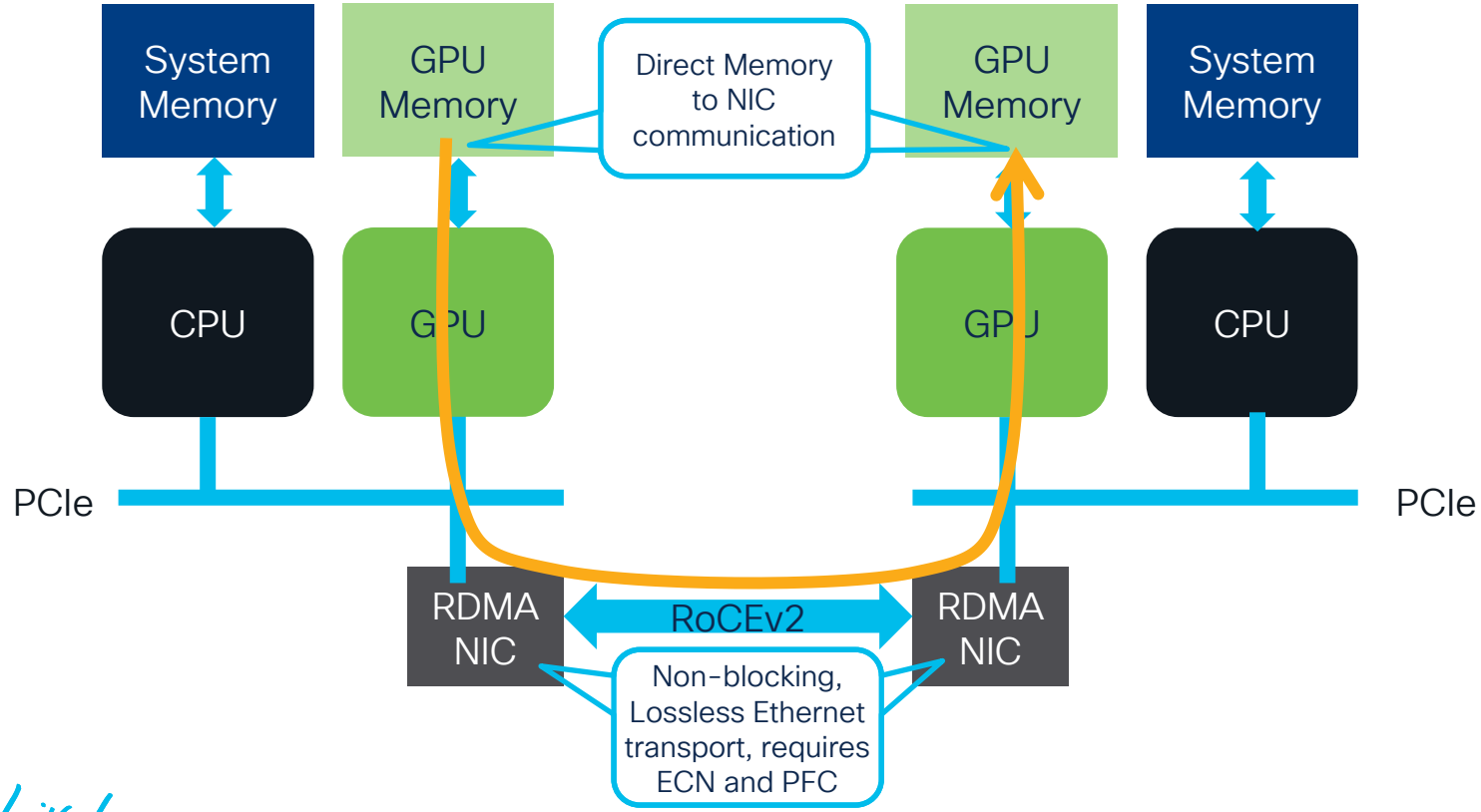


Source: https://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet

RoCEv2

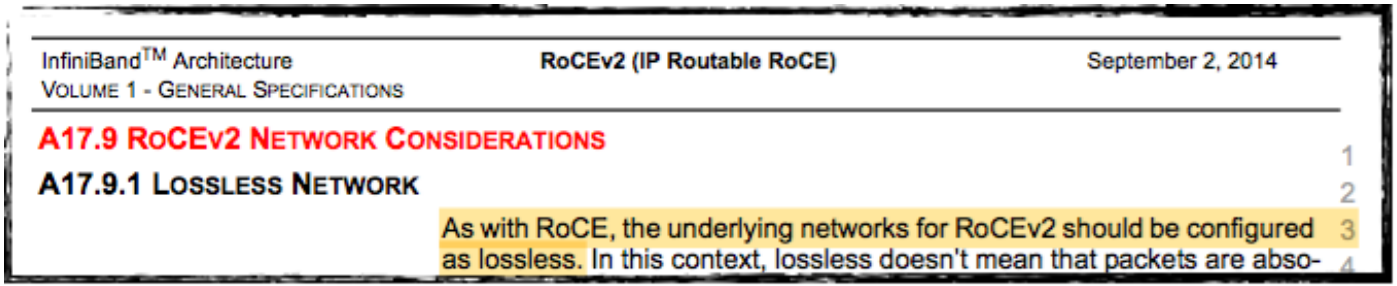
- Uses well-known UDP Destination Port (dport) value 4791
- UDP Source Port (sport) serves as opaque flow identifier that can be used by networking infrastructure for packet forwarding optimizations (e.g., ECMP)
- Supports both IPv4 and IPv6
- Makes use of ECN field in IPv4/6 header for signaling of congestion

RoCEv2 – Benefits and Requirements



RoCEv2 End-To-End Lossless Behavior

- Requires PFC to be enabled for RoCEv2 transport
- Traffic priority to be preserved between Layer 2 and Layer 3 network
  - Packet/Flow identification follows standard practices of IP/Ethernet networks (i.e., DSCP/802.1Q)
- ECN marking (WRED or consider AFD)
- Configure ETS
  - 802.1Qaz ETS



InfiniBand™ Architecture
VOLUME 1 - GENERAL SPECIFICATIONS

RoCEv2 (IP Routable RoCE)

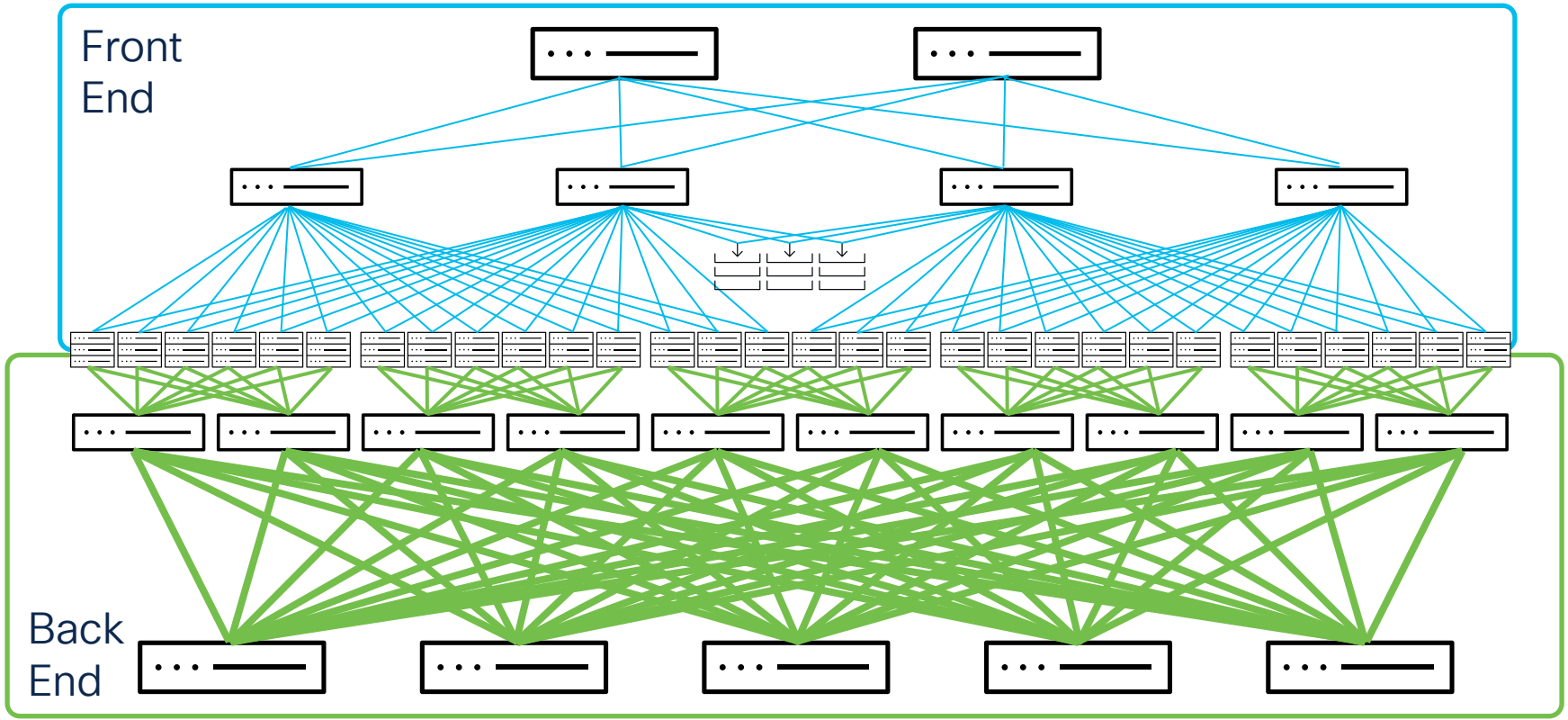
September 2, 2014

A17.9 RoCEv2 Network Considerations

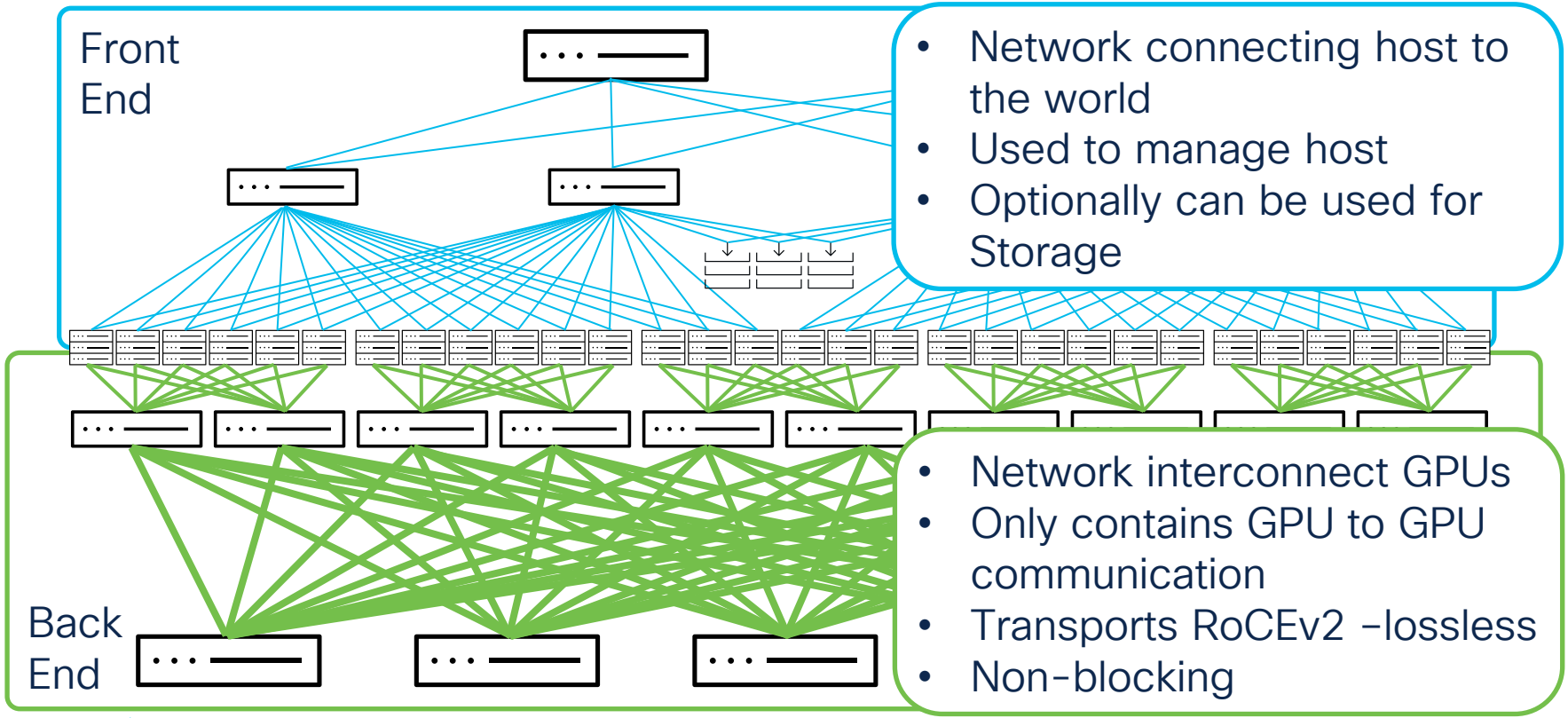
A17.9.1 Lossless Network

As with RoCE, the underlying networks for RoCEv2 should be configured as lossless. In this context, lossless doesn't mean that packets are abso-

Non-blocking Network



Non-blocking Network

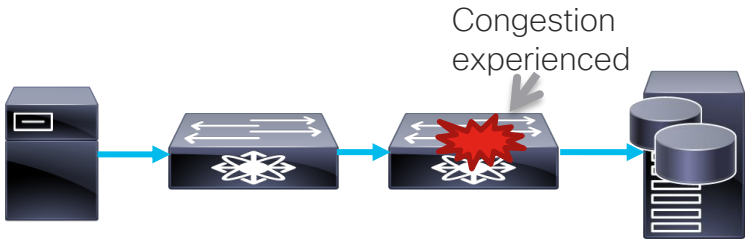


- Network connecting host to the world
- Used to manage host
- Optionally can be used for Storage

- Network interconnect GPUs
- Only contains GPU to GPU communication
- Transports RoCEv2 –lossless
- Non-blocking

Explicit Congestion Notification (ECN)

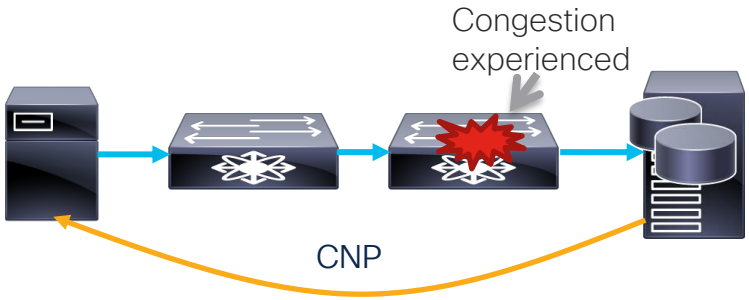
- IP Explicit Congestion Notification (ECN) is used for congestion notification.
- ECN enables end-to-end congestion notification between two endpoints on IP network
- ECN uses 2 LSB of Type of Service field in IP header



ECN	ECN Behavior
00	Non ECN Capable
10	ECN Capable Transport (0)
01	ECN Capable Transport (1)
11	Congestion Encountered

Explicit Congestion Notification (ECN)

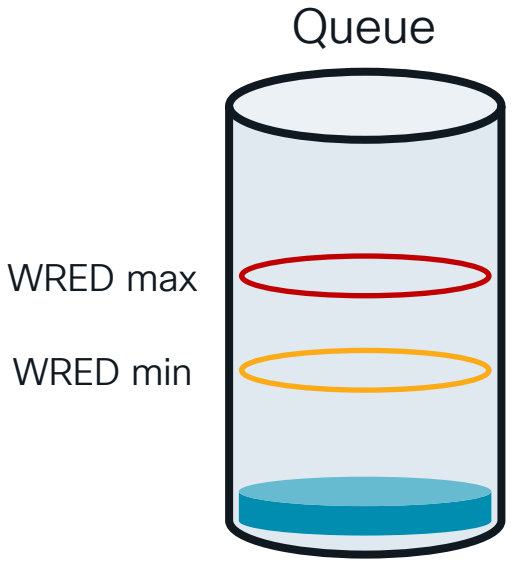
- IP Explicit Congestion Notification (ECN) is used for congestion notification.
- ECN enables end-to-end congestion notification between two endpoints on IP network
- ECN uses 2 LSB of Type of Service field in IP header
- In case of congestion, ECN gets transmitting device to reduce transmission rate using Congestion Notification Packet (CNP) without pausing traffic.



ECN	ECN Behavior
00	Non ECN Capable
10	ECN Capable Transport (0)
01	ECN Capable Transport (1)
11	Congestion Encountered

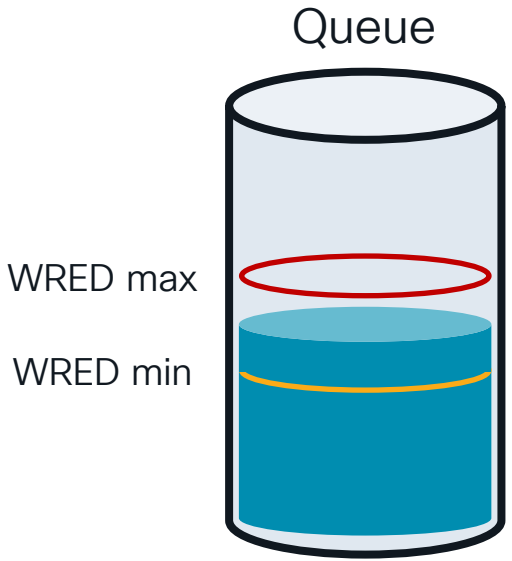
How does WRED ECN work?

- WRED (Weighted Random Early Detection) is used to signalize severity of congestion
- ECN is not marked when buffer usage is below WRED min threshold



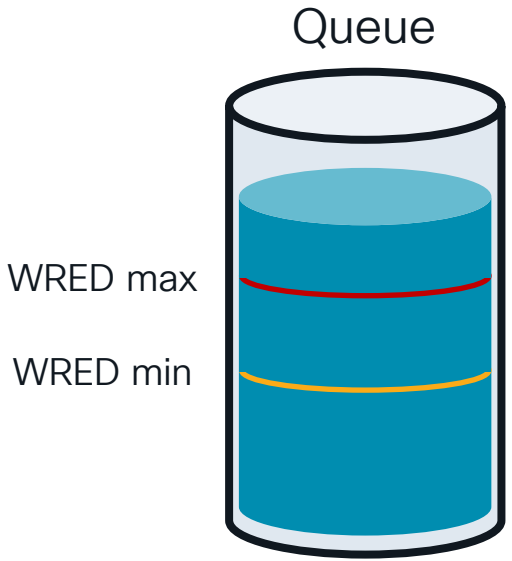
How does WRED ECN work?

- 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)

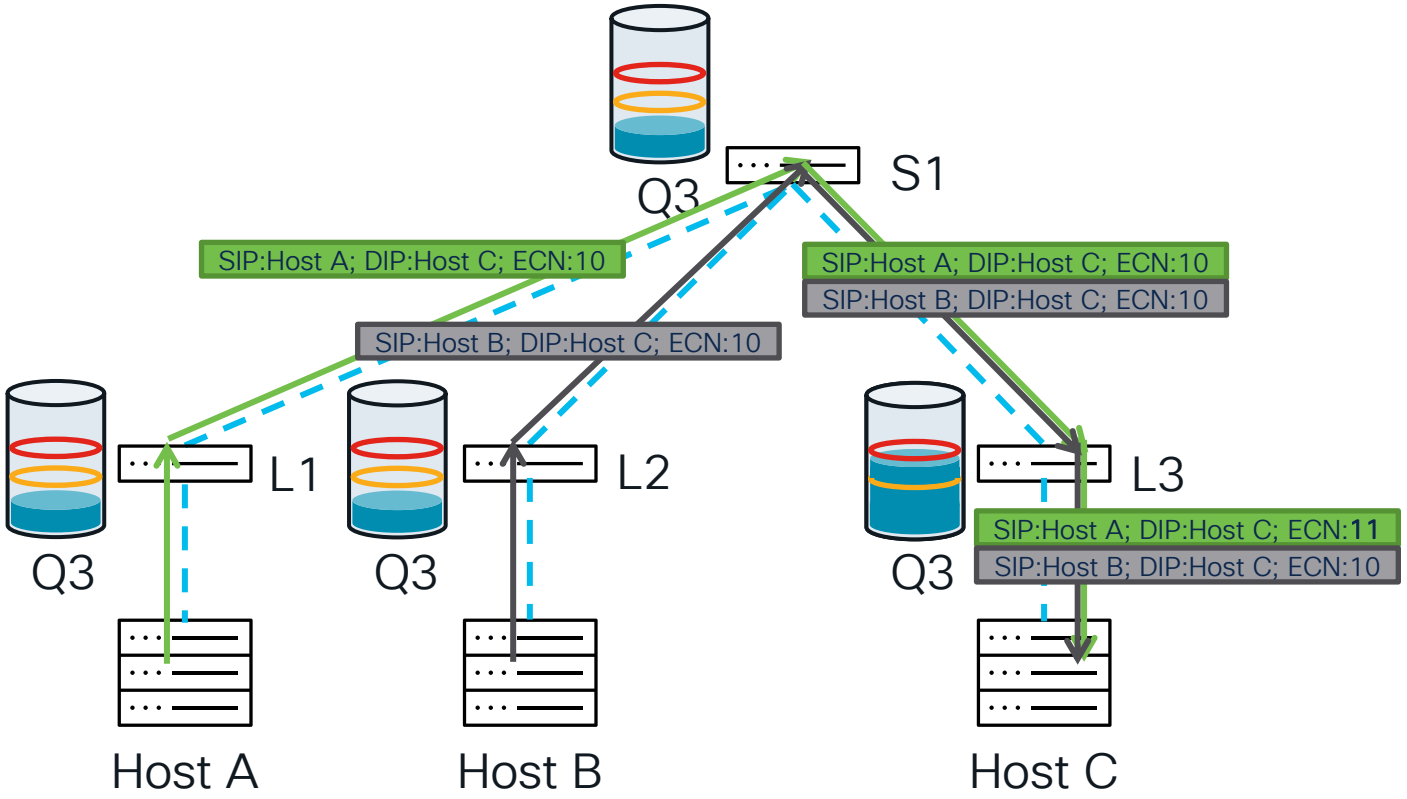


How does WRED ECN work?

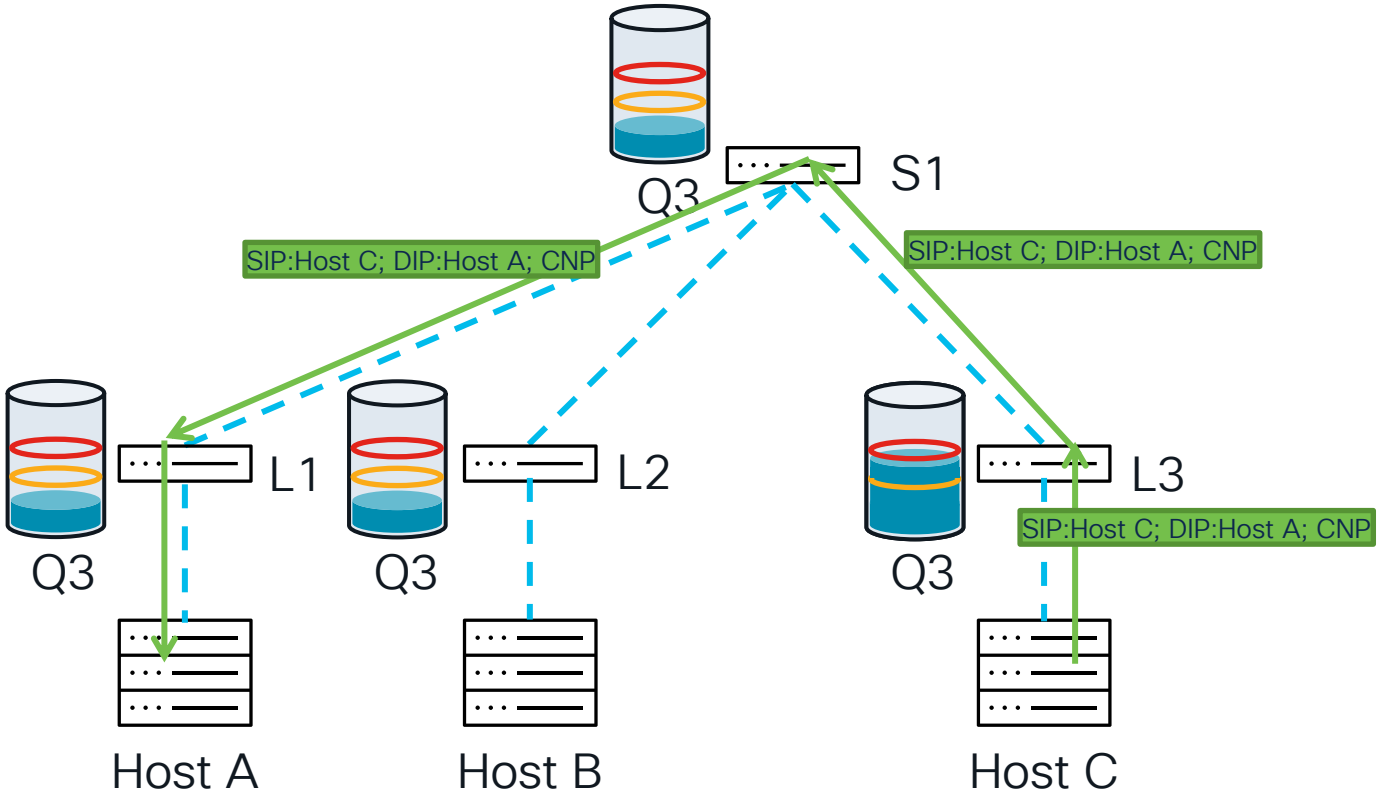
- WRED (Weighted Random Early Detection) is used to signalize severity of congestion
- ECN is not marked when buffer usage is below WRED min threshold
- When buffer usage is minimal threshold, Congestion Encountered will be marked on N number of randomly selected packets (probability parameter)
- After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



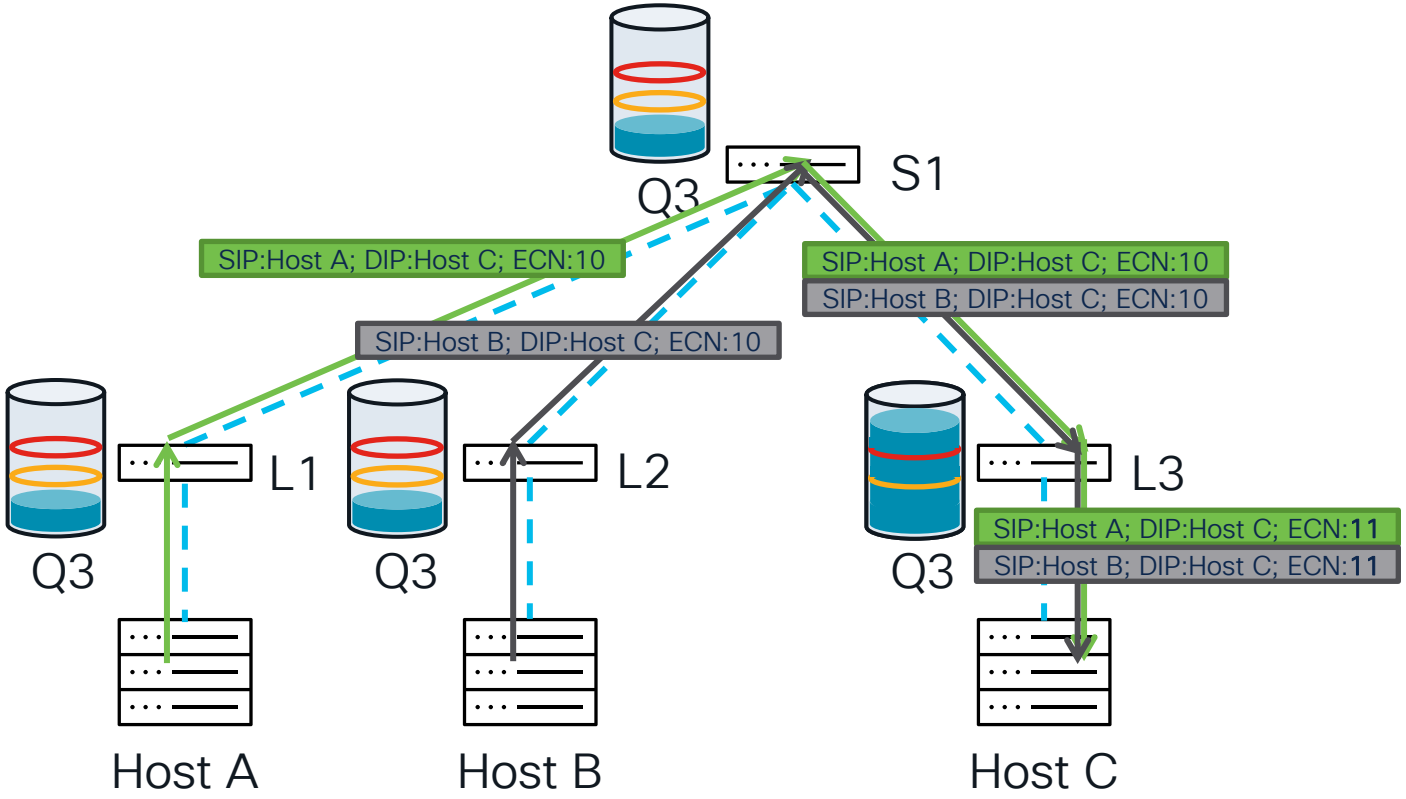
Explicit Congestion Notification – End to End



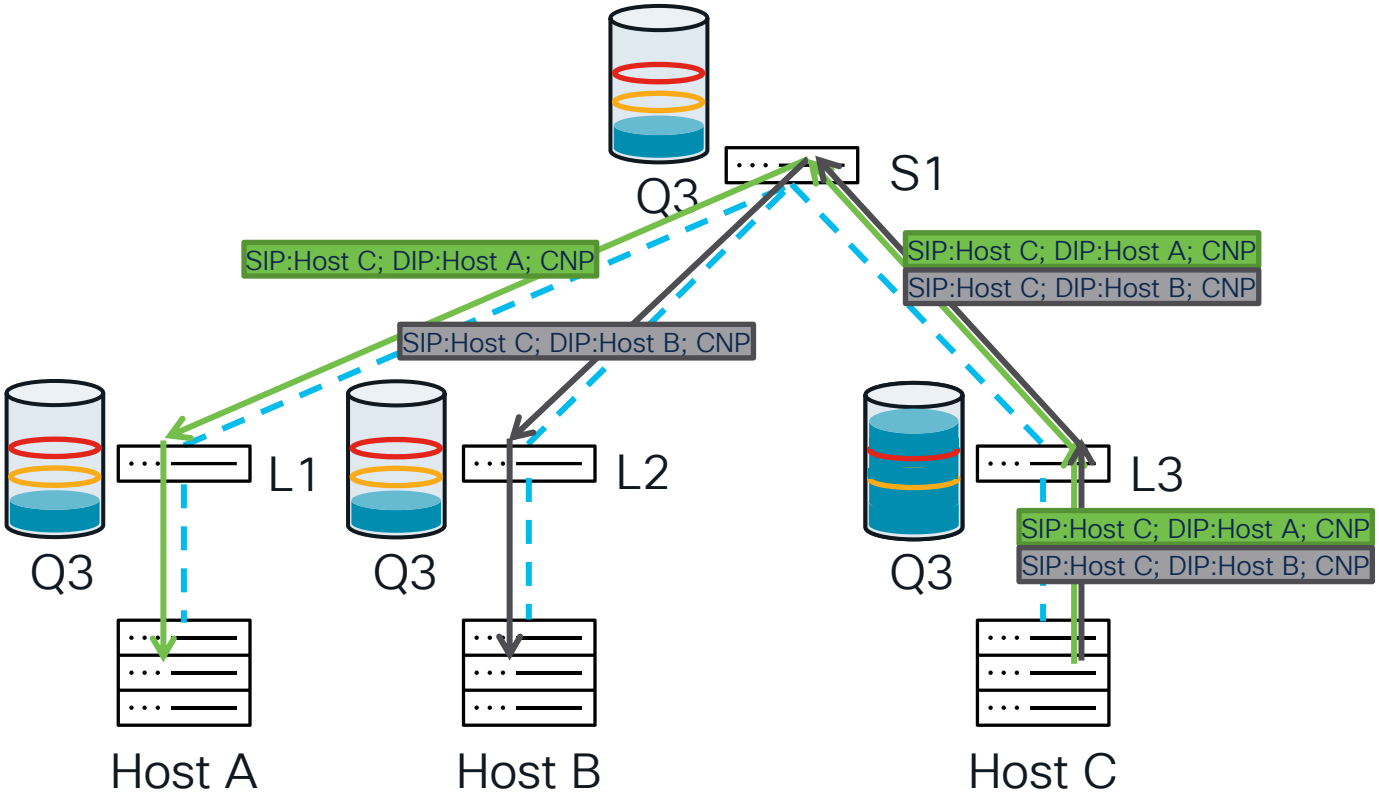
Explicit Congestion Notification – End to End



Explicit Congestion Notification – End to End

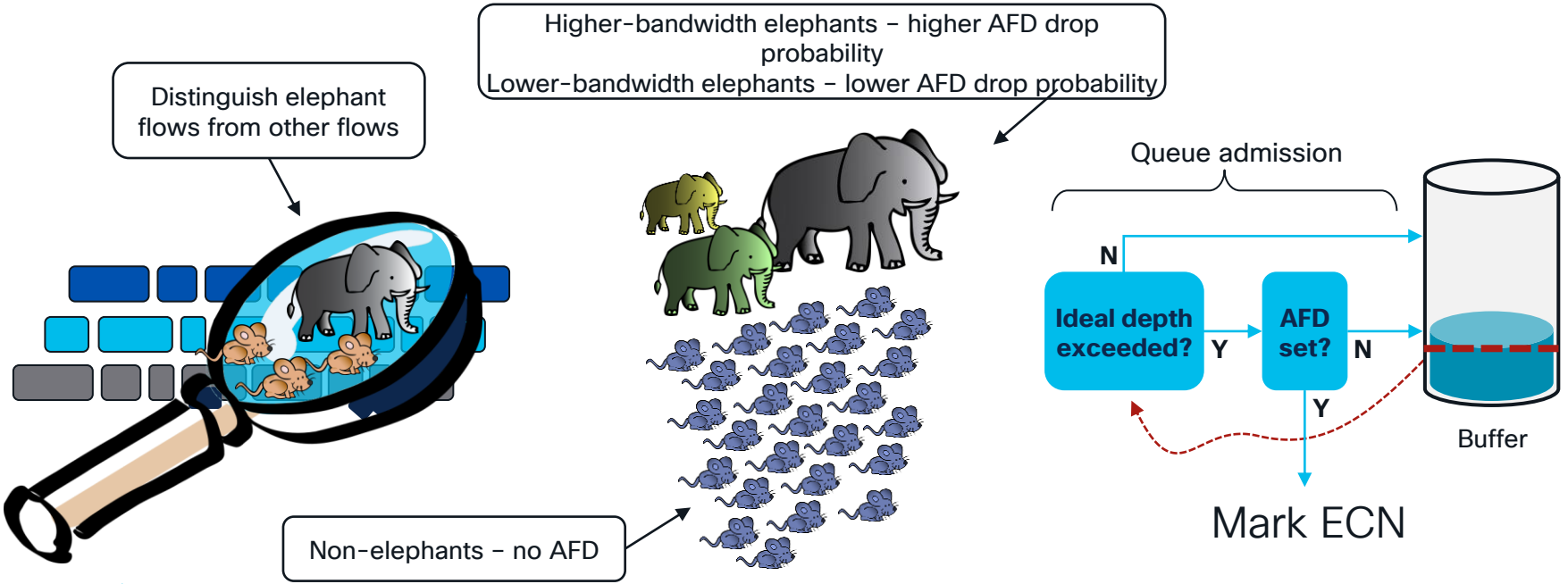


Explicit Congestion Notification – End to End



Approximate Fair Drop (AFD)

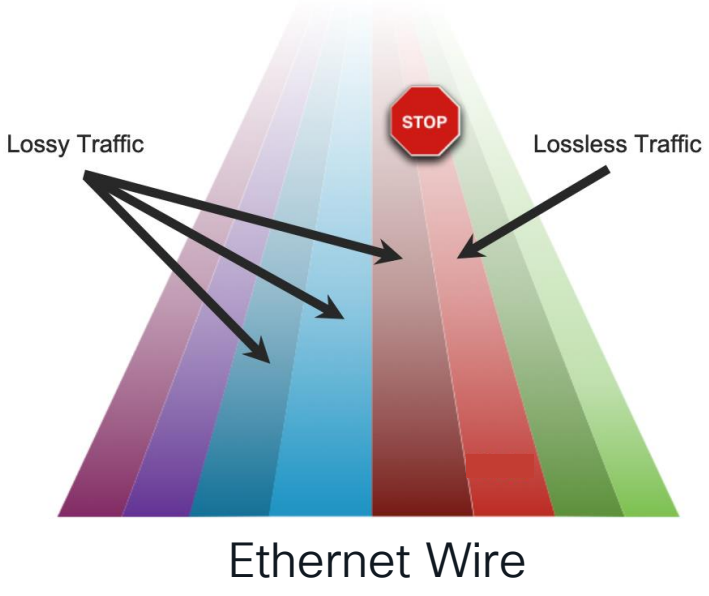
Maintain throughput while minimizing buffer consumption by elephant flows – **keep buffer state as close to the ideal as possible**



Priority Flow Control

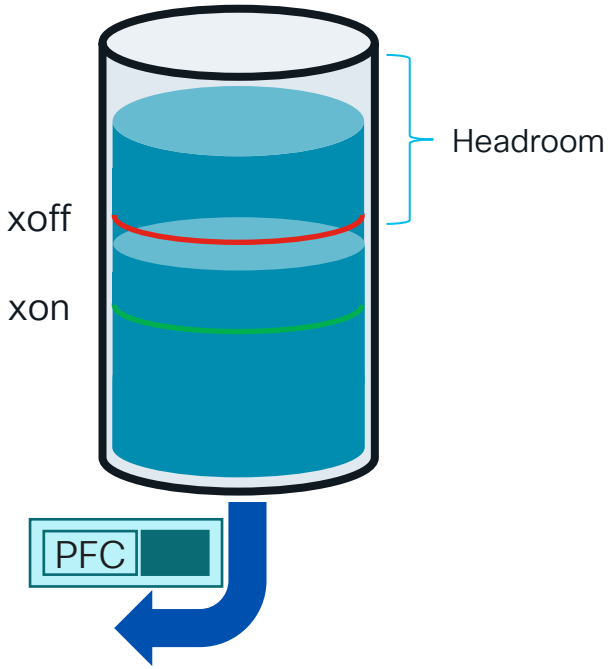
Flow Control Mechanism – 802.1Qbb

- A.k.a "Lossless Ethernet”
- PFC enables Flow Control on a Per-Priority basis
- PFC is also called Per-Priority-Pause
- Therefore, we have the ability to have lossless and lossy priorities at the same time on the same wire
- Allows traffic to operate over a lossless priority independent of other priorities
- Other traffic assigned to other priority will continue to transmit and rely on upper layer protocols for retransmission

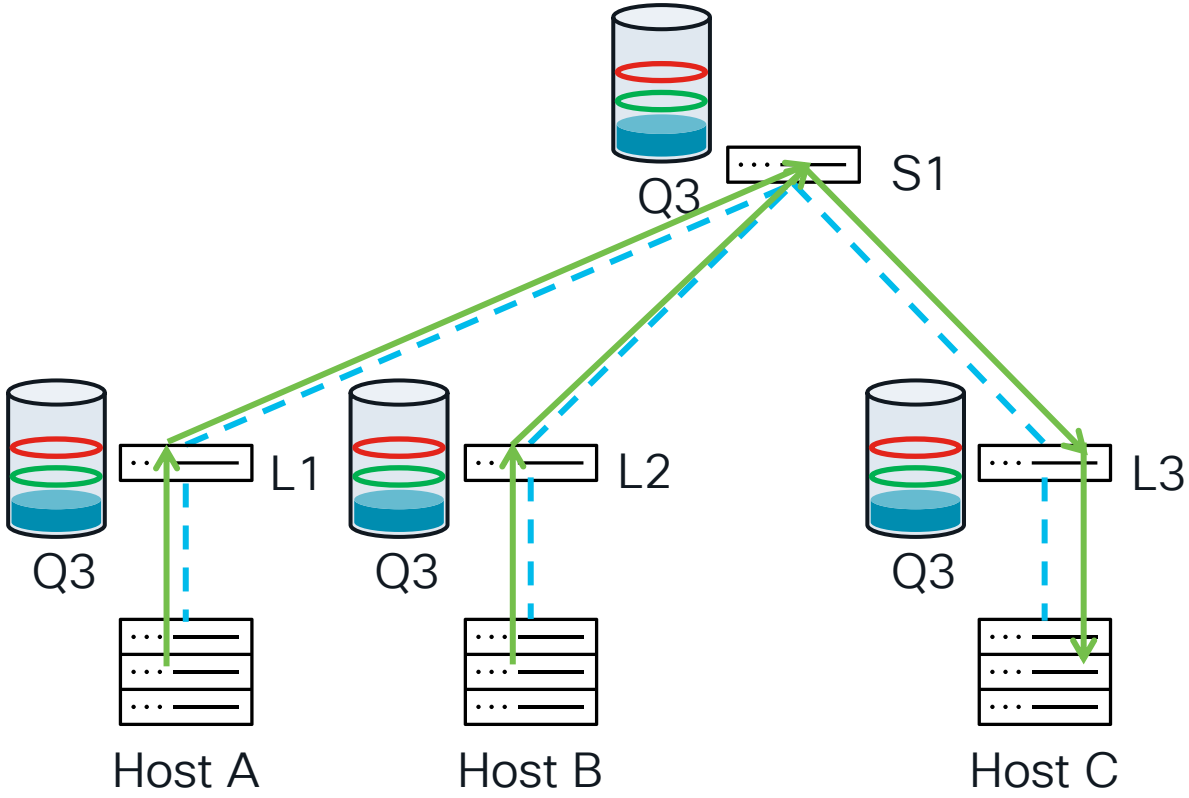


PFC – How pause frames are sent

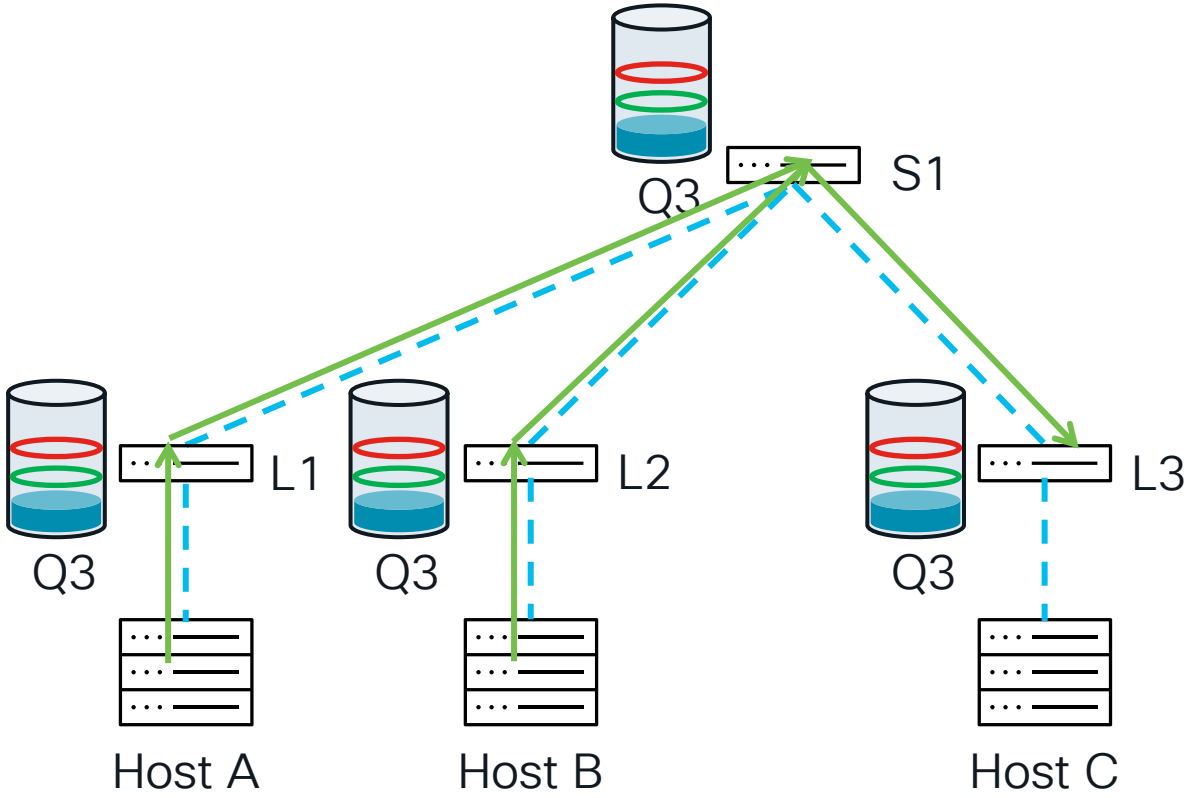
- PFC sets thresholds in no-drop queue
- Headroom is present to accommodate “in flight” packets
- Under congestion, traffic is buffered in non-drop queue
- PFC frames are sent toward sender after queue utilization exceeds *xoff* threshold
- While draining the queue, and utilization is below *xon* threshold system will stop sending PFC frames



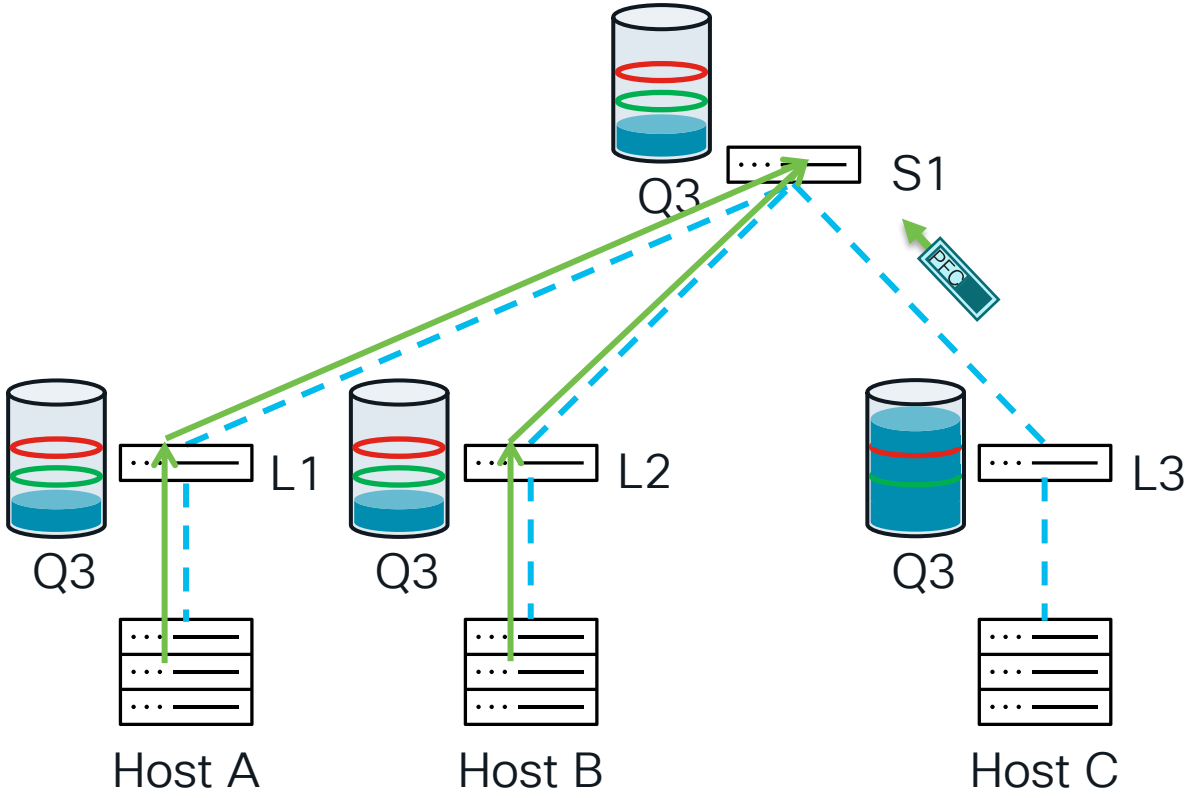
PFC Hop by hop



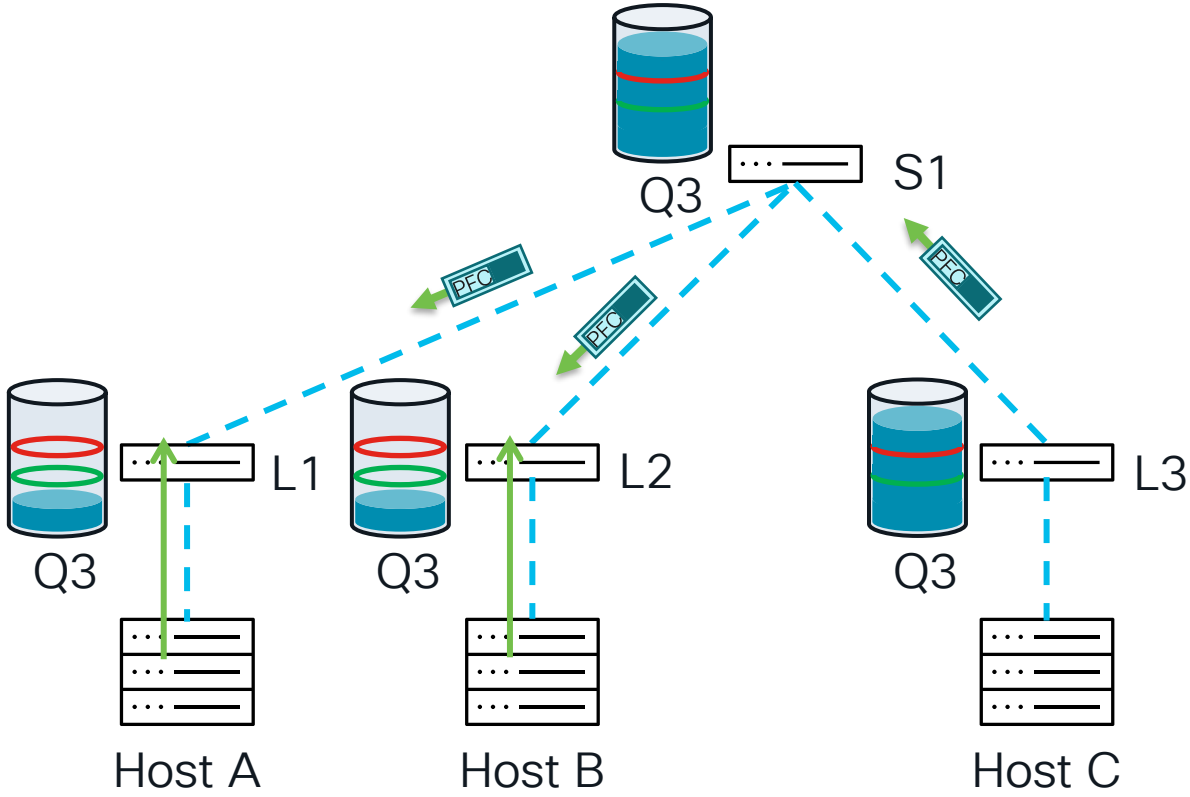
PFC Hop by hop



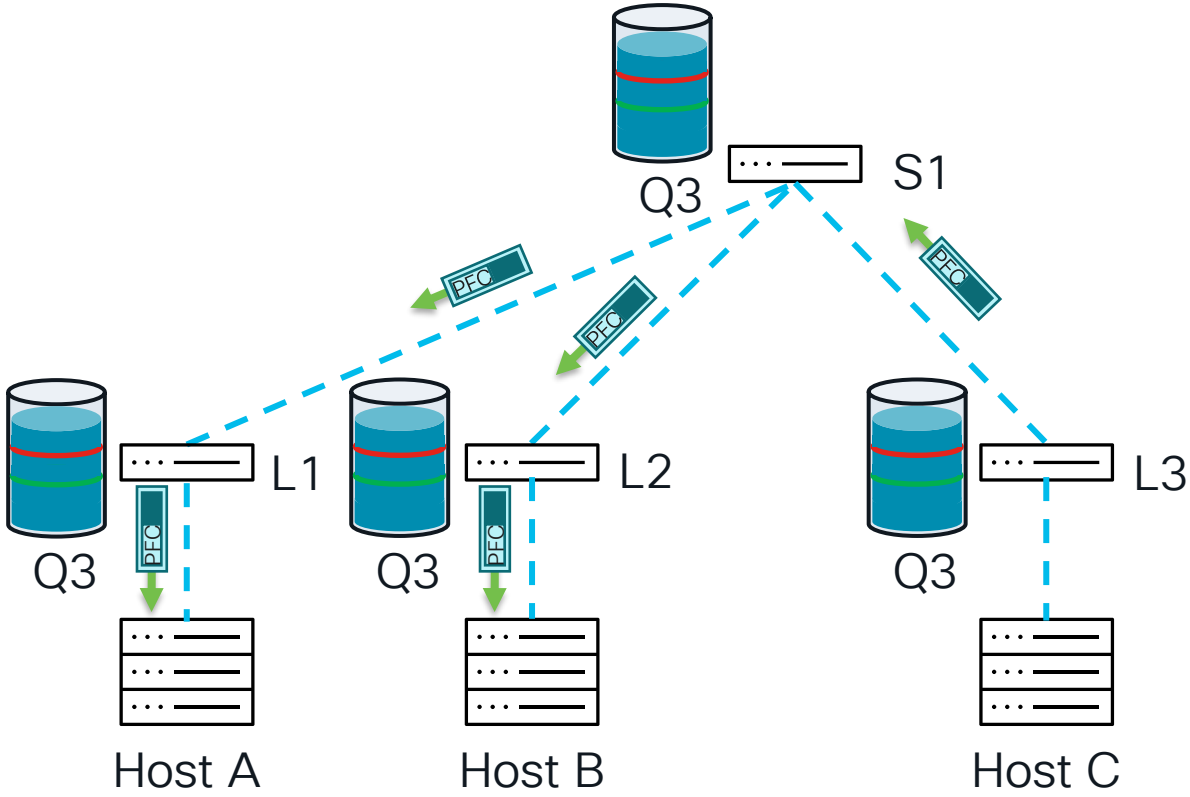
PFC Hop by hop



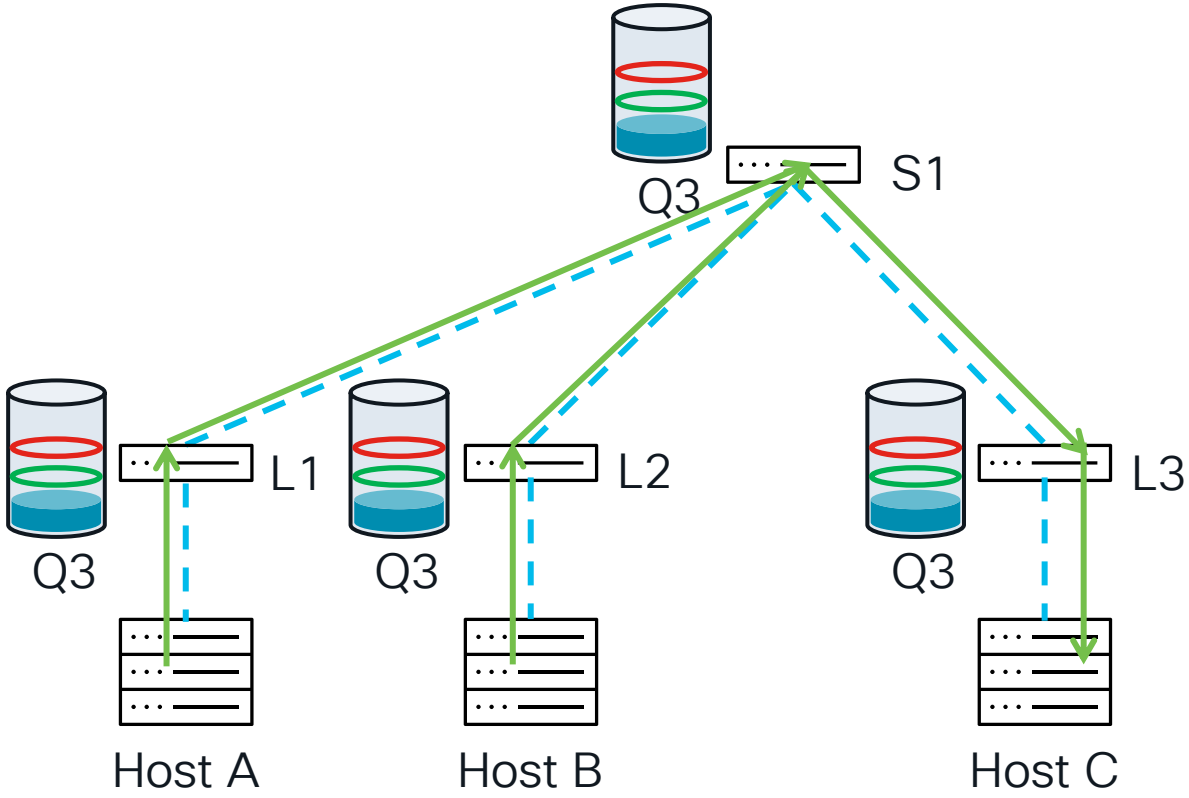
PFC Hop by hop



PFC Hop by hop

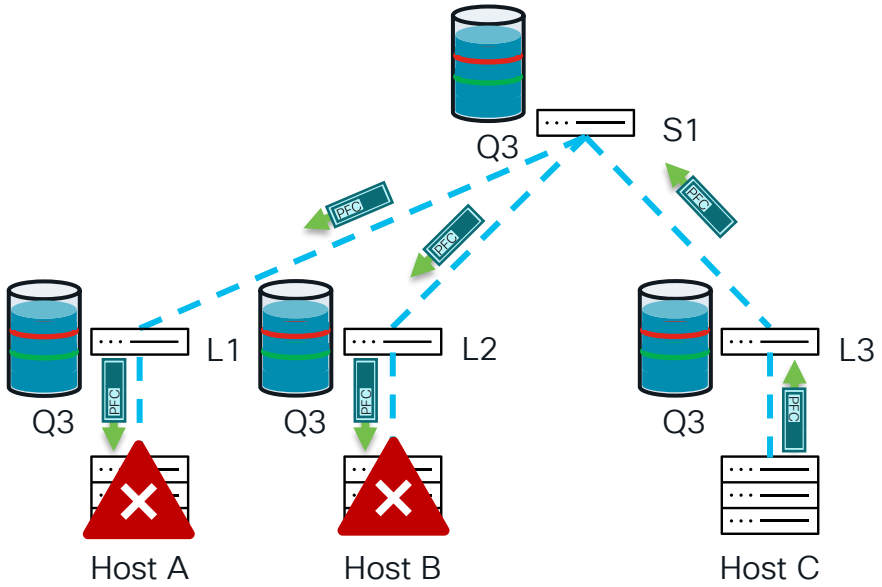


PFC Hop by hop



NIC PFC Storm

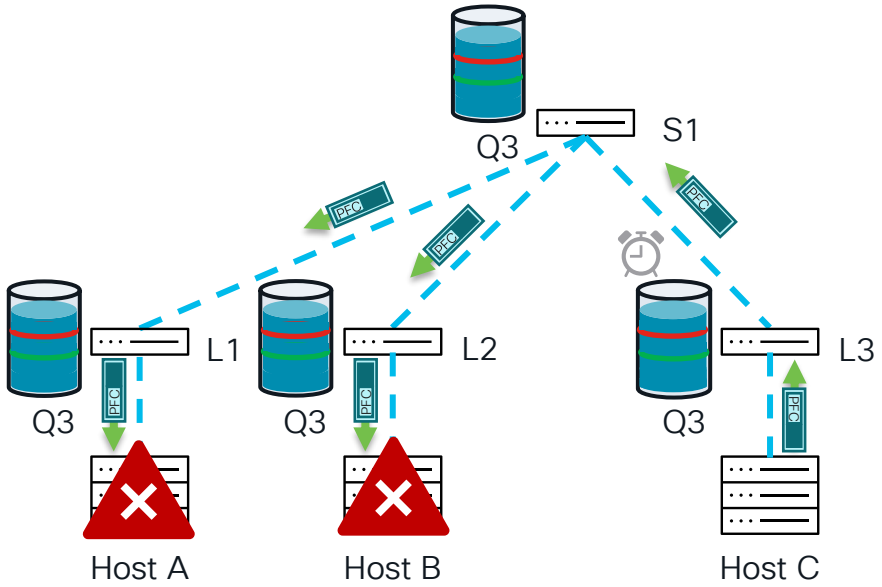
- In occasion of NIC malfunction, PFC storm can be triggered to send continues PFC frames in the network
- Network will propagate those frame to all senders
- PFC storm will stop traffic coming from sender
- PFC watchdog can drain the queue



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/rdma_sigcomm2016.pdf

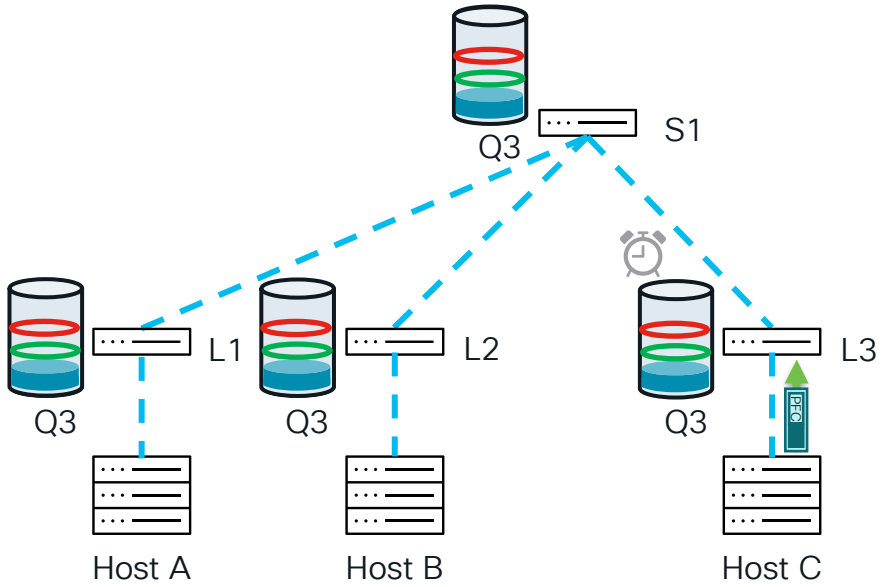
PFC Watchdog

- PFC watchdog sets a timeout, if a packet exceeds time out, all packets from a queue will be cleared
- The watchdog prevents PFC frames propagating to sender and blocking it
- PFC Watchdog is supported on Cisco Nexus 9000 switches



PFC Watchdog

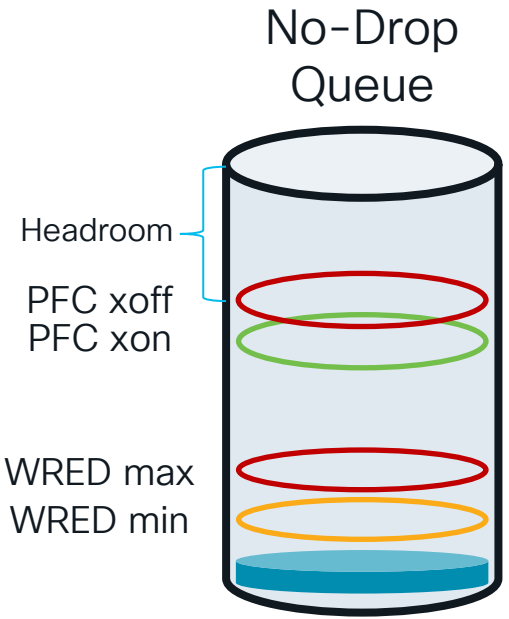
- PFC watchdog sets a timeout, if a packet exceeds time out, all packets from a queue will be cleared
- The watchdog prevents PFC frames propagating to sender and blocking it
- PFC Watchdog is supported on Cisco Nexus 9000 switches



RoCEv2: PFC and ECN together

How does it work?

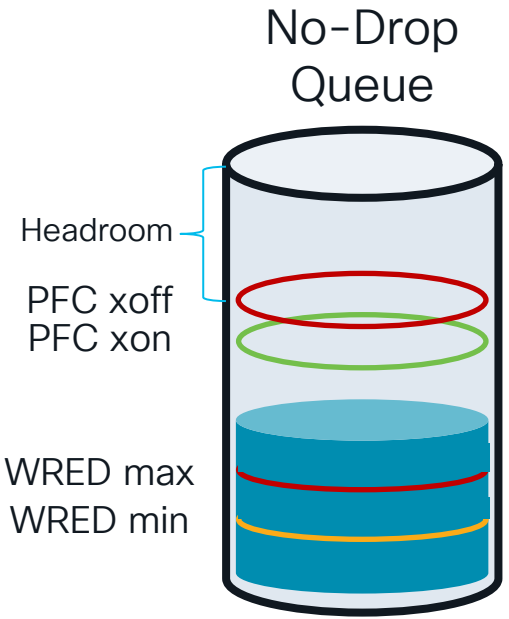
- WRED threshold are set low in no-drop queue
  - Signalize early for congestion, give enough time for end points to react
- PFC threshold are set higher than ECN
  - In case oversubscription buffers can be filled quickly without giving time to ECN to react
  - PFC will react and mitigate congestion



RoCEv2: PFC and ECN together

How does it work?

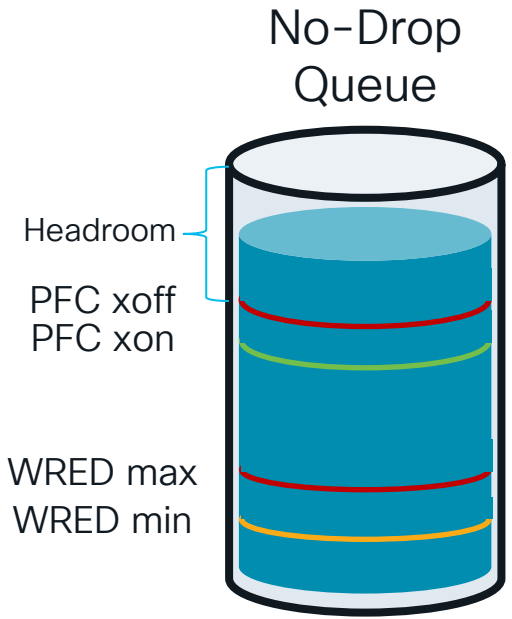
- WRED threshold are set low in no-drop queue
  - Signalize early for congestion, give enough time for end points to react
- PFC threshold are set higher than ECN
  - In case oversubscription buffers can be filled quickly without giving time to ECN to react
  - PFC will react and mitigate congestion



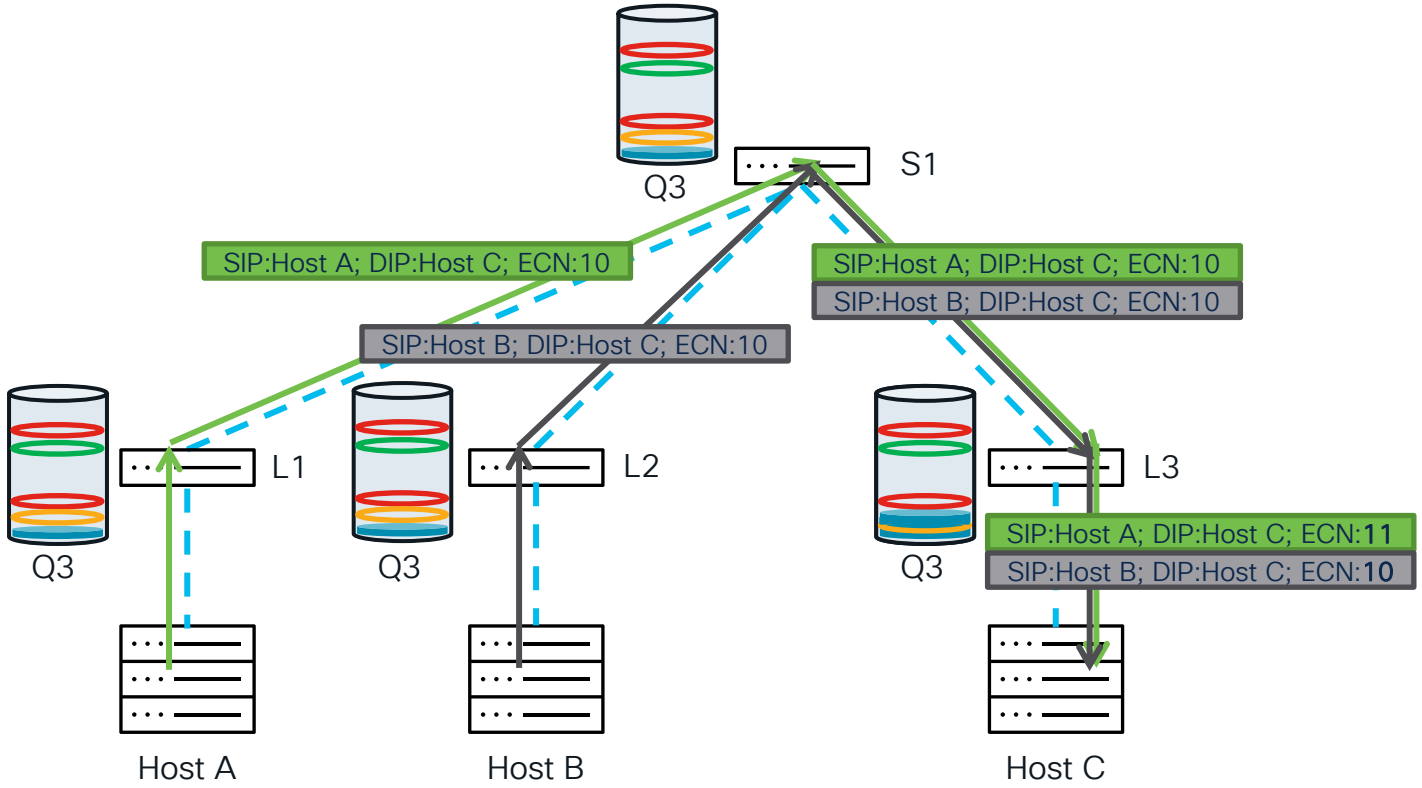
RoCEv2: PFC and ECN together

How does it work?

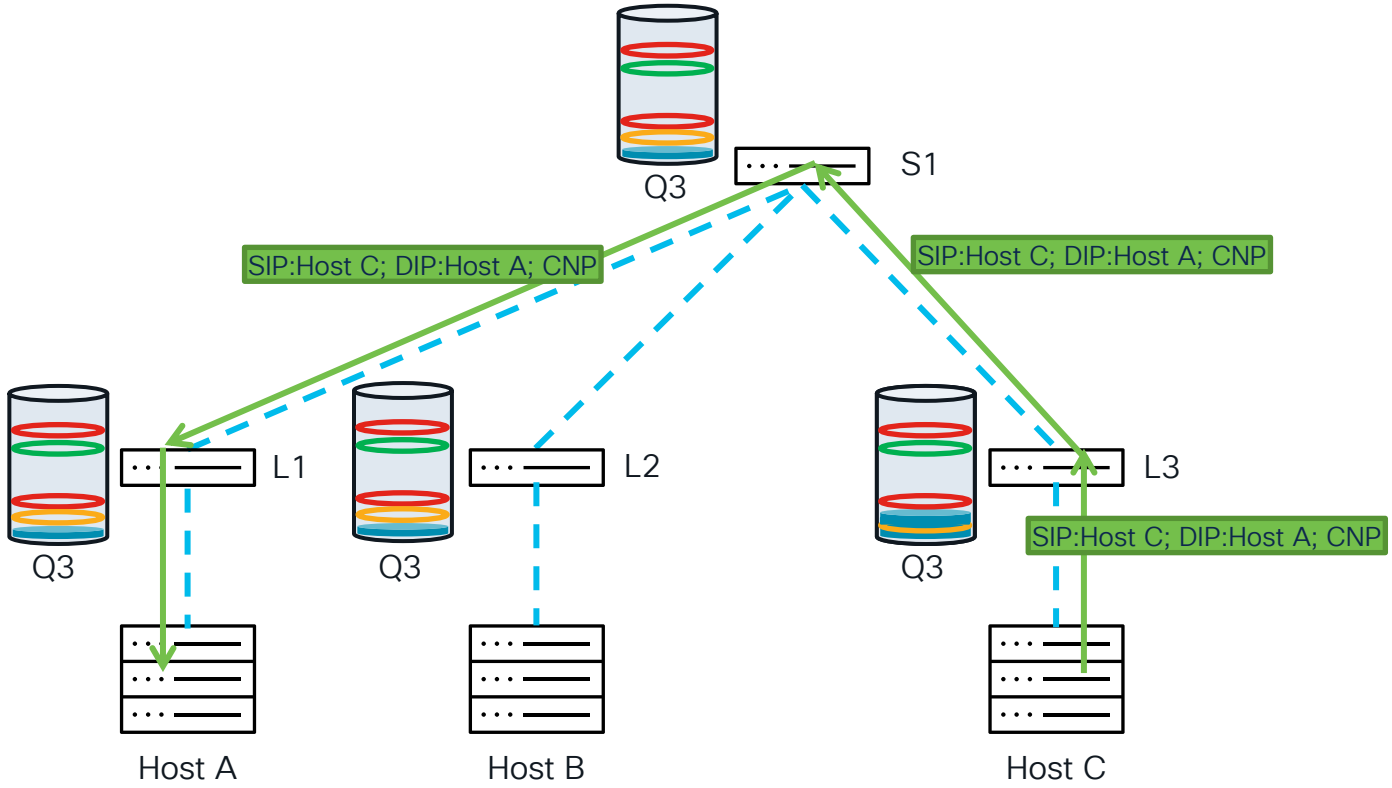
- WRED threshold are set low in no-drop queue
  - Signalize early for congestion, give enough time for end points to react
- PFC threshold are set higher than ECN
  - In case oversubscription buffers can be filled quickly without giving time to ECN to react
  - PFC will react and mitigate congestion



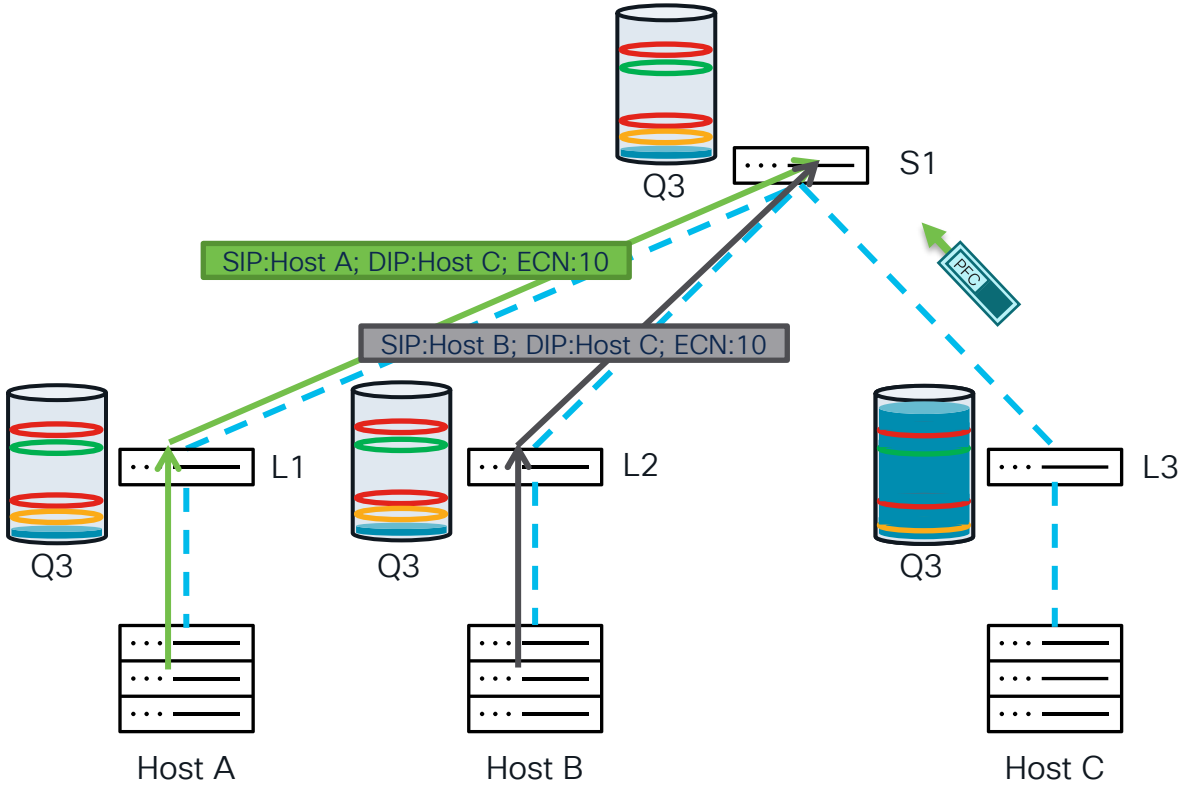
ECN and PFC make DCQCN



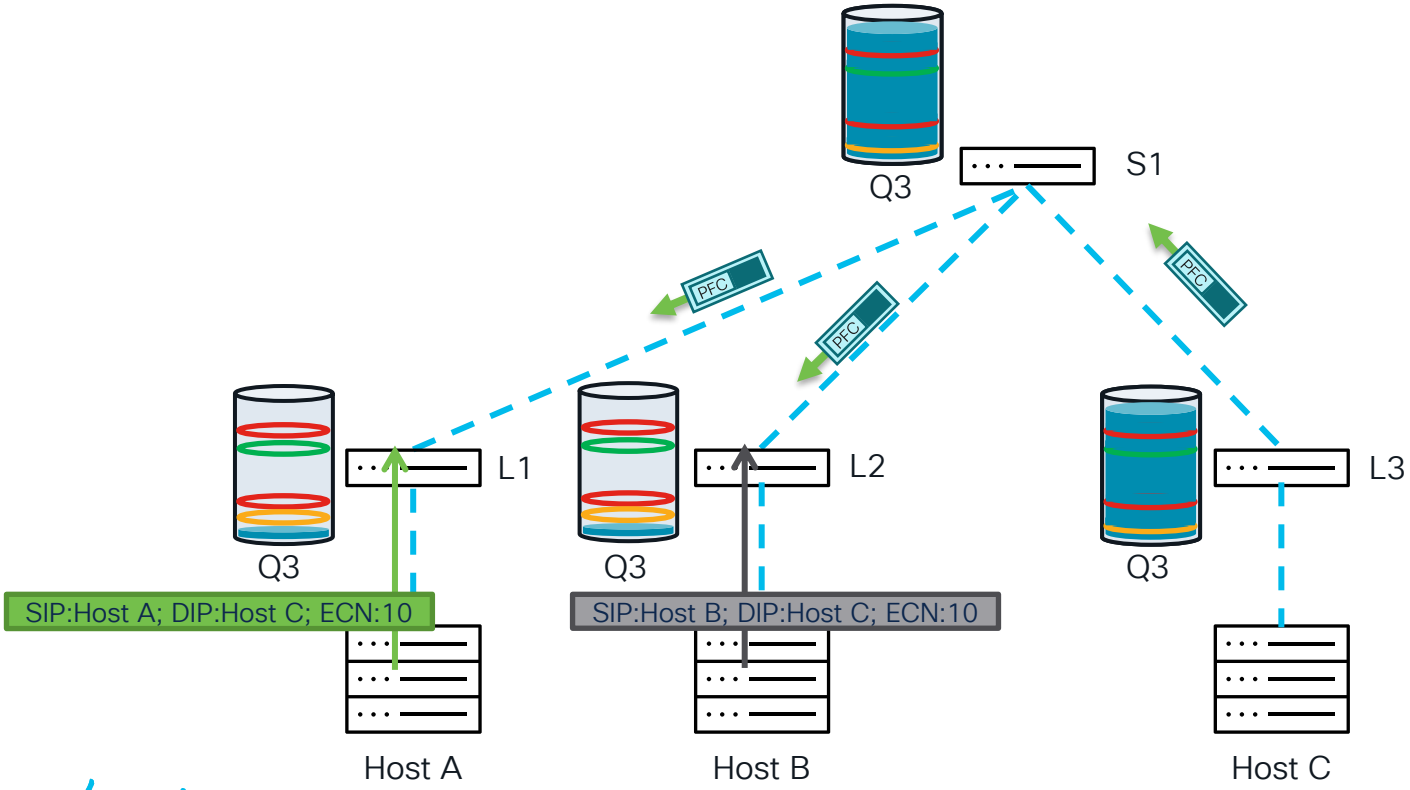
ECN and PFC make DCQCN



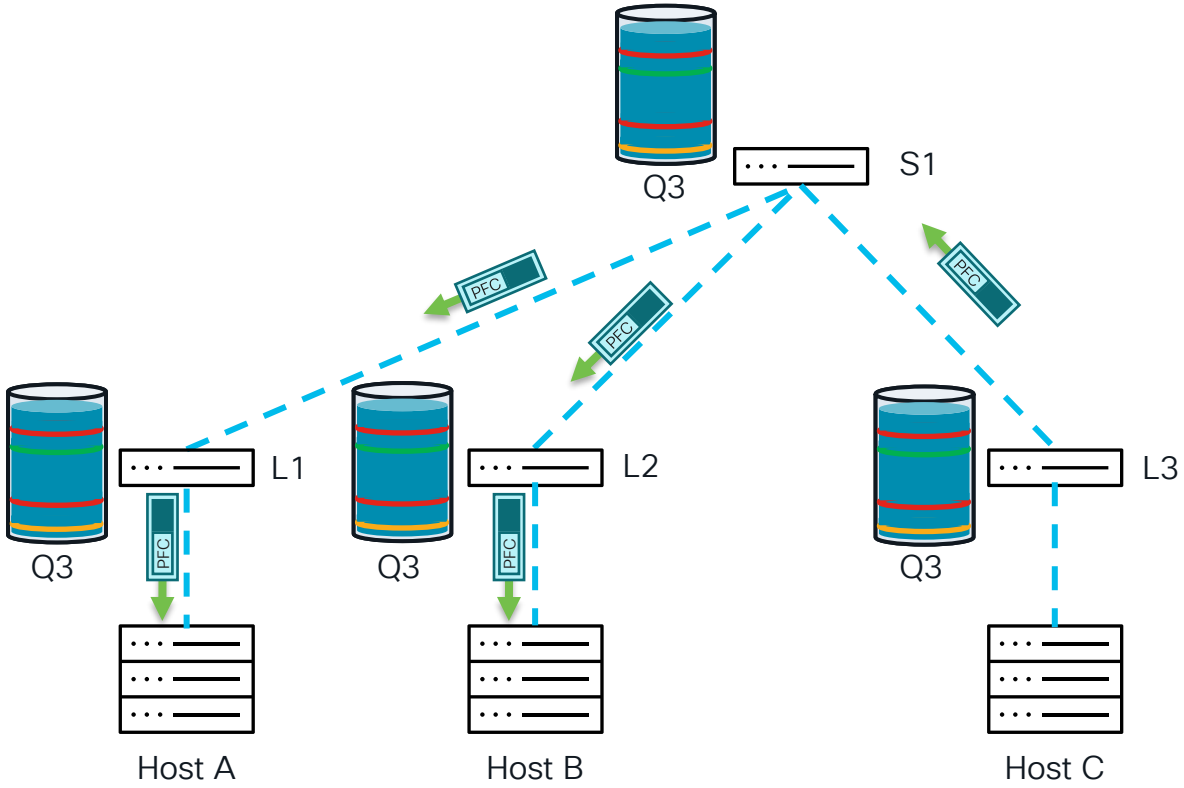
ECN and PFC make DCQCN



ECN and PFC make DCQCN



ECN and PFC make DCQCN



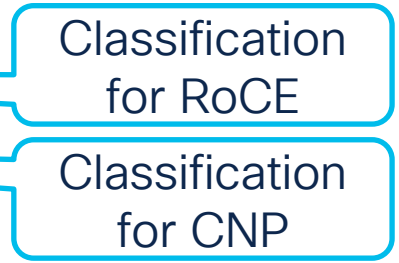
Quality of Service

- Required separate queue for RoCEv2 traffic
  - Distinguished from other traffic in the port
  - Provide dedicated scheduling resources, to reduce latency
  - No contention for buffer resources with other traffic
  - RoCEv2 is not be part of strict priority queue, high volume of it might affect control plane
- RoCEv2 traffic requires ECN, PFC on the queue, while other traffic does not
- CNP traffic is part of strict priority queue, to deliver congestion signaling in time

Quality of Service- Configuration

```
class-map type qos match-all class-roce
  match dscp 24
class-map type qos match-all class-cnp
  match dscp 48

policy-map type qos QOS_classification_policy
  class class-roce
    set qos-group 3
  class class-cnp
    set qos-group 7
  class class-default
    set qos-group 0
```



Quality of Service- Configuration

```
policy-map type queuing custom-8q-out-policy
  class type queuing c-out-8q-q7
    priority level 1
<snip>
  class type queuing c-out-8q-q3
    bandwidth remaining percent 99
    random-detect minimum-threshold 150 kbytes maximum-threshold 3000 kbytes drop-probability 7
weight 0 ecn
<snip>
  class type queuing c-out-8q-q-default
    bandwidth remaining percent 1

policy-map type network-qos custom-8q-nq-policy
<snip>
  class type network-qos c-8q-nq3
    mtu 9216
    pause pfc-cos 3
<snip>
```

WRED min

WRED max

WRED random marked packets

Enable PFC

Quality of Service- Configuration

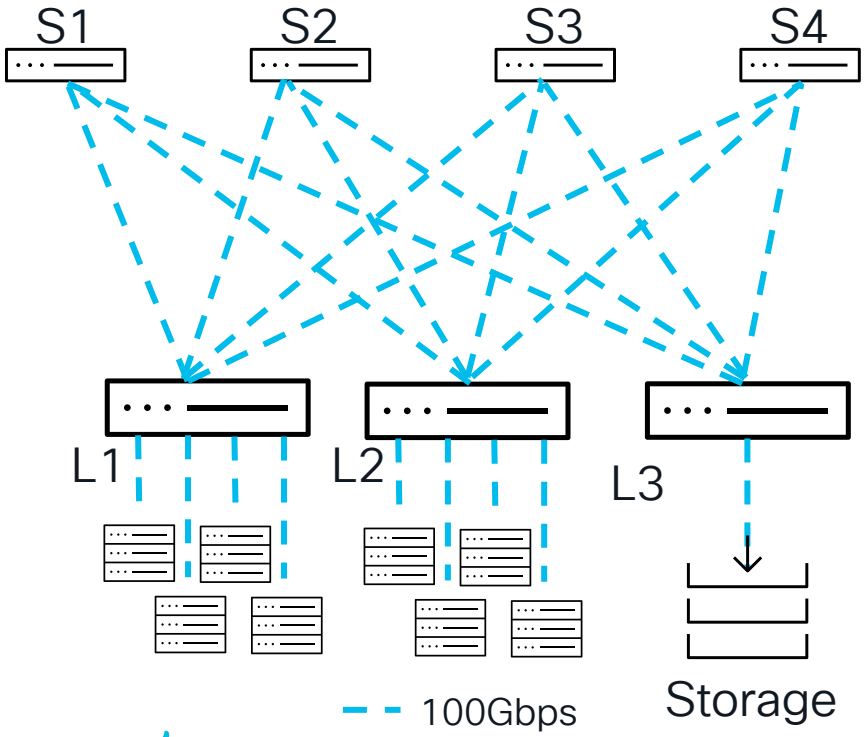
```
system qos
  service-policy type network-qos custom-8q-nq-policy
  service-policy type queuing output custom-8q-out-policy

interface Ethernet1/1
  service-policy type qos input QOS_classification_policy
  priority-flow-control mode on
  priority-flow-control watch-dog-interval on
```

Enable WRED ECN

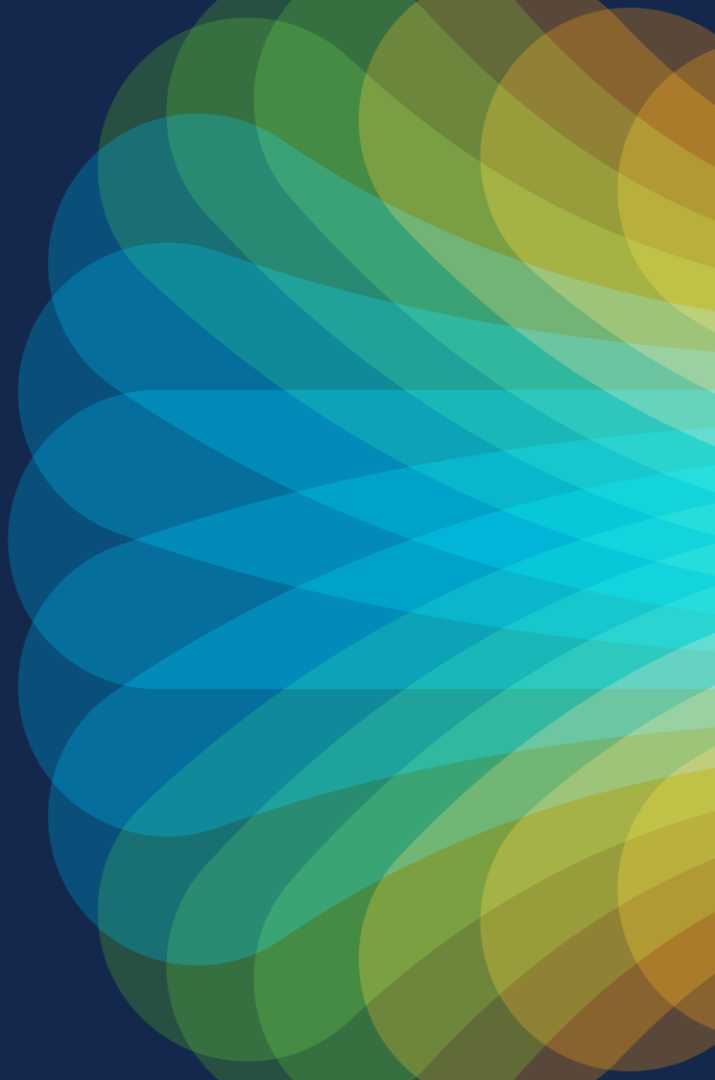
Enable PFC per interface

Non-blocking Network



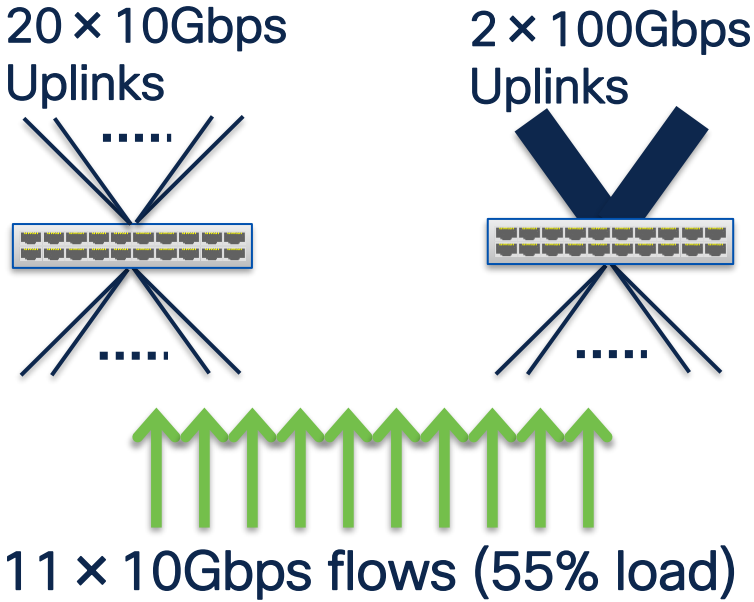
- Non-Blocking Network, allow host to talk to other hosts at full bandwidth
- Leaf: Same bandwidth to the host as to the spine
- Reduces need for congestion management to increase performance

Traffic Load-balancing



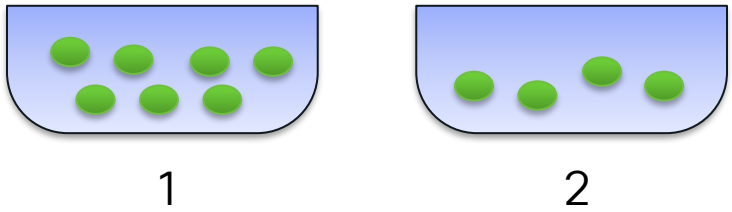
Hashing – Where traffic goes?

- Avoid oversubscribing link, allow multiple paths to ECMP



Prob of 100% throughput = 3.27%

Prob of 100% throughput = 99.95%



Default ECMP algorithm

- By default, ECMP looks at source and destination IP and Layer 4 ports
- Hosts in AI fabric may belong to uniform subnets
- Layer 4 ports, destination port is 4791 for RoCEv2
- Entropy comes from Layer 4 Source Port

```
N9K-switch# show ip load-sharing
IPv4/IPv6 ECMP load sharing:
Universal-id (Random Seed): 2467474893
Load-share mode : address source-destination port source-destination
Rotate: 32
```

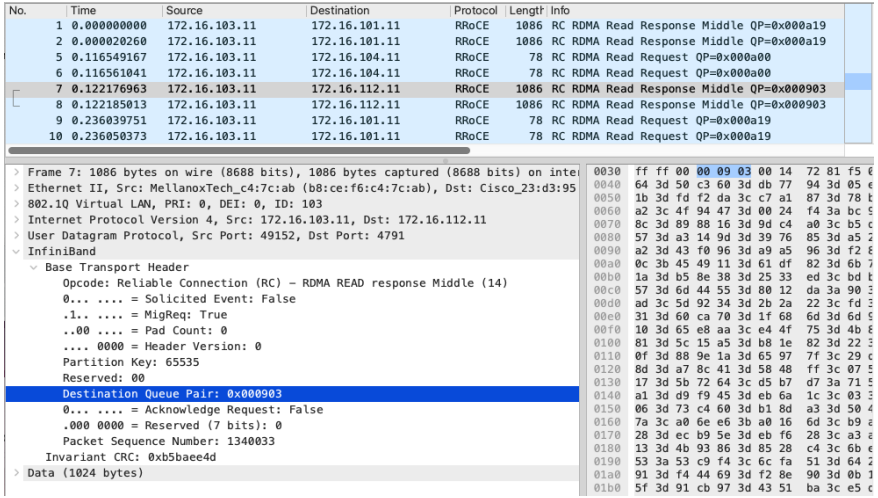
User Defined Field (UDF) ECMP algorithm

- UDF ECMP looks at source and destination IP and User Defined Field in a packet
- User can choose what field to look at to enhance entropy
- Every RoCE conversation is identified by Destinate Queue Pair, in IB header
- Destination Queue Pair is 3-byte field

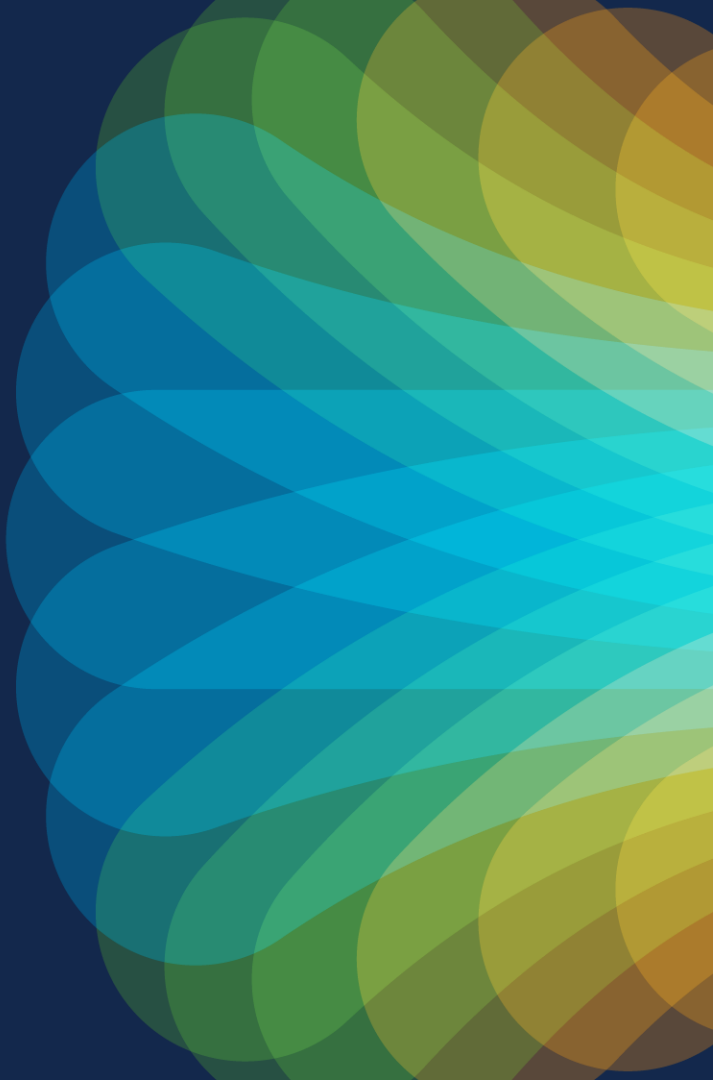
```
N9K-switch# show ip load-sharing
IPv4/IPv6 ECMP load sharing:
Universal-id (Random Seed): 908907021
Load-share mode : address source-destination udf offset 33 length 24
Rotate: 32
```

User Defined Field (UDF) – Destination Q Pair

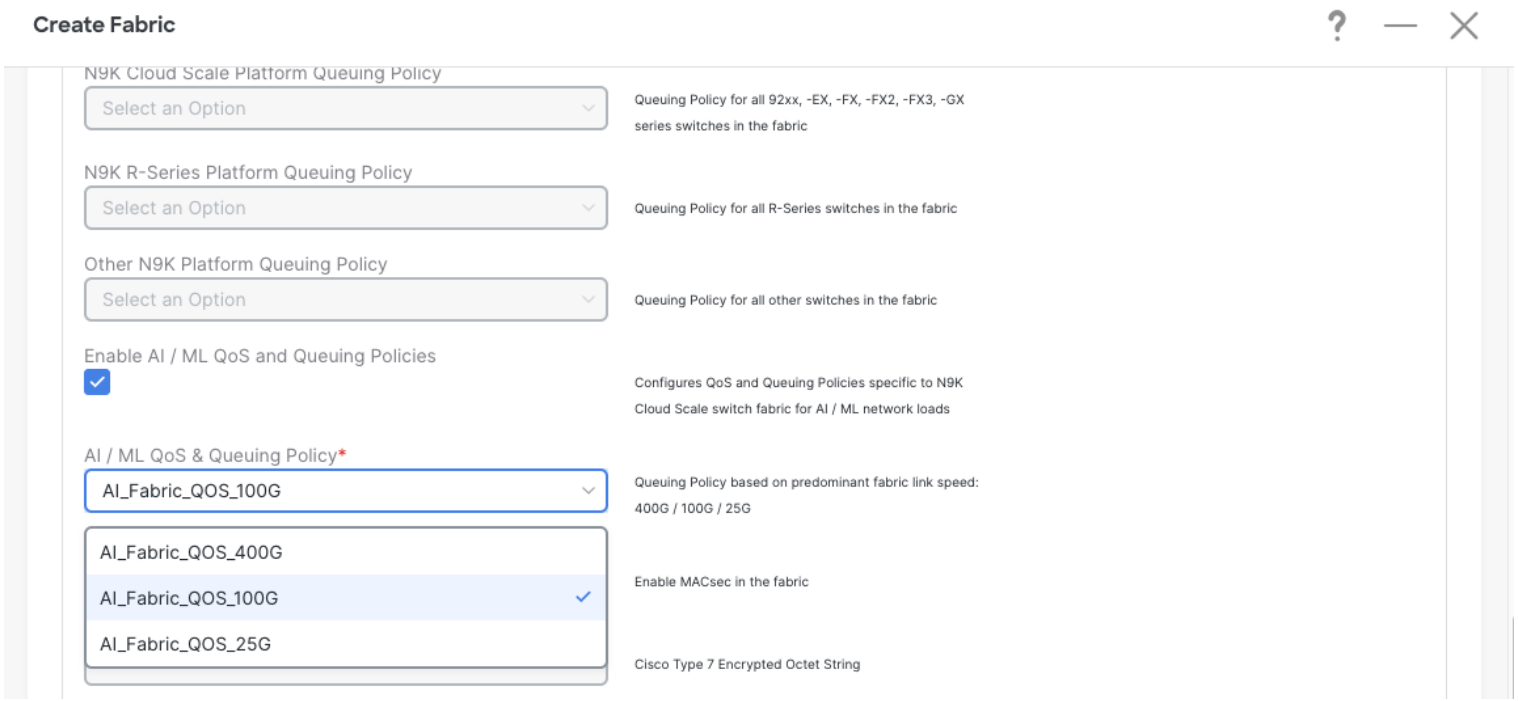
- UDF offset of Nexus 9000 switches starts from first byte of IP header
- Destination Q Pair is 33 bytes from beginning of IP field (IP (20B) + UDP (8B) + IB (5B)) or 6th byte in InfiniBand header



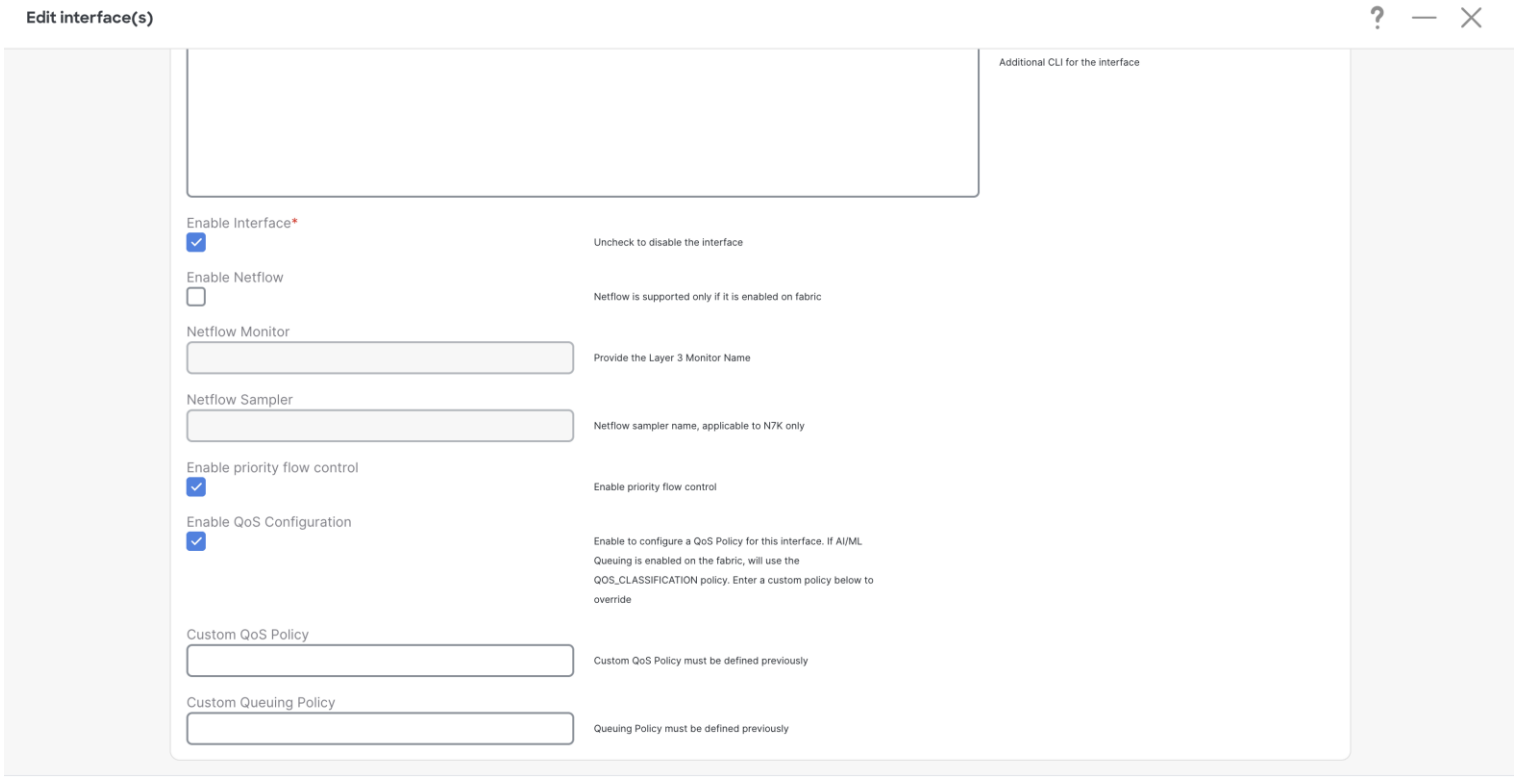
Automation and Visibility



Nexus Dashboard Fabric Controller



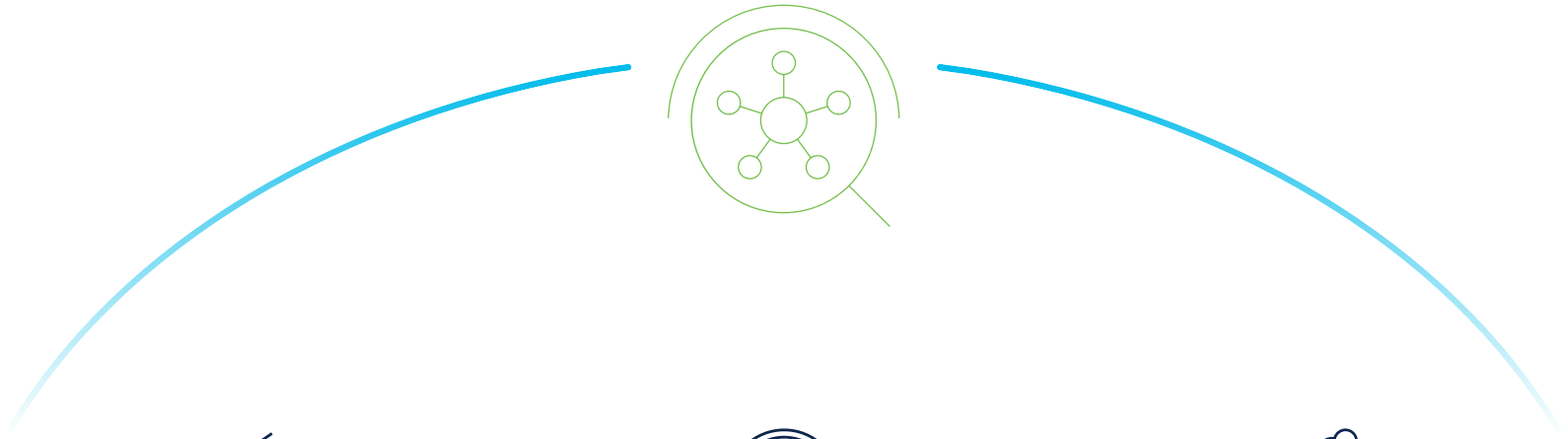
Nexus Dashboard Fabric Controller



Visibility – Flow Table

- Collects full flow information plus metadata
  - 5-tuple flow info
  - Interface/queue info
  - Flow start/stop time
  - Packet disposition (drop indicators)
  - Burst measurement
- Export date to collector
- Leveraged by Nexus Dashboard Insights

Cisco Nexus Dashboard Insights

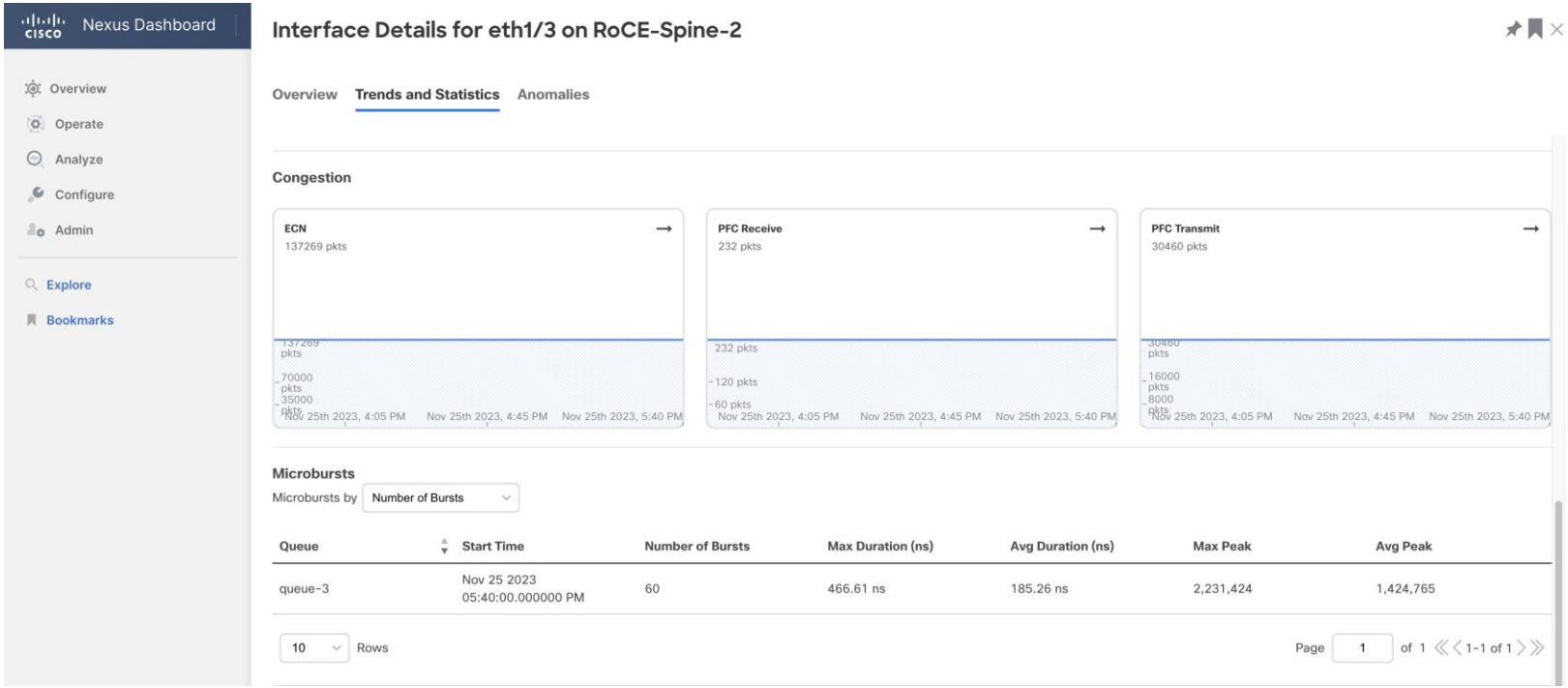


With the granular visibility provided by Cisco Nexus Dashboard Insights the network administrator can observe drops

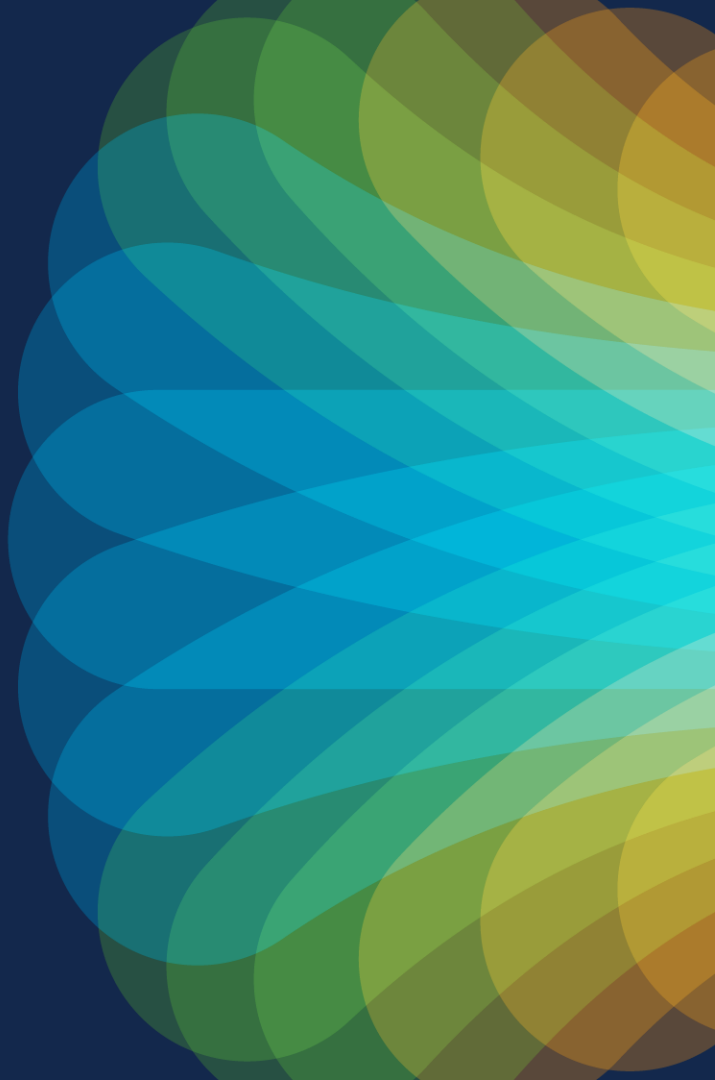
Tune thresholds until congestion hot spots clear and packet drops stop in normal
traffic conditions

This is the first and most important step to ensure that the AI/ML network will cope with regular traffic congestion occurrences effectively

Nexus Dashboard Insights – Congestion Visibility



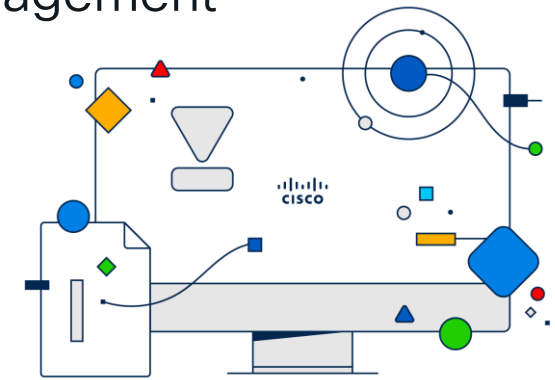
Blueprint of Today



Choosing the right infrastructure

- Build a Clos Fabric / Spine-Leaf
- Fixed Switches
  - Lower Latency, single AISC
  - Power Efficient
- Right Congestion management

- Routed Fabric
  - Use BGP for control plane
- Scalable design
  - Two tier design
  - Three tier design



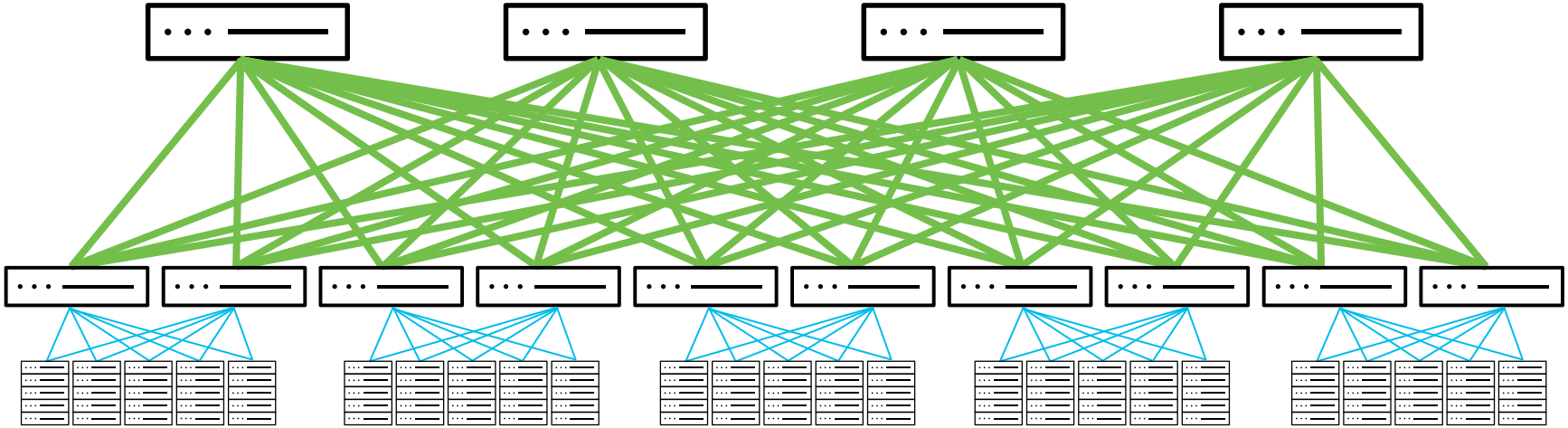
Customer Request #1

- Build Cluster of 260 GPUs, for small training model use case
  - Cluster is bult with stand alone server (e.g. Cisco UCS-C240-M7)
- Build non-blocking network, for GPU communication
- Predictable and low latency for efficient training
- Host connectivity at dual100Gbps
- Fabric at 400Gbps for efficient load-balancing

Customer Request #1 – Proposal

- Standalone server can have up to 2 GPUs
  - As required is 260 GPUs, 130 Stand alone servers are needed
  - Each standalone server is equipped with dual 100G port NIC
- 260 x 100G ports required for host connectivity in leaf layer
  - 26 x 100G host interfaces per leaf switch, for 10 leaf switches
  - For seeped up have 8 x 400G uplinks per switch for Spine connectivity
- Total of 4 spines 20 x 400G ports used per Spine
- Leave room for future expansion

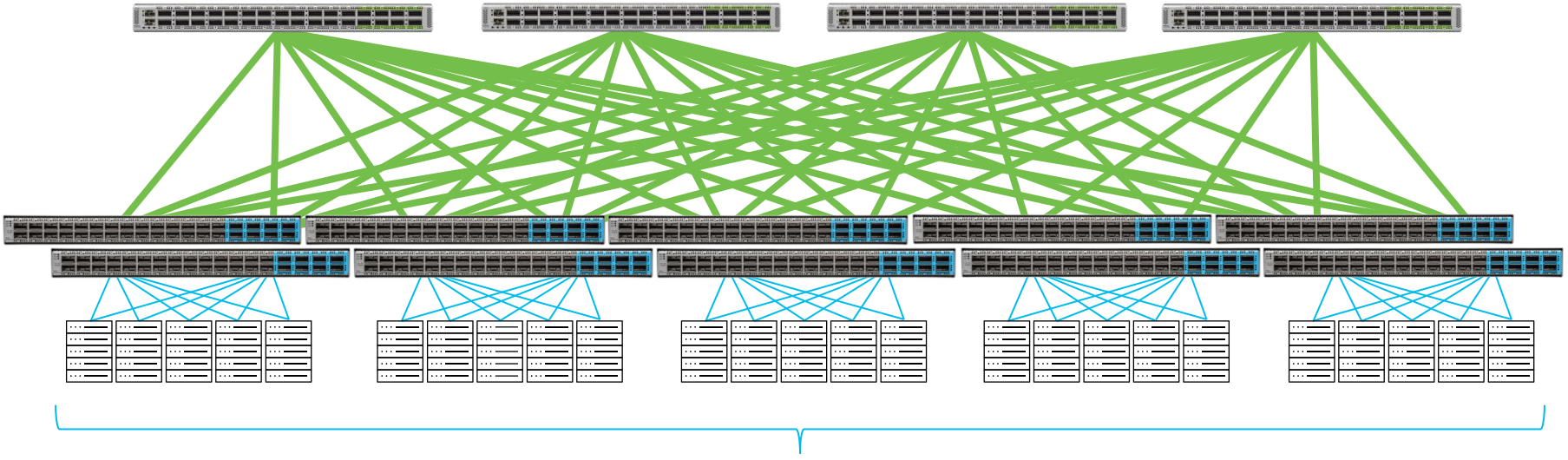
Customer Request #1 – Design



130 Hosts each with 2 x GPU and 2 X100G NIC

100 Gbps
2 x 400Gbps

Customer Request #1 – Design



130 Hosts each with 2 x GPU and 2 X100G NIC

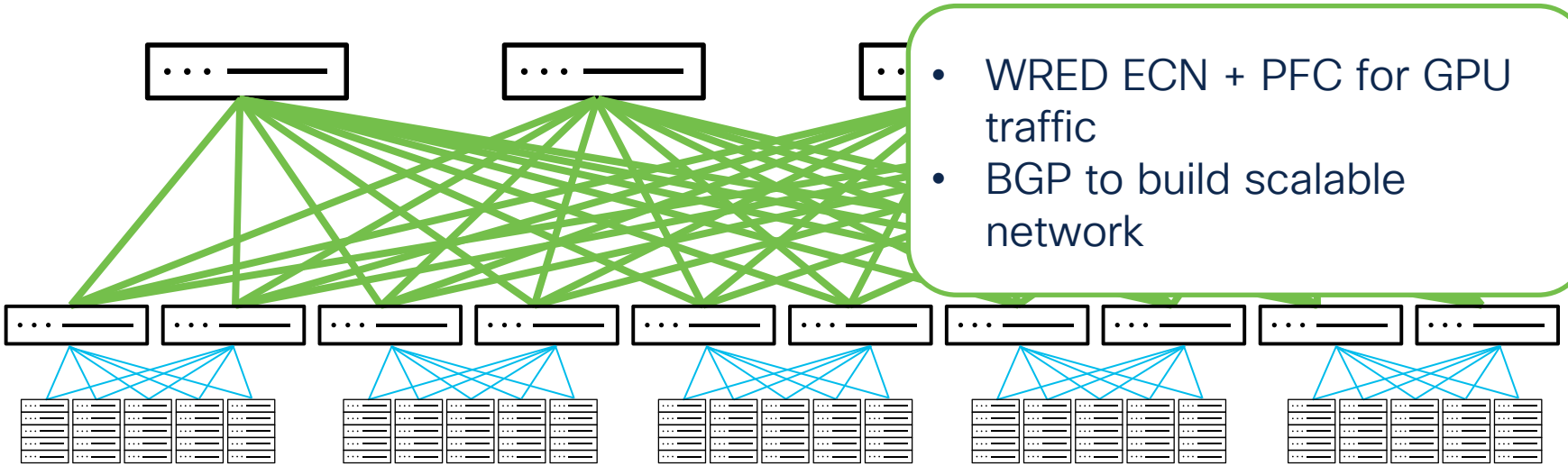
N9K-C93600CD-GX

N9K-C9332D-GX2B

100 Gbps

2 x 400Gbps

Customer Request #1 – Design



- WRED ECN + PFC for GPU traffic
- BGP to build scalable network

130 Hosts each with 2 x GPU and 2 X100G NIC

100 Gbps
2 x 400Gbps

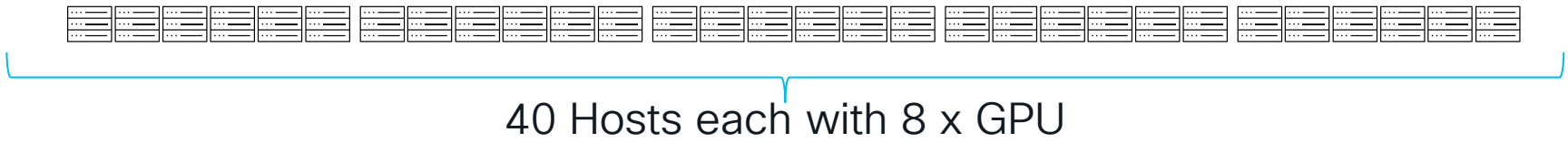
Customer Request #2

- Build Powerful Cluster of 320 GPUs, for large training model use case
  - Cluster is bult with powerful accelerators (e.g. NVIDIA DGX, Intel Gaudi)
- Build non-blocking back-end network, for GPU communication
  - Predictable and low latency for efficient training
  - Host connectivity at 8 ports of 4 x 100Gbps (Logical 400Gbps)
  - Fabric at 400Gbps for efficient load-balancing
- Build front-end network, for server-to-server interaction, and storage connectivity
  - 2 x 100Gbps ports for connectivity, non-blocking and future expansion

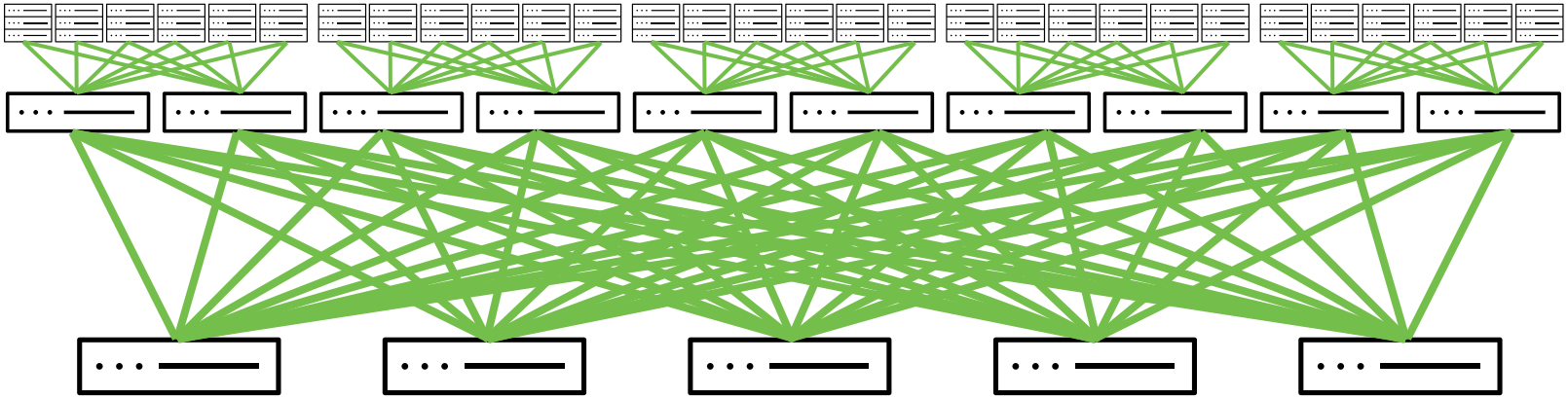
Customer Request #2 – Proposal for Back-End

- Powerful AI training accelerator, has 8 GPUs per server
  - As required is 320 GPUs, 40 host are sufficient
  - Each host is equipped with 4 NICs each with dual QSFP-DD ports
- 320 x 400G ports required for host connectivity in leaf layer
  - 32 x 400G host interfaces per leaf switch, for 10 leaf switches
  - Non-blocking network 32 x 400G uplinks per leaf for Spine connectivity
- Total of 5 spines 64 x 400G ports used per Spine

Front-End and Back-End Cluster Network

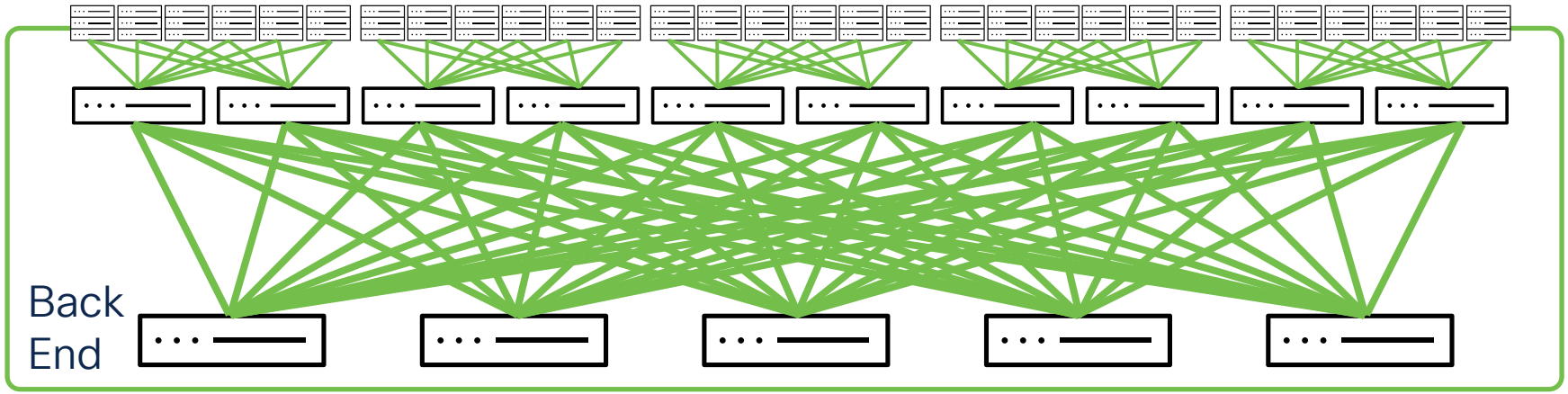


Back-End Cluster Network

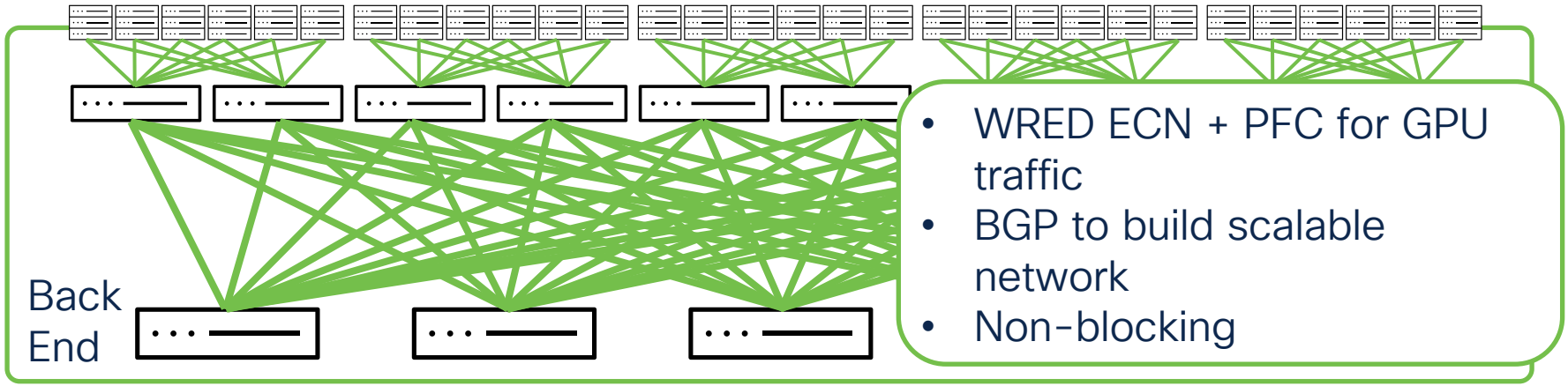


4 x 4 x 100 Gbps — 6 x 400Gbps

Back-End Cluster Network

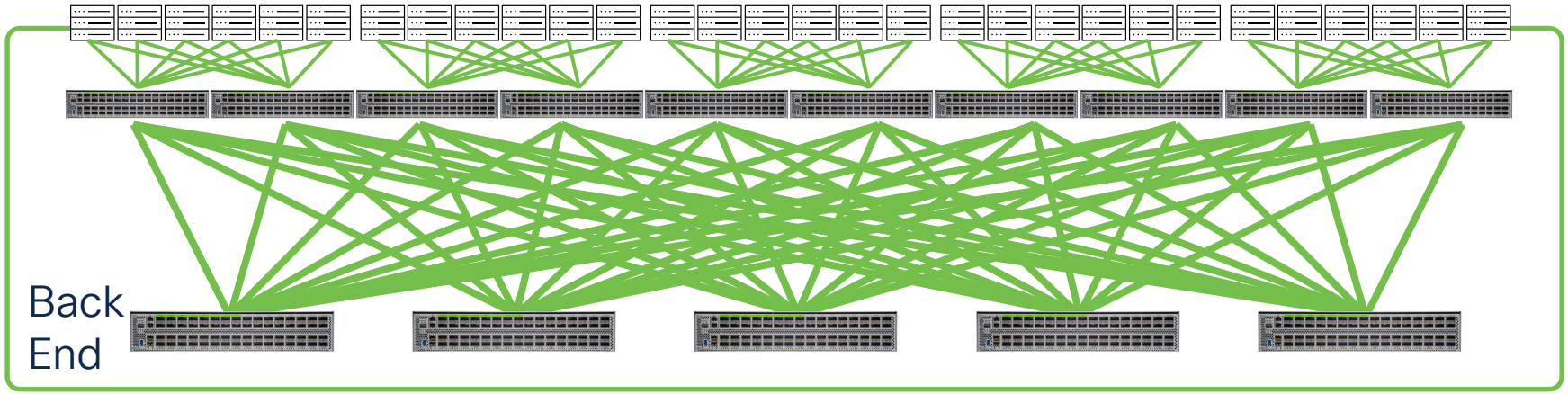


Back-End Cluster Network



Back-End Cluster Network

N9K-C9364D-GX2A

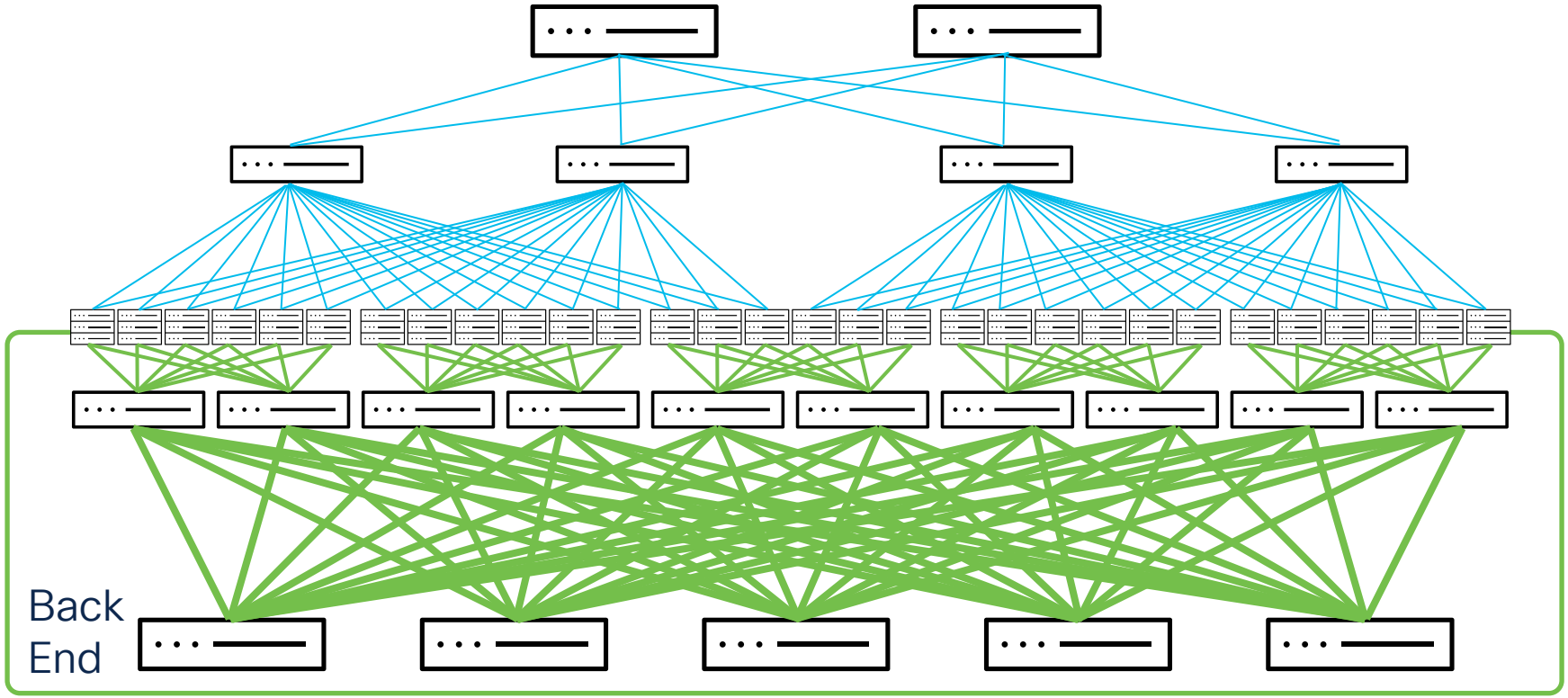


4 x 4 x 100 Gbps 6 x 400Gbps

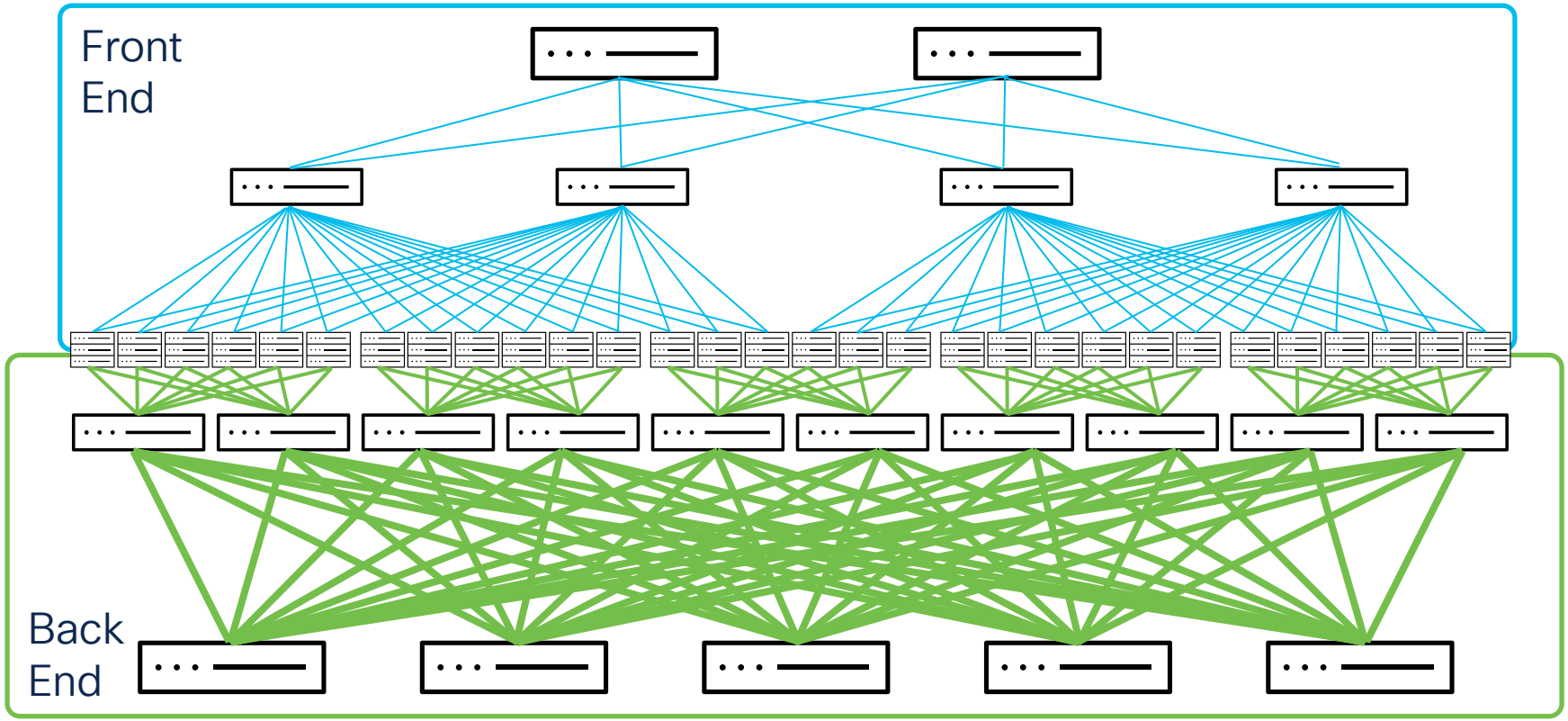
Customer Request #2 – Proposal for Front-End

- Front End network for host communication, and storage
  - Each server has 2 x 100G ports
- 80 x 100G ports required for host connectivity in leaf layer
  - 20 x 100G host interfaces per leaf switch, for 4 leaf switches
  - Non-blocking network 20 x 100G uplinks per leaf for Spine connectivity
- Total of 2 spines 64 x 100G ports used per Spine
- Storage NFS network in Front End
  - 3 Storage array connected to leaf
  - RoCEv2 for storage networks

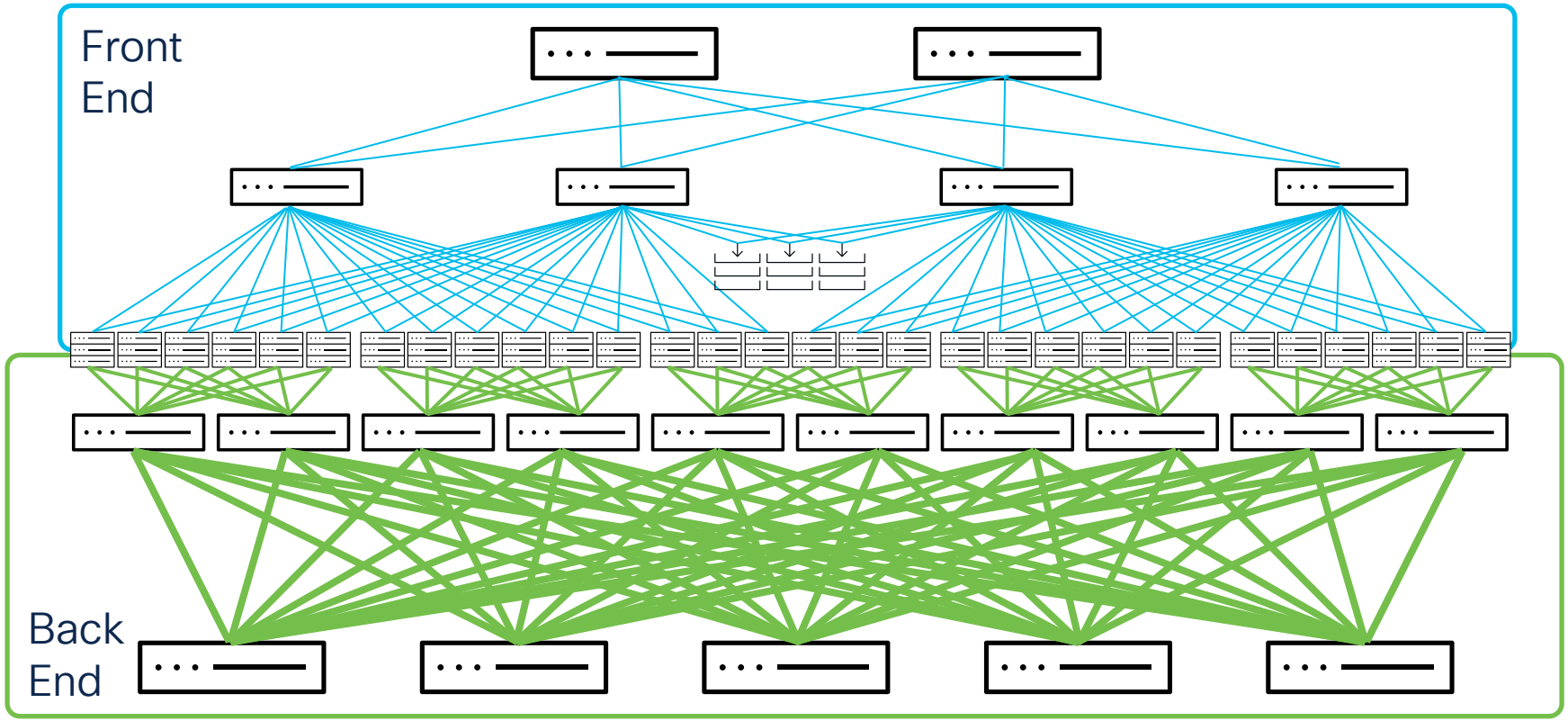
Front-End and Back-End Cluster Network



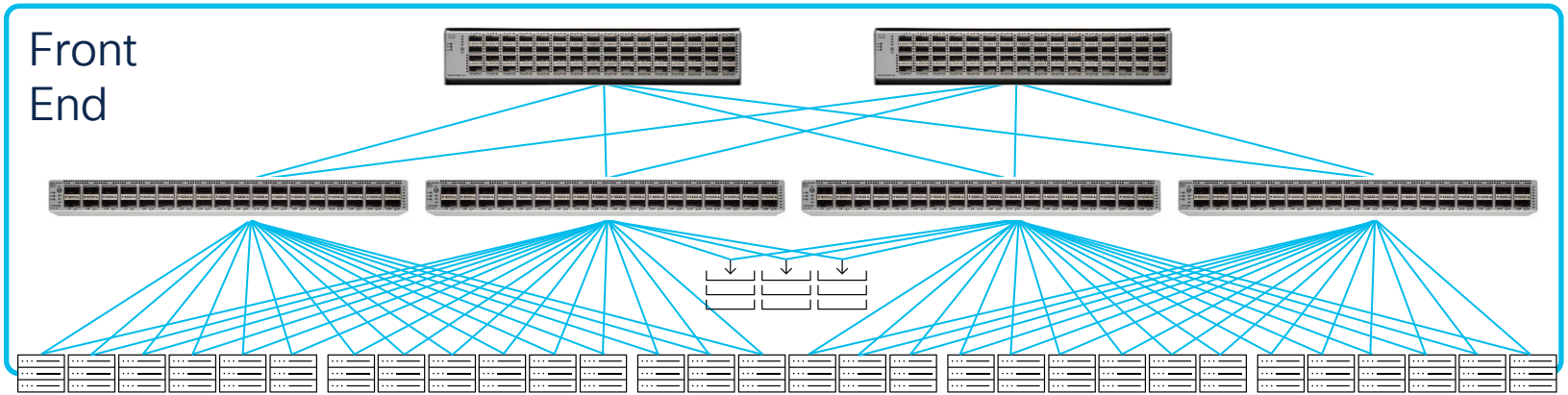
Front-End and Back-End Cluster Network



Front-End and Back-End Cluster Network



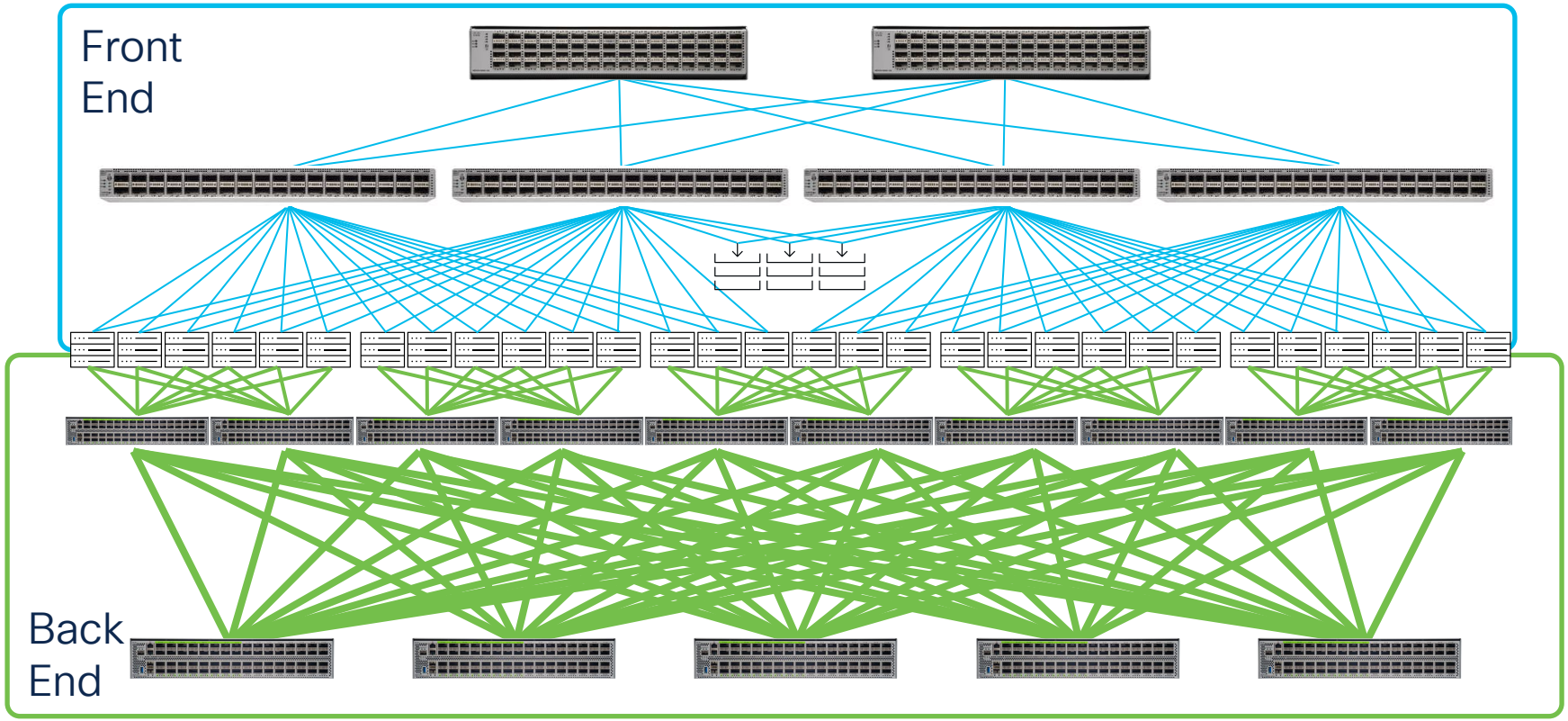
Front-End Cluster Network



N9K-C9364C-GX

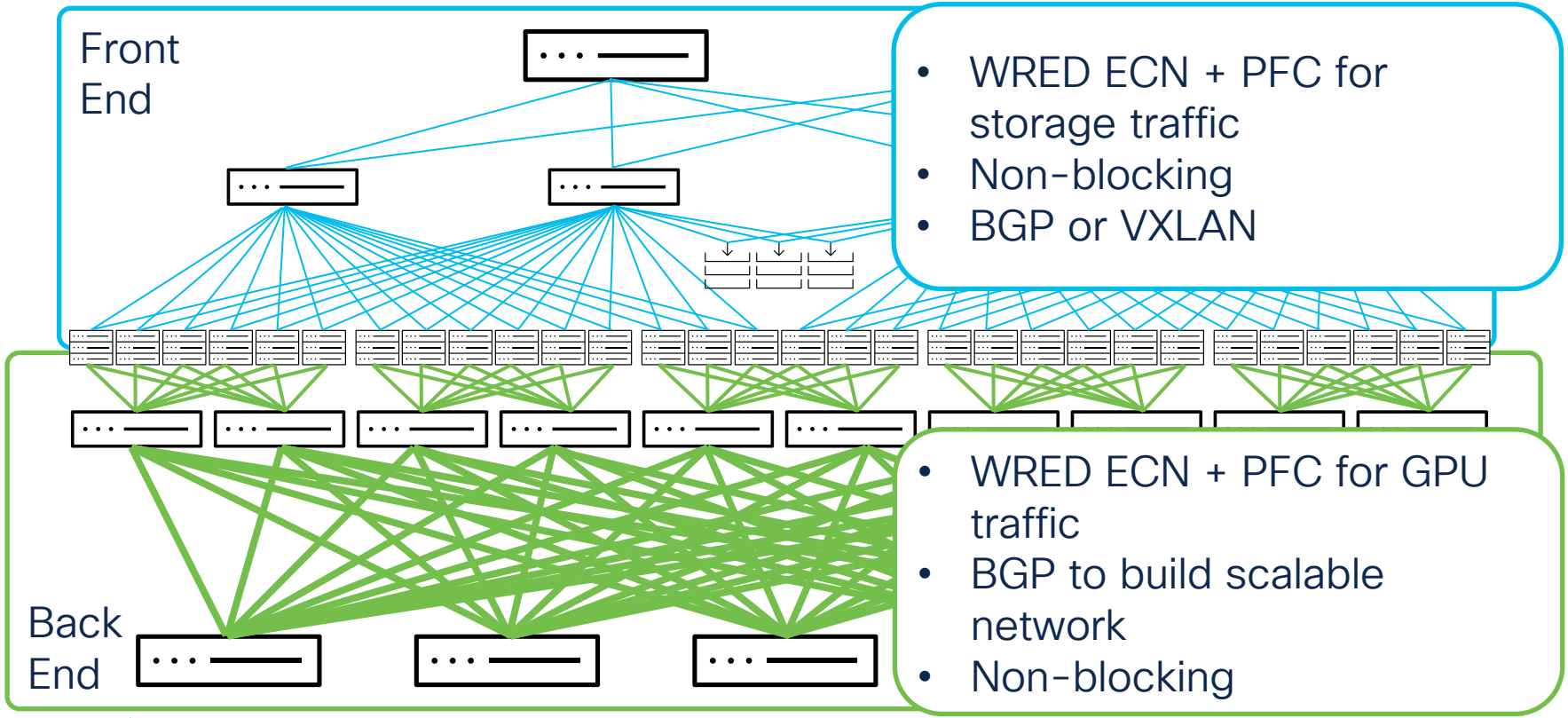
N9K-C9336C-FX2

Front-End and Back-End Cluster Network

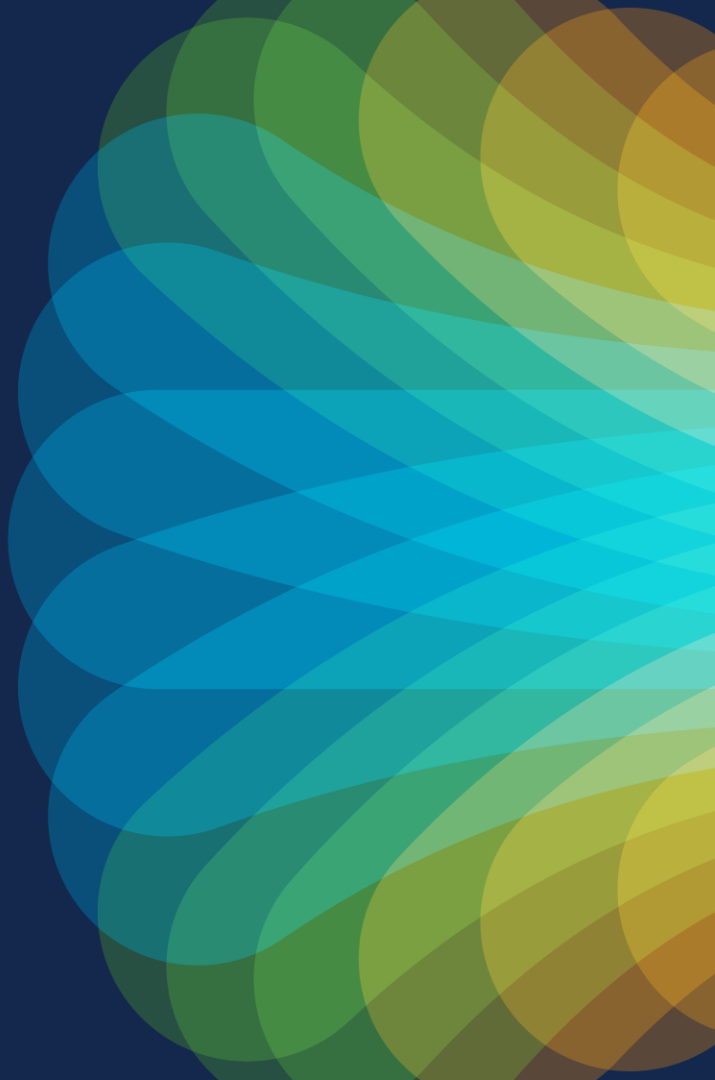


100 Gbps — 4 x 4 x 100 Gbps — 6 x 400Gbps

Front-End and Back-End Cluster Network

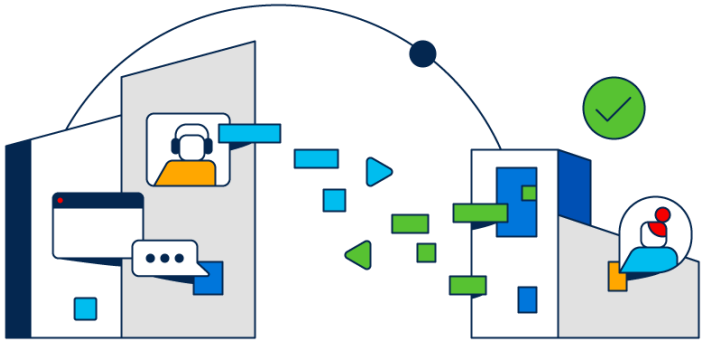


Conclusion

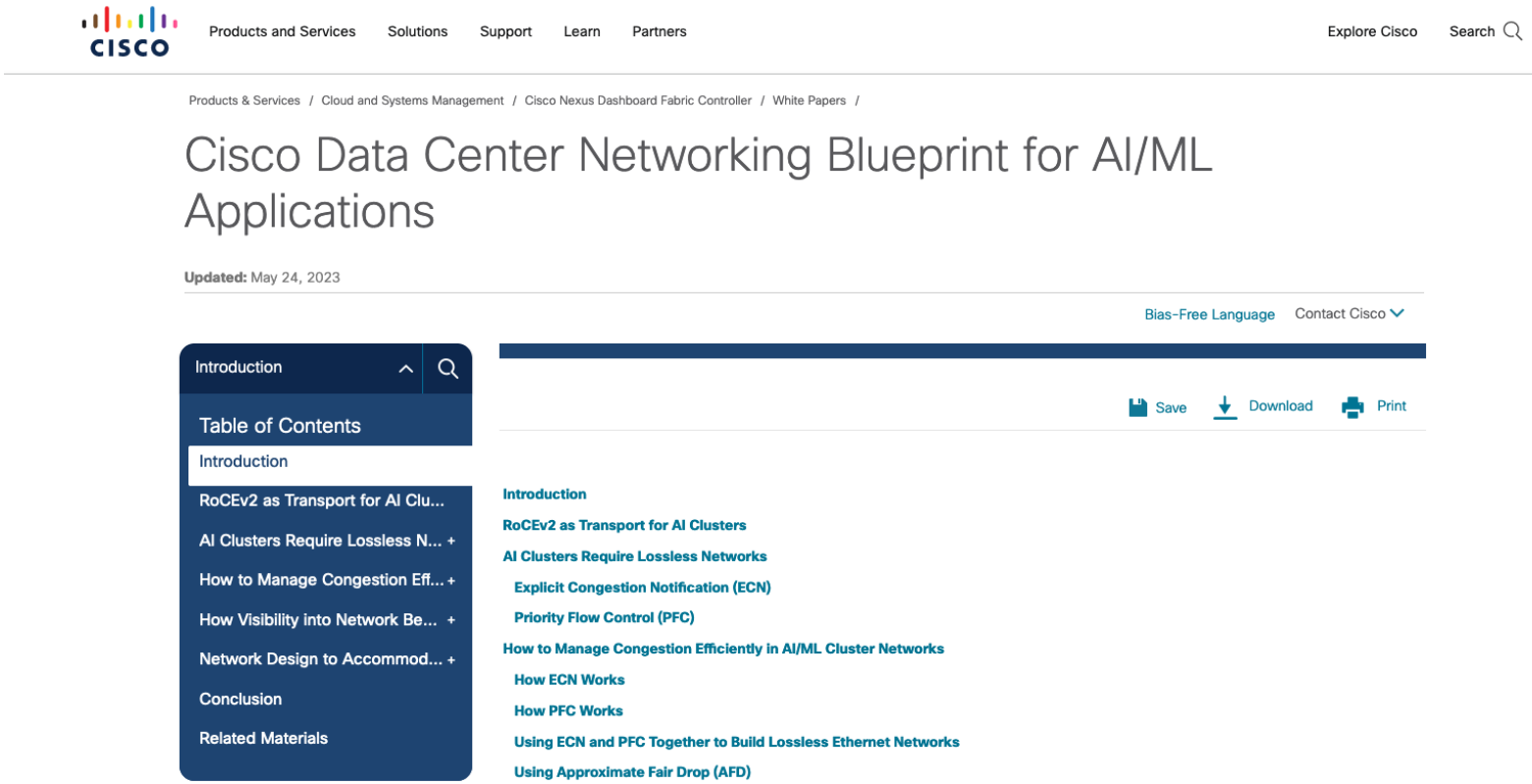


Takeaways

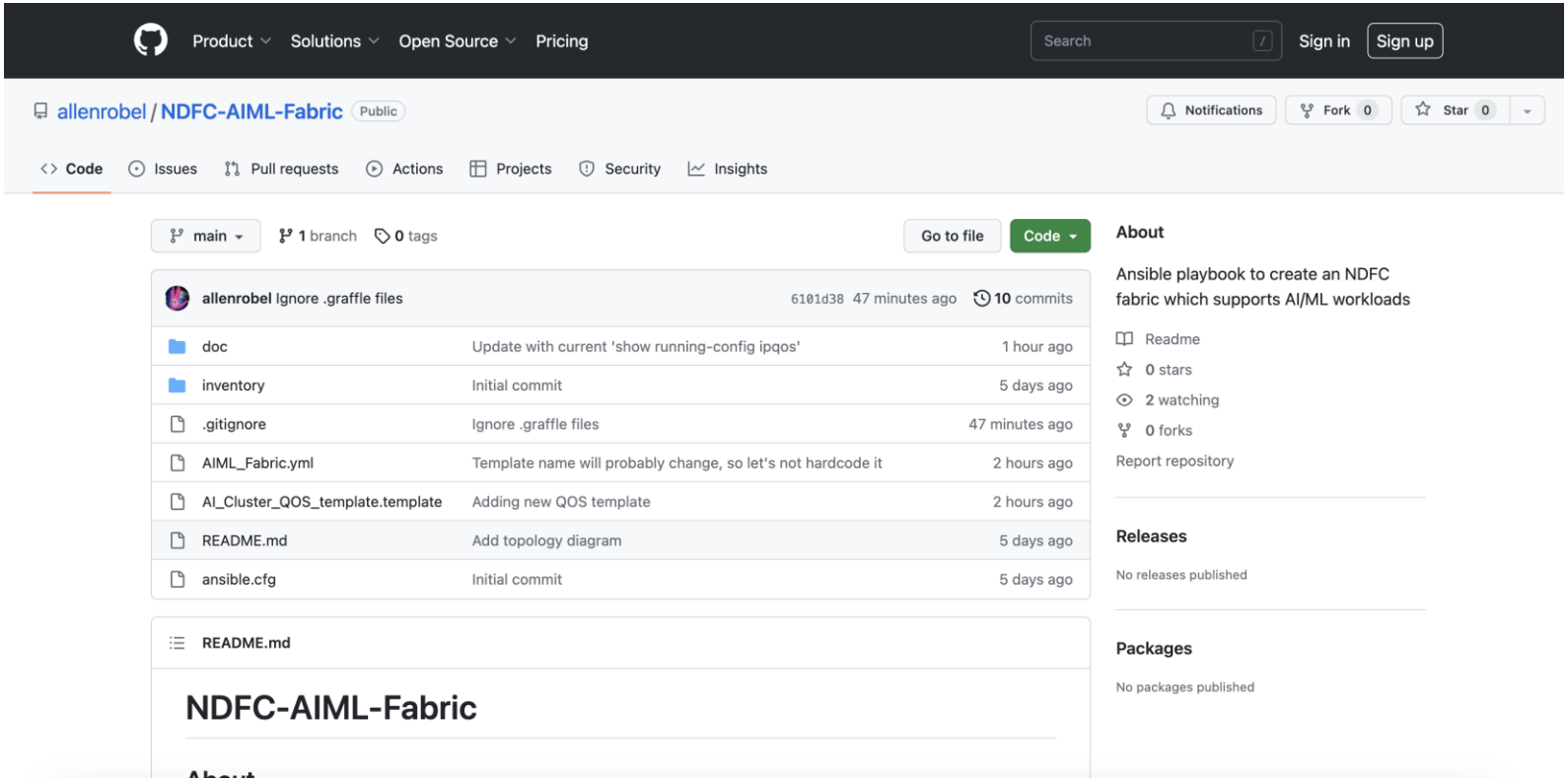
- Robust back-end network, and flexible front-end network
- Familiar data center fabric technologies, BGP or VXLAN
- Automate network for easier bring up and operation
- Visibility in network congestion, and bottleneck to troubleshoot and optimize



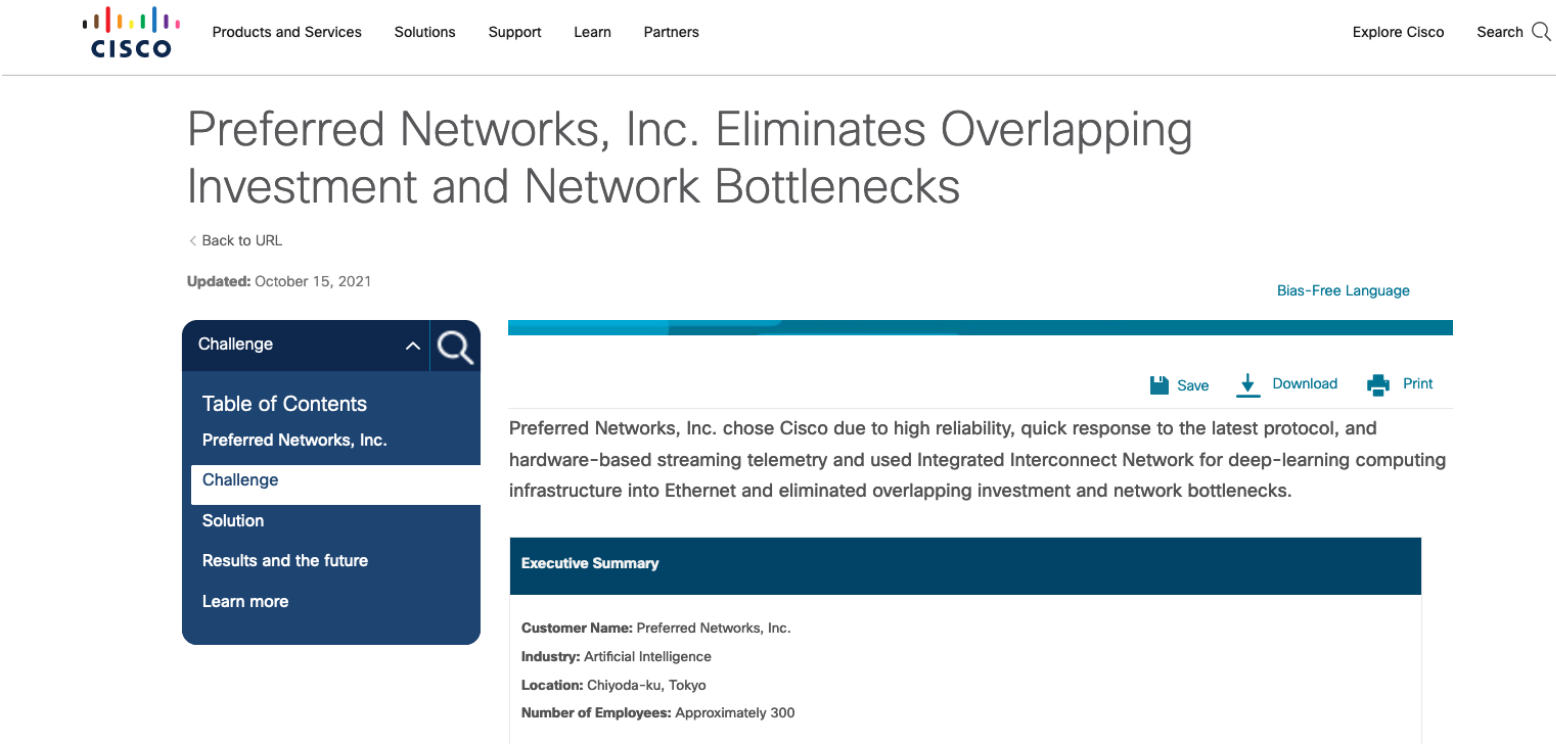
The Blueprint For Today



The Blueprint For Today



The Blueprint For Today



Related sessions

Session ID	Session Title	Day and Time
IBODCN-1010	An Interactive Conversation: AI/ML Networking Requirements and Blueprint	Wednesday, Feb 7, 11:30 AM Thursday, Feb 8, 5:15 PM
BRKDCN-2999	Multi-Tier Fabric-Networks Designs for the Modern Data Center	Thursday, Feb 8, 10:45 AM
BRKDCN-1619	Introduction to NDFC: Simplifying Management of Your Data Center	Tuesday, Feb 6, 3:30 PM
PSODCN-1732	Unlocking the Potential of AI/ML Workloads in Cisco Data Center Networks With Cisco Nexus 9000 Series Switches	Wednesday, Feb 7, 2:10 PM


cisco
The bridge to possible
Thank you
cisco Live!


cisco Live!
Let's go