WEBINAR:

OPEN ETHERNET NETWORKING FOR MODERN AI/ML WORKLOADS

A BLUEPRINT FOR BUILDING THE AI FACTORY

ALAN HUANG
Senior Product Manager
ipinfusion

SUJAY GUPTA
Senior Solutions Manager
Edge-corE NETWORKS

SEPTEMBER 10, 2025 | 14:00 AM BST | 6:00 AM PT

Table of Contents

- Company Introductions
- AI Solution Ecosystem
  - Hardware
  - Software
- Use Case Deep Dive - Ethernet Fabric for AI
- Ethernet vs InfiniBand Brief
- Summary and Future Development
- Q&A

ALAN HUANG
Senior Product Manager
ipinfusion

SUJAY GUPTA
Senior Solutions Manager
Edge-corE NETWORKS

IP Infusion Corporate Overview

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**

25 Years of Network Innovation

HQ: Silicon Valley With Global Presence	**Technology Leader in Open Networking** Telecom Infra Project, GigaOm, O-RAN Plugfest		
R&D Centers Bengaluru, Ottawa, Montreal, Israel	**2.4M** Ports Shipped	**120K+** Licenses Sold	**10,000s** Carrier Deployments
400 Employees	**+100** New Customers in 2022 and 2023 each	**600+** Customers	**15** of Top 88 Global Telecoms

Product and Technology Leadership

- **OcNOS** Carrier Grade NOS
- Control Plane
- White Box Solutions
- **IP Maestro** Point-and-Click EMS

Total Network Disaggregation

- Service Provider
- Data Center
- OEM

IP Infusion Advantages for Open Networking

OcNOS

Most Comprehensive Open NOS

The Widest HW Solution Ecosystem

Open Optics Ecosystem

Centralized Monitoring and Management

24/7 Professional Support

600+
Modern Networking Features

40+
Supported Hardware Platforms

100+
Qualified Optical Transceivers

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**

IP Infusion Client Roster

NETWORK OPERATORS

afribone, africell, AIST, Amazonet, AMPLEX INTERNET, archtopfiber, ASOM NET, AST, AT&T, atc Broadband, BROADSTAR

BCS, Bedroq, BT, Cable One, COMSOL, CONCORD, Dataduct networks, diggio, DIGI, dish, Disney

dojo, eletronet, EN, encoo COMMUNICATIONS, ENet, EXATEL, EQUINIX, FASTWAY, FH TECHNIKUM WIEN, FIXED MOBILE TELECOMS, fit, FTC

GOLD DATA, gridtelekom, GT, hSo:, HARMONY Telephone Company, Himachal Futuristic Communications Limited, HIRSCHMANN, HUBER+SUHNER, INSIST GLOBAL, indosat OOREDOO HUTCHISON

inverso, KOREN, KWIKOM CONNECTING COMMUNITIES, LIBERTY, lintasarta, LIVEOAK FIBER, MAINSTREAMING, MOMENTUM TELECOM, multinet, mundo, MTN, NASA

napatech, Netforward, NEWROZ, NEXGEN, nbn, NORTHWEST, NTT, NHK, North Atlantic Industries, PETRON COMMUNICATIONS LTD., point BROADBAND, OKINAWA OPEN LABORATORY

Open Coconut, OPENX, PURESTORAGE, RelAix Networks, RNP, RURAL NET, SaskTel, SCOTTDATA, SIEMENS, Sinergy, SK telecom, SKY

SOLITON telmec, softax, SQUILD, エネコム, syrex, Tachyum, TARGO, Taiwan Mobile, TATA COMMUNICATIONS, TRUESPEED

Telenet, ufinet, UGRID, Uniti, VIAVI, Viettel, VIRTUAL TECHNOLOGIES AND SOLUTIONS, VOCALTRANSIT, Vyve, WISPER INTERNET, WOW!, YouFibre

NETWORK EQUIPMENT MANUFACTURERS

alcatel, Alcatel·Lucent, Allied Telesis, AVAYA

BENU NETWORKS, BROCADE, casa systems, CERAGON networks

ciena, CISCO, citrix, DELL, ERICSSON

f5, FUJITSU, Gilat, HITACHI

HUAWEI, IBM, infinera, intel, McAfee

NEC, SAMSUNG, Selex ES, SONICWALL

STARENT NETWORKS, Telcordia, THALES, UGRID

Viasat, VYATTA

About Edgecore

- Portfolio of Open Networking Products, Solutions and Services
- Delivering to Large Tier1's and Enterprise Customers Worldwide
- Independent Branded Company Accton Owned Subsidiary Since 2010
- Worldwide Sales and Support, Headquartered in Hsinchu Taiwan
- Flexible Business Model Solutions Provider


Telecom


AI & Data Center


Enterprise

OcNOS
NOS Software

About Accton: The Parent Company of Edgecore

With over 35 years of experience, Accton is a well-known technology ODM/JDM provider for global enterprises, recognized for innovative technologies and manufacturing excellence, earning a distinguished industry reputation.

- Established in **1988**
- Global operating sites extend across **North America, Europe, and Asia**
- Number of Employees: **6,500**
- 2024 Revenue: **USD3.4 billion**

Manufacturing in **China**
Space: 71,040 m²

Manufacturing **in Zhunan, Taiwan**
Space: 15,518 m²

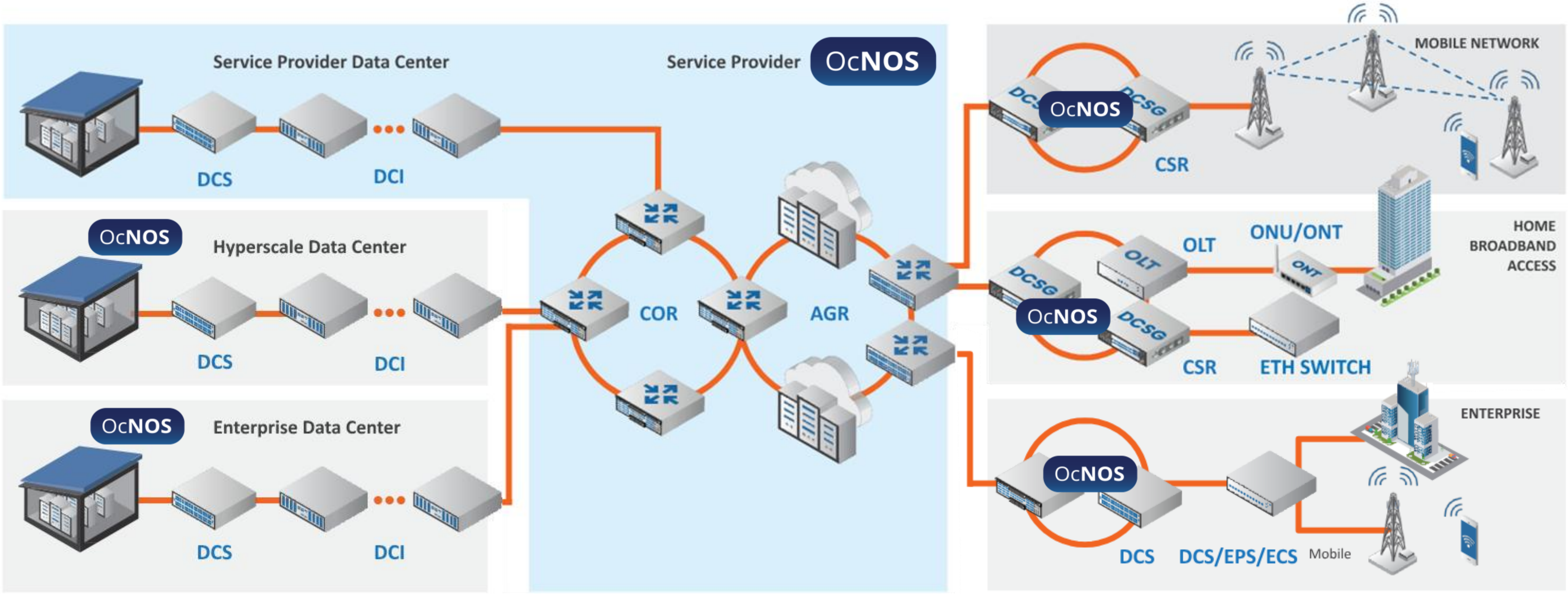
Manufacturing in **Vietnam**
Space: 11,340 m²

Office and Warehouse in **the United States**

Brand new facility launched in 2024, in Zhubei, Taiwan

Open Networking Solutions from Edge to Core

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**



AI/Cloud Data Center Solutions

Service Providers Solutions

Enterprise Solutions

Edgecore Data Center Portfolio

IP Infusion OcNOS Qualified

Edge-corE NETWORKS | ip infusion

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**

2022

More than 50% of DC switches in 2nd largest marketplace

2025

World's largest Payment Gateway DC using Edgecore

Spine Switches
Tomahawk family

DCS501
32x 100G – TH1 – 3.2T

DCS500
64x 100G – TH2 – 6.4T

DCS511
32x 400G – TD4 – 12.8T

DCS510
32x 400G – TH3 – 12.8T

DCS520
64x 400G – TH4 – 25.6T

AIS800-64O/D
64x 800G – TH5 – 51.2T

AIS800-32O/D
32x 800G – TH5 – 25.6T

Leaf Switches
Trident family

DCS202

DCS201
6x 100G, 48 x 10G – TD3 – 1.08T

DCS203
8x 100G, 48x 25G – TD3 – 2.0T

DCS204
32x 100G – TD3 – 6.4T

DCS240
32x 400G – TD4 – 12.8T

DCS230
8x 400G, 48x 100G – TD4 – 8.0T

DC Mgmt/ Enterprise Switches
Trident family

EPS202

EPS201
48x 1G, 4x 25G – TD3 – 480G

EPS203
36x 2.5G, 12x 10G, 4x25G – TD3 – 560G

EPS121
48x1G, 6x10G – TD3-X2 – 108G

EPS122
48x1G(POE), 6x10G - TD3-X2

AI DC – a. High Radix b. Suits Leaf and Spine c. Low Latency d. E-W traffic

Qualified with IP Infusion OcNOS

Enterprise/Cloud DC – a. 25/100/400G b. Over-Sub c. E-W and N-S traffic

Edgecore Data Center Portfolio

IP Infusion OcNOS Qualified – by Use Case

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**

Trident Family
a. Higher Buffer than Tomahawk
b. Better QOS
c. Feature Rich(Virtualization, IP)

AS7326-56X | DCS203
8x 100G, 48x 25G – TD3 – 2.0T

AS9736-64D | DCS520
64x 400G – TH4 – 25.6T

AS9817-64D | AIS800-64O/D
64x 800G – TH5 – 51.2T

AS5835-54T | DCS202
6x 100G, 48x 10G – TD3 – 1.08T

AS7816-64X | DCS500
64x 100G – TH2 – 6.4T

AS9726-32DB | DCS511
32x 400G – TD4 – 12.8T

AS4625-54T | EPS121
48x1G, 6x10G – TD3-X2 – 108G

AS5835-54X | DCS201
6x 100G, 48x 10G – TD3 – 1.08T

AS7726-32X | DCS204
32x 100G – TD3 – 6.4T

AS9716-32D | DCS510
32x 400G – TH3 – 12.8T

128 Gbps → 1.08 – 2.0 Tbps → 3.2 – 6.4 Tbps → 12.8 – 25.6 Tbps → 51.2 Tbps

Management Switch | Storage Fabric Switch | AI Fabric Switch

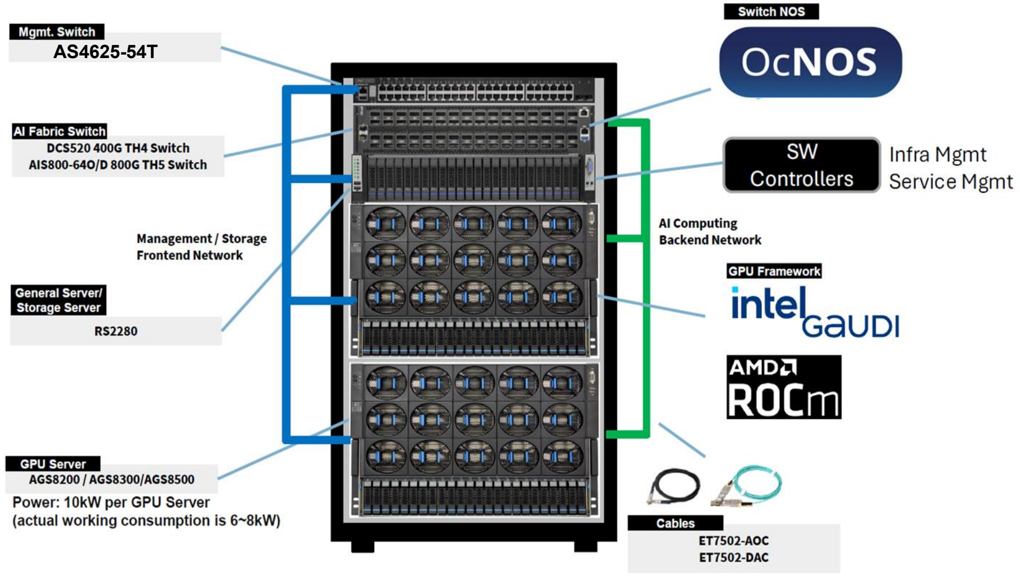
Edgecore AI Solution

OcNOS

Deployment proven, open standards-based disaggregated Networking OS providing high performance, extensive programmability, flexibility and interoperability

Data Center Switches
High performance, low latency switches for GPU interconnect and leaf/spine use cases, bringing advanced load balancing and congestion control features needed for the critical parts of your network

Edgecore AI Rack Total Solution



Edgecore GPU Server Portfolio

GPU Servers (AGS Series)
State-of-the-art GPU servers for AI, machine learning, and data analytics to accelerate your most demanding workloads
Intel Gaudi 2, Gaudi 3
AMD MI 300, MI 325

Transceivers and Cables
Enhance your network's performance and reliability with our high-quality transceivers and network cables, designed for seamless connectivity and superior data transmission

The Modern - Accelerated Compute Systems

Edge-corE NETWORKS ipinfusion

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**

- Next phase in Evolution of Computer Systems
- Every new modern Server/Workstation now has compute accelerators to power today's modern applications

Types of AI customers
- Cloud service providers
- Colo providers
- Enterprises (various verticals: logistics, oil exploration, chemical, government, etc.)

BAIDU AI CLOUD
IBM Cloud
Azure
Alibaba Cloud
Tencent Cloud
aws
ORACLE CLOUD

Availability of Cloud Based AI Acceleration Systems Today

***Public cloud hosted GaaS vs on-prem AI DC cost comparison:**
Average 3-Year Reserved H100 public cloud price:
8,000 GPUs
8000 * $3.00/hour * 24 hours/day * 365 days/year * 3 years = $631M

***On-prem AI DC**
1,000 H100 GPU servers (8 GPUs per server):
1,000 * $120K = $120M
64 IP Infusion TH5 bundles + frontend-network/storage-fabric switches < $3M
3Y power cost ~ $37M (US industrial avg)
3Y TCO savings > 74%

From the Cloud → To Office → To Datacenter → To Edge

Cloud Based GPU

GPU Accelerated Data Analytics
Data scientists, researchers and developers

Multi-GPU Based Data Centers
Gives Data Locality, Model Training and Tuning Capabilities – Serves as the foundation for organizations in the AI maturity cycle

Deployment Platforms which can Infer at Real Time

AI Lifecycle

* - based on estimations

AI Stack and Performance

- Application
- Platform
- Acceleration Libraries
- System Software
  Example: CUDA/DOCA/Magnum IO/Base Command/Forge
- **Hardware**
  GPU's, CPU's, DPU's, NIC, Switch, Optics

GPU's have different architectures for different workloads:

- Large Scale LLM Training and Inference

NVIDIA B200, H100
AMD MI 300, MI 325
Intel Gaudi

- Data Analytics, Conversational AI, Language Processing

NVIDIA H100

- Gaming, 3D Rendering

NVIDIA L405

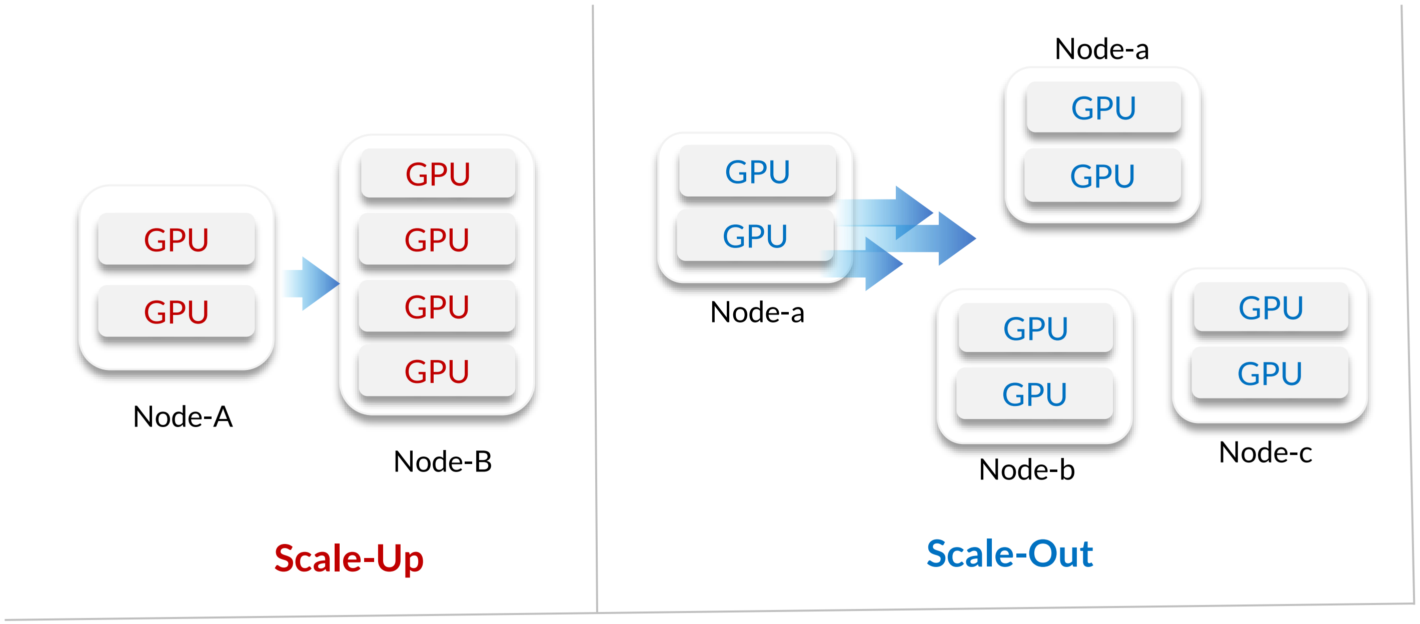
- Machine Learning

NVIDIA Grace

Nature of GPU workloads

- GPU's perform parallel processing, to maximize GPU efficiency the data must always be available. Which in turn requires High bandwidth with low latency and low jitter.
- As AI models and related datasets are growing, there is a need for multi-GPU systems.
- Certain AI models can be efficiently run on multi-GPU systems

Multi-GPU Systems & Performance



- Overall performance of multi-GPU dependent on:
 ‘ Data Must Always be Available for the GPU’

➢ Hardware
➢ Data Management
➢ GPU utilization
➢ Network Configuration

- GPU to GPU communication All to All – PCIE not sufficient
- Chip-to-Chip Interconnect technologies such as (‘Nvlink + NvSwitch’, AMD Infinity Fabric, *UA link*)

Multi GPU Systems

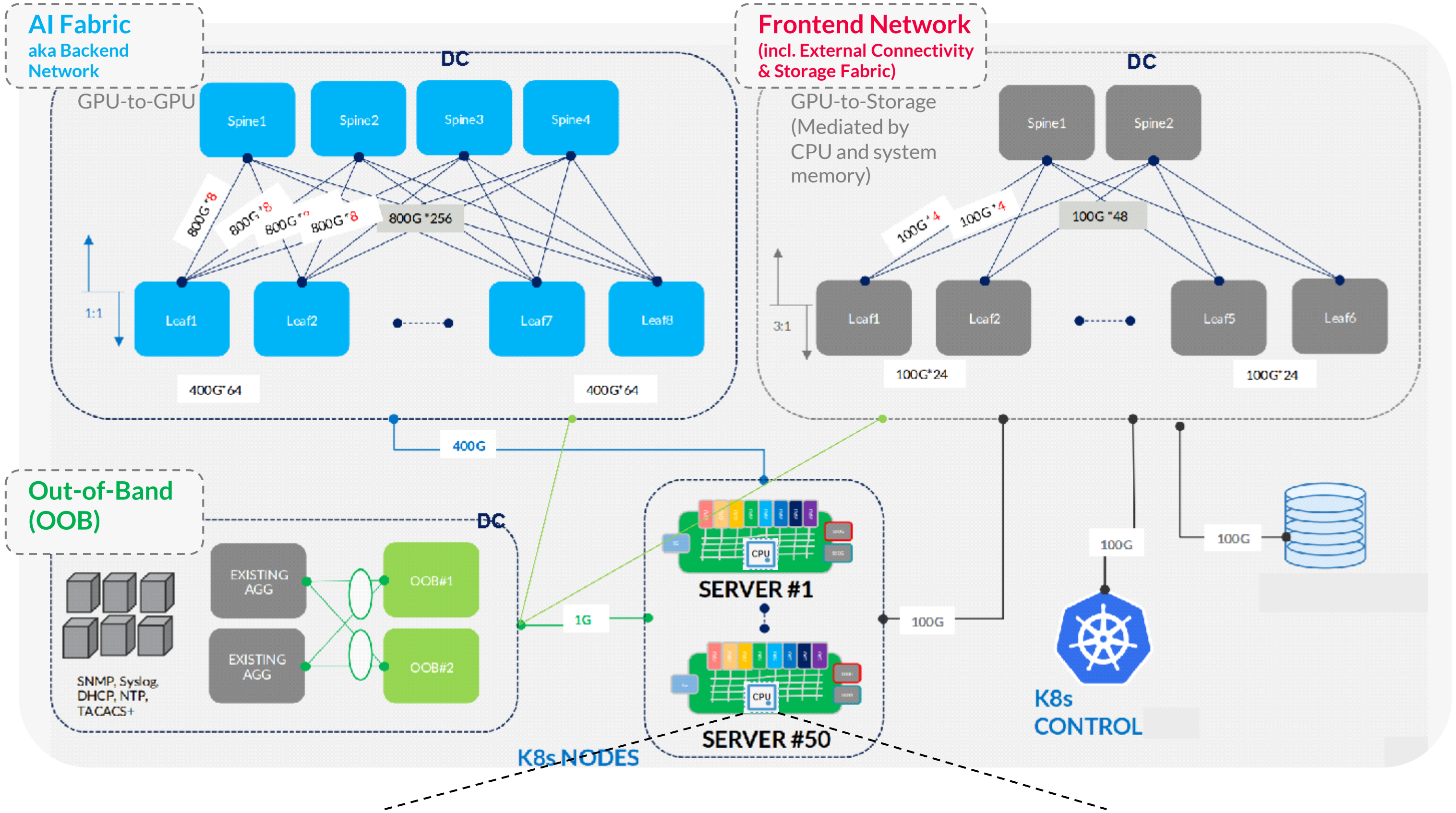
- Scale-Up - Has inherent Weak Fault Tolerance
- Scale-Out – Has Robust Fault Tolerance

Scale-Out

- Network Topology
- Bandwidth and Latency
- Network Protocols
- Data Transferring Techniques
- Management Methods

OcNOS AI/ML Use Case Review

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**



Data → **Data stack**
Transform data into features & labels
Pandas, Polars, Spark, DBT, Flink, Bytewax

Kubeflow
Pipeline orchestration

Training stack
Train models with features & labels
PyTorch, TensorFlow, Scikit-learn, XGBoost, Jax

OcNOS Network Service Highlights

Powering AI/ML data center network

AI Fabric (aka Backend Network)

- Ethernet based Layer 3 IP network
- Dynamic load balancing to avoid collision of long lasting elephant flows
- Lossless ROCEv2 (*RDMA over Converged Ethernet*) transport via PFC (*Priority Flow Control*) and ECN (*Explicit Congestion Notification*)
- Efficient support for mixed traffic types via ETS (*Enhanced Traffic Selection*)

Frontend Network (incl. External Connectivity & Storage Fabric)

- EVPN-VxLAN overlay network
- Dynamic load balancing to avoid collision of long lasting elephant flows
- Lossless ROCEv2 (*RDMA over Converged Ethernet*) transport via PFC (*Priority Flow Control*) and ECN (*Explicit Congestion Notification*)
- Efficient support for mixed traffic types via ETS

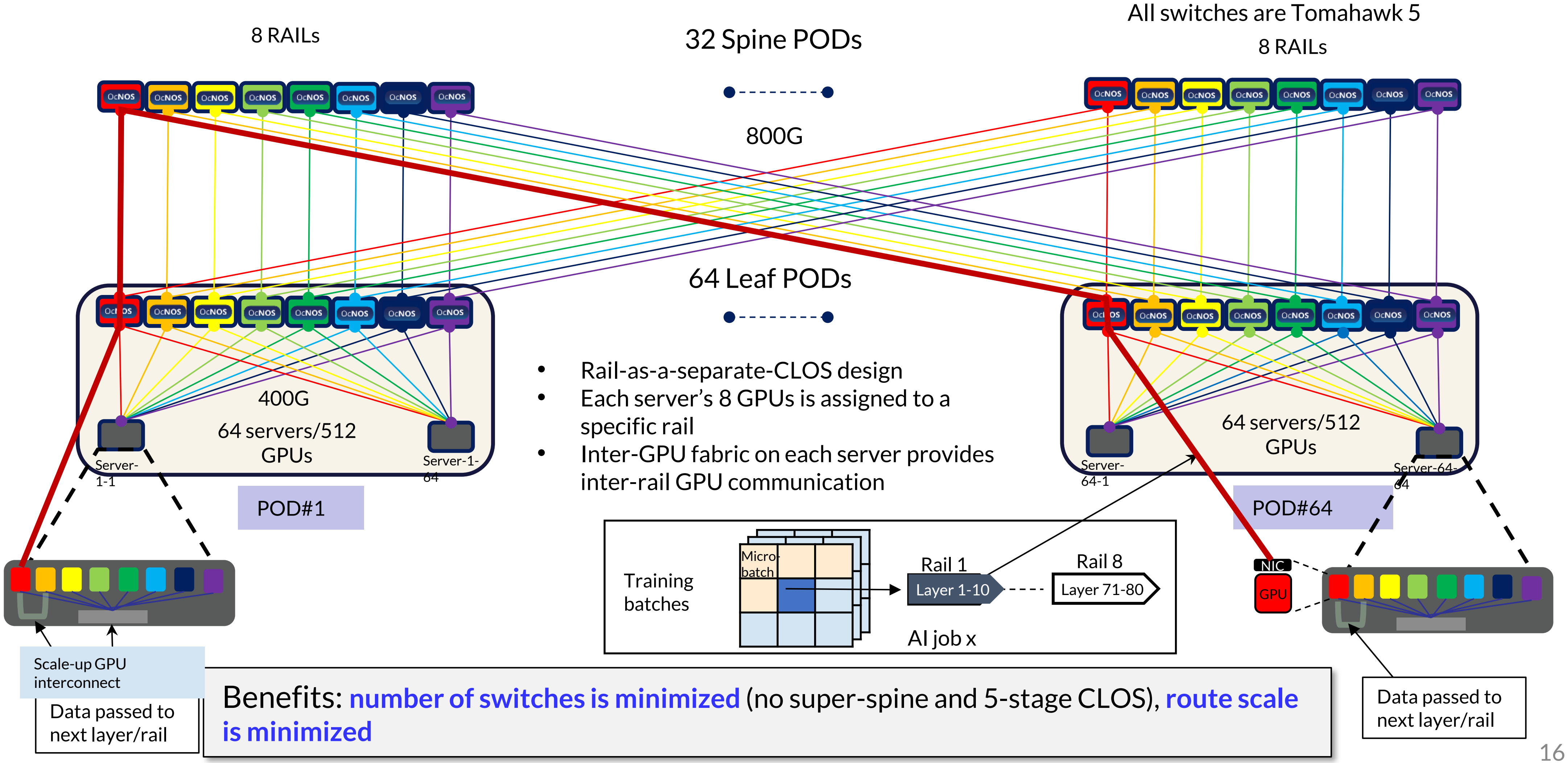
Out-of-Band (OOB) Device Management

- Layer 2 and layer 3 feature support
- Redundancy and availability
- Access control and security

Scaling the AI Fabric – A Modular Approach

Rail Architecture – 32k GPU Example

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**



- Rail-as-a-separate-CLOS design
- Each server's 8 GPUs is assigned to a specific rail
- Inter-GPU fabric on each server provides inter-rail GPU communication

Benefits: **number of switches is minimized** (no super-spine and 5-stage CLOS), **route scale is minimized**

Route Scale for 32K GPU AI Fabric

What each leaf and spine must hold

	Leaf	Spine
Advertises	64 /32 local GPU IPs	Own loopback
Receives	**4,064 routes** (from all other leaves of the same rail + loopbacks of spine peers)	**4,160 routes** (local GPU IPs on each of 64 leaf neighbors + loopbacks of leaf neighbor)
Adj-RIB-In	**129,056 paths** (32-way ECMP per remote /32, single path for each spine peer loopback)	**4,160 paths** (single path per each leaf neighbor for its local GPU IPs, single path for each leaf neighbor loopback)
FIB	**4,064 routes** (one shared 32-way ECMP next-hop group comprised of spine peers reused by all remote prefixes)	**4,160 routes** (single next-hop switch for each route)
Re-advertises	None	**4,160 /32s** to 64 leaf neighbors

Note: OcNOS supports BGP peer group, BGP graceful restart and HW based fast link failover to reduce BGP updates resulted from events like link failure and BGP control plane restart

AI/ML Workload and Management/Control Traffic Types

Traffic Type	Typical Volume	Frequency	Purpose	Characteristics	Transport Fabric
1. Gradient Synchronization / All-Reduce	Very High	Per step or iteration	Sync model parameters	Long-lived, high-throughput, latency sensitive	GPU <-> GPU
2. Activation and Feature Map Data	Very High	Per step or iteration	Exchange intermediate tensors during model/ pipeline parallelism	Long-lived, high-throughput, latency sensitive	GPU <-> GPU
3. Checkpointing	Moderate to High	Periodic (every N minutes/steps)	Save model snapshots	Large, bursty file transfers	CPU <-> storage
4. Bulk I/O	Moderate to High	Periodic (every N minutes/steps)	Load training data / write results	Large-volume, often parallel	CPU <-> storage
5. Control Messaging	Low	Continuous, small bursts	Job coordination, sync	Small packets, periodic or bursty	worker nodes <-> monitoring/ management system(s)
6. Logs / Telemetry	Very Low	Steady or bursty	Record metrics or events	Low rate, asynchronous	worker nodes <-> monitoring/ management system(s)

AI Fabric (rows 1–2)

Frontend Network/ Storage Fabric (rows 3–6)

QoS setting Example for Each AI/ML Traffic Type

AI Fabric

ETS Setting (DSCP through Scheduling)

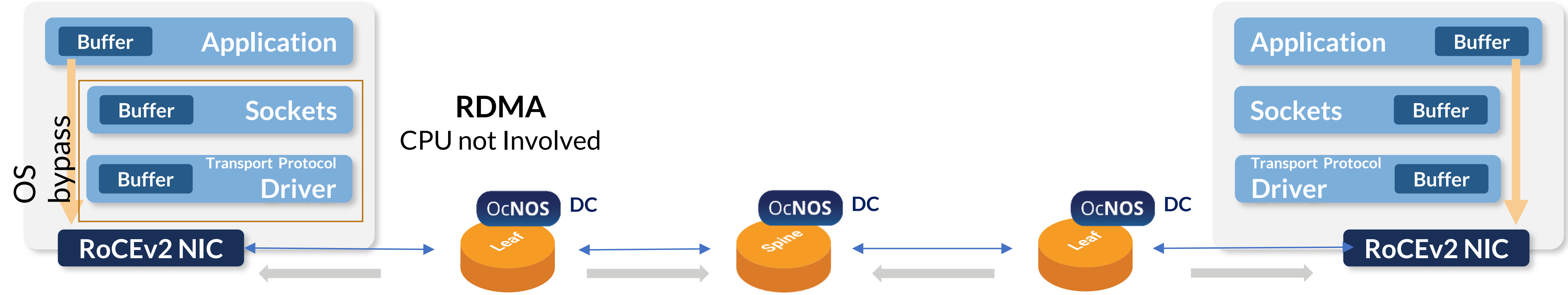
Traffic Type	DSCP	802.1p PCP	Forwarding Class	Queue Number	Scheduling (within class)	PFC Enabled	ECN Enabled
CNP	46 (EF)	5 or 7	CNP-LOSSLESS	7 (Highest Priority)	Strict Priority (SP)	YES	(N/A - source of ECN signal)
RoCEv2 Data	26 (AF31)	3 or 4	ROCE-LOSSLESS	4 (High Priority)	WRR / SP (if truly single class)	YES	YES

Frontend Network/ Storage Fabric

ETS Setting (DSCP through Scheduling)

Traffic Type	DSCP	802.1p PCP	Forwarding Class	Queue Number	Scheduling (within class)	PFC Enabled	ECN Enabled
CNP	46 (EF)	7	CNP-LOSSLESS	7 (Higher SP)	Strict Priority (SP)	YES	(N/A – source of ECN signal)
AI/ML Storage I/O (RoCEv2)	26 (AF31)	4	STORAGE-ROCE	4	WRR (min bandwidth)	YES	YES
AI/ML Management/Control	46 (EF)	5	AI-CONTROL	5 (Highest SP)	Strict Priority (SP)	NO	YES
Logs / Telemetry	0 (Default)	0	BEST-EFFORT	0	WRR / DRR	NO	YES

Enabling Hop-by-Hop Lossless Transport via DCBX



Leaf pushing PFC and ETS settings and application priority mapping to host via DCBX

PFC:
Which priorities have PFC enabled (e.g. priority 3)

ETS:
Priority-to-TC mapping, bandwidth allocation per TC

Application Priority:
Map UDP port to priority (e.g. request host to map RoCEv2 → CoS 3)

Leaf pushing PFC and ETS settings to spine via DCBX

PFC:
Which priorities have PFC enabled (e.g. priority 3)

ETS:
Priority-to-TC mapping, bandwidth allocation per TC

Leaf pushing PFC and ETS settings to spine via DCBX

PFC:
Which priorities have PFC enabled (e.g. priority 3)

ETS:
Priority-to-TC mapping, bandwidth allocation per TC

Leaf pushing PFC and ETS settings and application priority mapping to host via DCBX

PFC:
Which priorities have PFC enabled (e.g. priority 3)

ETS:
Priority-to-TC mapping, bandwidth allocation per TC

Application Priority:
Map UDP port to priority (e.g. request host to map RoCEv2 → CoS 3)

DCBX ensures every switch and NIC along the path reserves consistent bandwidth and maps CoS values to queues the same way

Dynamic Load Balancing (DLB)

- Traditional ECMP hash-based link selection is fixed throughout the flow even when port load and port queue size change

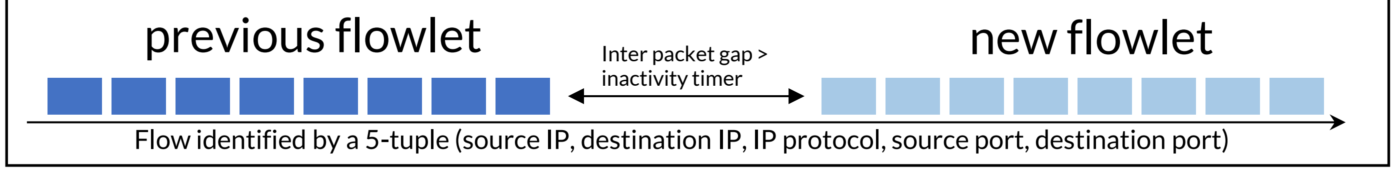
  (destination prefix, packet hash) → output link/nexthop

- DLB dynamically selects output member link in an ECMP group (i.e. group of next hops for a destination prefix) for a flow

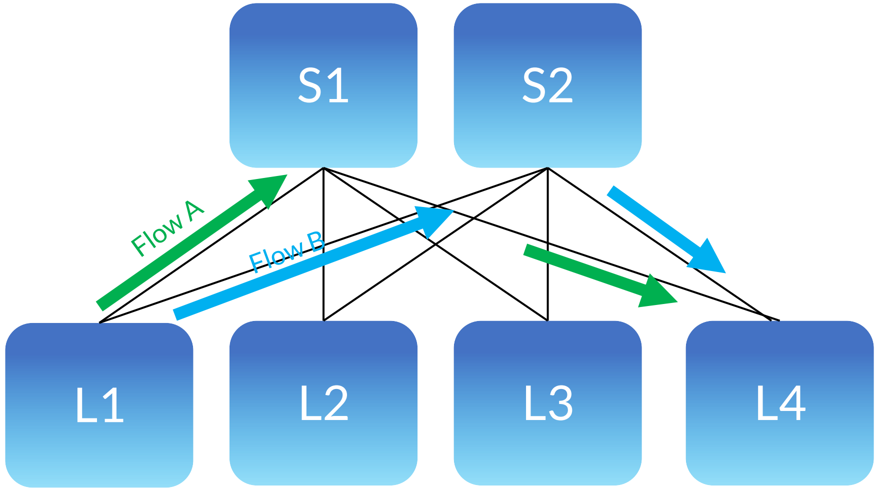
  (destination prefix, dynamic index) → output link/nexthop

  Dynamically change based on the following conditions
  Link utilization
  Queue depth / buffer pressure
  Packet drops
  LAG/ECMP member availability

- Change of output link for a flow only takes effect for new flowlet to preserve in-order delivery



- Reactive Path Rebalancing (RPR) mode of DLB probabilistically reassigns a continuous incoming stream to a better quality (less loaded) egress member if quality is good by a configured delta



Network Observability – Example Sensor Paths

Per leaf

Metric category	Paths per interface	Interfaces	Total paths
Interface counters	1	80 (64 downlinks+16 uplinks)	80
Optics metrics	1	80	80
Queue stats (per-queue)	8	80	640
Buffer depth (per-port)	1	80	80
BGP neighbor state (1 per uplink port)	1	16	32
Grand Total			912 paths

Per spine

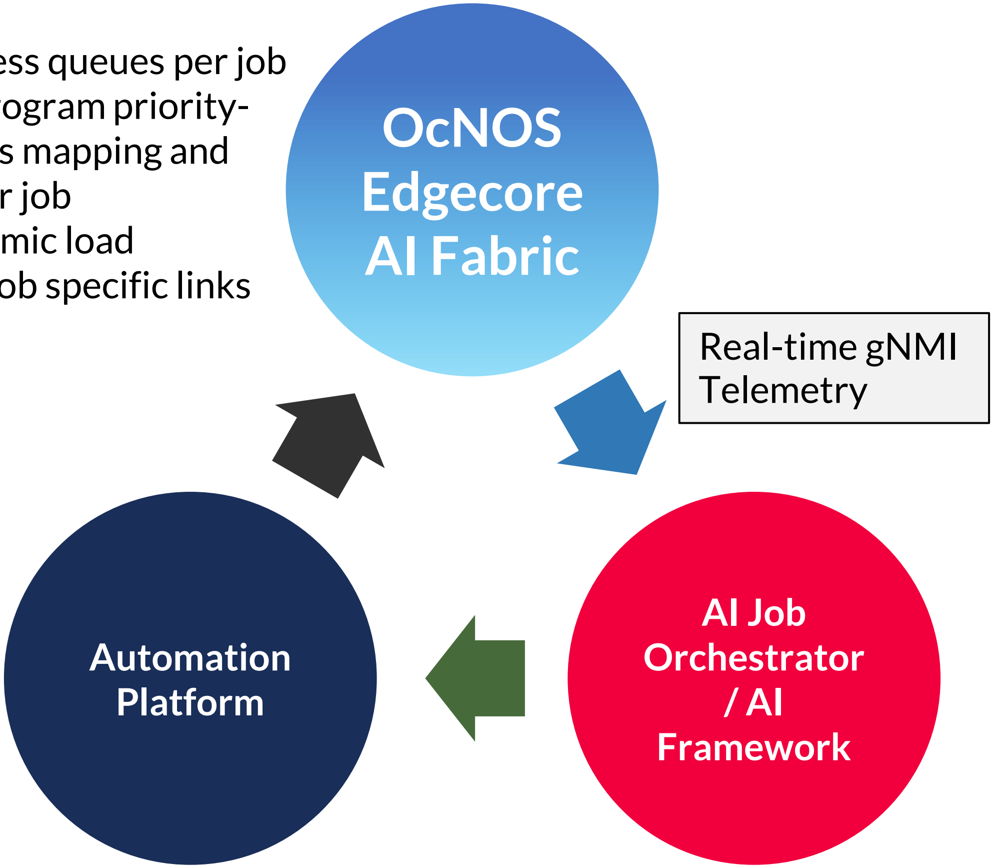
Metric category	Paths per interface	Interfaces	Total paths
Interface counters	1	32	32
Optics metrics	9 (1 module + 8 per-lane)	32	288
Queue stats (per-queue)	8	32	256
Buffer depth (per-queue)	8	32	256
BGP neighbor state (1 per port)	1	32	64
Grand Total			896 paths

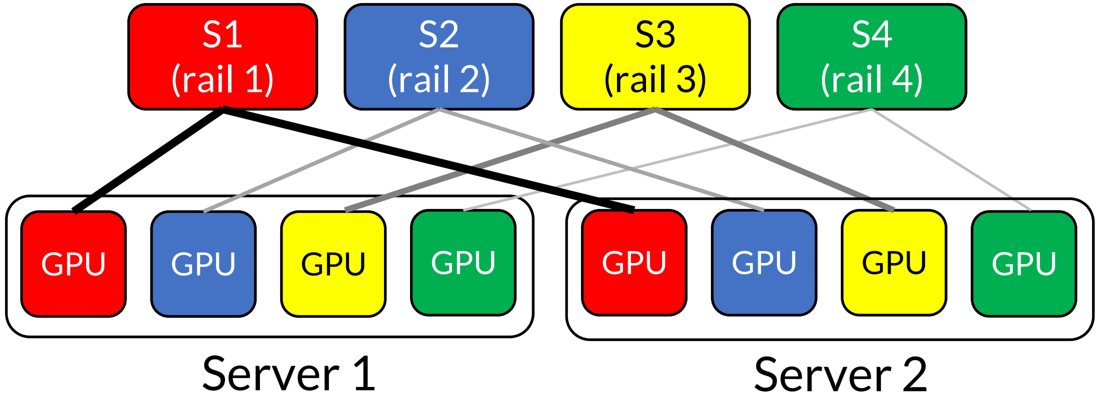
Optional paths (for more granularity)

- **Per-lane optics metrics** for media side (8 lanes/port) in addition to per-port module level optics metrics
- **Per-queue buffer depth**
- Optional per-lane optics can be enabled only on spines to monitor fiber links
- Although per-queue buffer depth monitoring is more critical on leaves, leaves already have many sensor paths

Closed-loop OcNOS Fabic and AI Workload Orchestrator Integration

- Reserve lossless queues per job
- Program/reprogram priority-to-traffic-class mapping and bandwidth per job
- Activate dynamic load balancing on job specific links





Rail	Link Utilization	Buffer Depth	Action
1	90%	High	Offload gradient aggregation to rail 2 GPUs temporarily
2	30%	Low	Additionally perform Layer 1's gradient aggregation temporarily
3	50%	Medium	Normal scheduling
4	10%	Low	Start next micro-batch from rail 4 instead of rail 1 (i.e. rail 4 -> Layer 1)

Driving GPU Efficiency with Efficient Networking

Key Design Considerations

- Usage of RDMA – High Bandwidth flows and Utilization
- Usage of Low Jitter Tolerance
- Design Network for Non-Blocking paths with High Bandwidth
- Predictive Performance

AI Factories
- Single or few workloads
- Extremely Large Models
- One/Few users

→ InfiniBand

AI Cloud
- Multi Tenant
- Variety of Workloads
- Less complex jobs

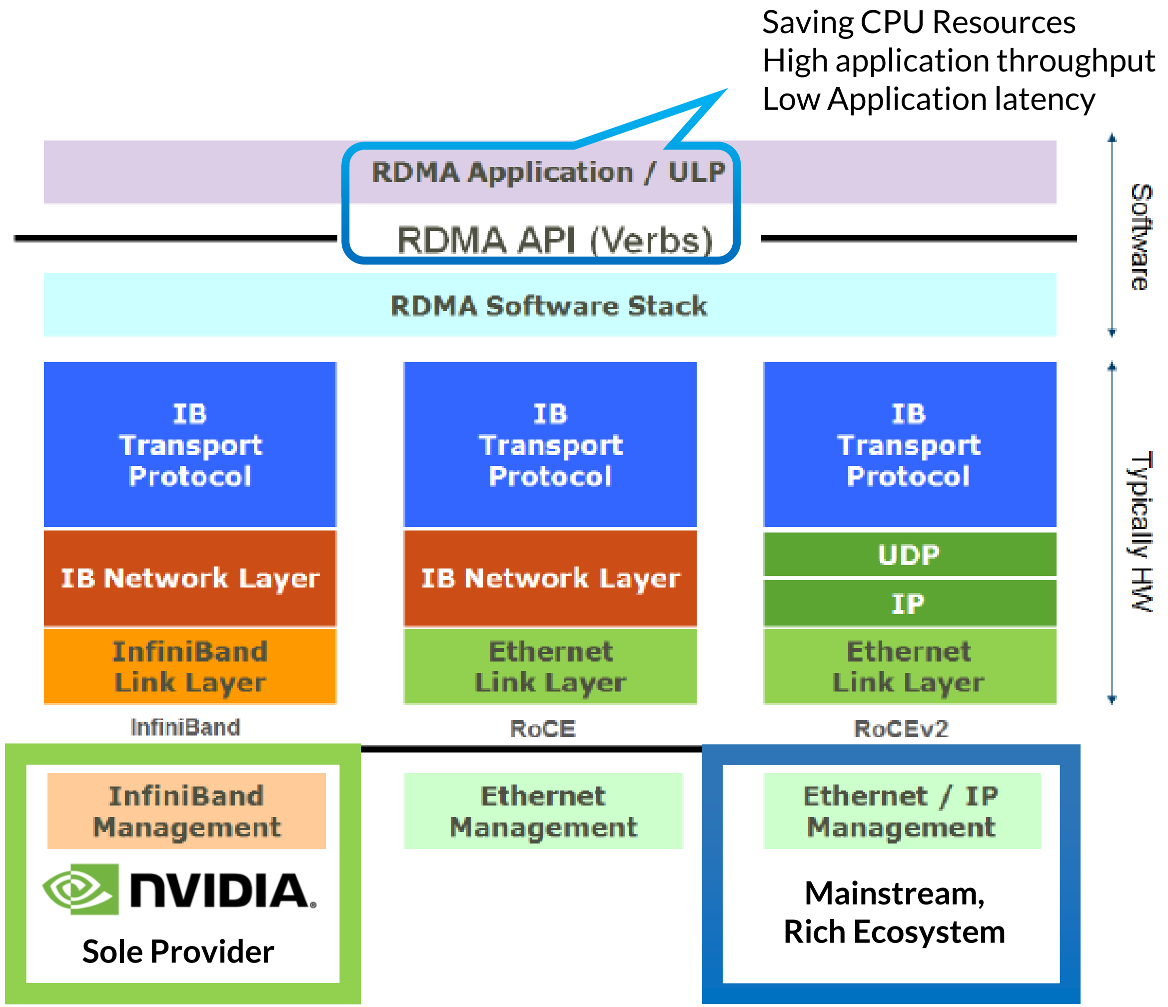
→ Ethernet

Increasing AI workloads and Large-Scale Gen AI training has shown standard Ethernet to be slow.

ROCE (RDMA over Converged Ethernet)

- ROCEv2 (RDMA over Converged Ethernet) uses IB packet header and encapsulates with UDP header
- Efficient data transfer where the OS is bypassed and enables fast access to remote data
- Supports message passing, sockets, and storage protocols
- Support by all major operating systems
- ROCE is an Open Source and a formal IBTA (Infiniband Trade Association) standard

RDMA, RoCEv2 and UEC



BROADCOM AMD

Ultra Ethernet Consortium

UEC is enhancing RoCEv2 drawbacks and improve in many layers that ideal for mixing workloads

Driving GPU Efficiency with Efficient Networking

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**

Dimension	InfiniBand (IB)	iWARP	RoCE v2
Specification / Release	IBTA Spec 1.0 (2000)	IETF RDDP (2003)	IBTA Annex 17 (2014)
Wire format	Native IB frame	IB frame carried in **TCP/IP**	IB frame carried in **UDP/IP** (UDP 4791) *[RoCE v1 was L2 Hdr, with v2 it supports L3 is routeable with ECMP and DLB]*
Layer reach	Proprietary L1/L2 switched fabric	Routable Layer-3; crosses subnets	Routable Layer-3; crosses subnets
Switching & control	IB switches + Subnet Manager	Standard Ethernet/IP switching & routing; **no PFC required**	Lossless Ethernet switches with **PFC**
Lossless guarantee	Built-in credit-based flow control	**No lossless fabric needed**; TCP is reliable & in-order, tolerates loss/retransmit	PFC
Congestion control	IB-CC (link-level credits)	**TCP window/ECN** (Reno/CUBIC/DCTCP, etc.) — **window-based**	DCQCN most common (Reactive CC w/ ECN)
CPU involvement	Near-zero copy RDMA; minimal CPU	Same as IB	Same as IB
Scalability limits	Tens of thousands of nodes per fabric (topology-dependent)	Scales with IP routing—data-center or multi-DC	Scales with IP routing—data-center or multi-DC
Typical deployments	HPC supercomputers, AI clusters that value lowest latency	-	Large AI clusters, cloud RDMA services, multi-site fabrics
Strengths	Lowest latency, mature HPC software stack	**No PFC required**; works on standard IP networks; resilient to loss/reorder (TCP)	L3 routeability, coexists with traditional IP, flexible
Weaknesses	Requires dedicated hardware & management; higher CapEx	Higher TCP/IP latency, small ecosystem	Adds UDP/IP overhead; still needs PFC/ECN tuning for true lossless ness *[Go-Back N Retransmission – requires lossless and In-order delivery]*

Latency: IB<RoCEv2<i-WARP

UEC – Building Blocks

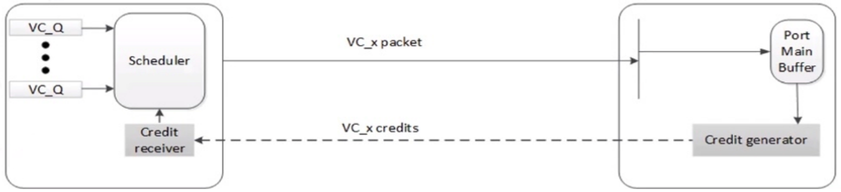
AI FH – Allows for smaller packet size required and sufficient for Fabric communication

LLR – Mechanism to allow for link partners to request retransmission upon FEC failures

CBFC – Targets to provide link layer level lossless operations for VC

Credit Based Flow Control

- Targets to provide link-layer lossless operations for each lossless VC over links between two peer devices to enable lossless buffering in Rx devices.
  - Virtual Channel represents parts of port traffic and can be flexibly configured as lossless or lossy
- At receiver end
  - Pre-allocate for a port in Rx device headroom buffer space for lossless traffic
  - Generate credit generation based on port buffer availability in receiver
  - Advertise credits to Tx device
- CBFC Credit messages are used for transmission of credits from receiver to sender
- The sender keeps track of the available credits and its scheduler is allowed to schedule a VC queue only if it has credits



AI Fabric Header for Routed Flows

- **Reduced IPG (Inter-Packet Gap)**
  - 1B to 8B based on packet alignment (vs. Ethernet standard 12B)
- **Optimized Fabric Header: Fields [] Are Optional**

DA (6B)	SA (6B)	[VLAN (4B)]	AFH Ethtype (2B)	[AFH_Extension (0B - 4B)]

- **Retains Ethernet-like structure for coexistence with IPv4/v6**
  - Minimize overhead for small packets by combining L2 (MAC) and L3 (IP) headers
  - Addressing is overlaid on SA/DA, usable for single-tier (eg scale-up) and multi-tier (eg scale-out) fabrics
  - Ethertype indicates the presence of a AFH_Hdr or a standard header such as IPv4/v6
  - VLAN tag is optional (eg for security)
  - AFH_Hdr includes fields commonly used for routing (hop count, traffic class, congestion, etc.)
  - Allows for coexistence with standard IPv4/v6 packets and interop with standard MACs
  - Supports ECN and other fabric notifications
- **AFH format is user-defined**
  - AFH Address space (# address bits) can be defined by system designer
  - AFH EtherType determines AFH Extension Size, which can be 0, 2, 3, or 4 bytes
  - TU can simultaneously support two different AFH formats with different AFH Ethertypes

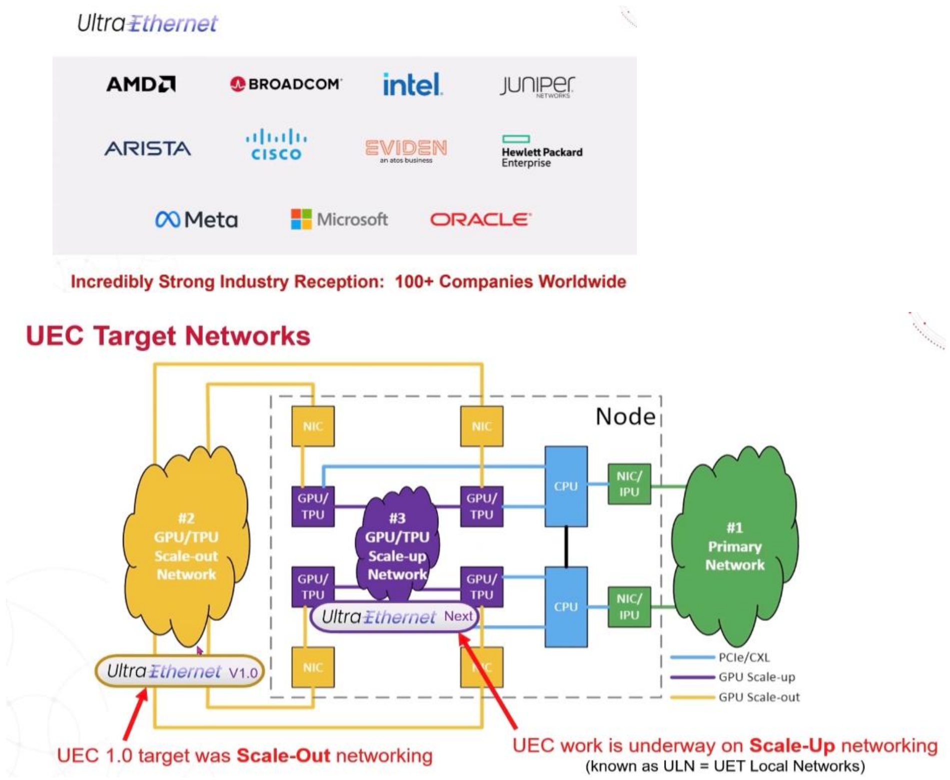
NOTE: AFH was developed prior to UEC, and while AFH and the UEC's Unified Forwarding Header (UFH) have some similarities, they are distinct and not equivalent

Link-Layer Retry Architecture

- **LLR Scope**
  - LLR retransmits packets due to FEC/CRC errors on a full duplex Ethernet link
    - Much faster recovery than end-to-end "TCP level" retransmission
  - LLR does not protect against dropped packets due to buffer congestion
- **LLR Architecture**
  - Ethernet extensions:
    - A sequence number is placed in each packet's preamble.
    - Data receiver sends ACK/NACK messages (8B Control Ordered Set) for correctly or incorrectly received packets.
  - MAC TX contains replay buffer to support retransmission upon receiving NACK.
    - After receiving NACK, packet stream replays from lost or corrupted packet
    - It is a Go-back-N packet-based protocol.
  - Initialization Sequence
    - Handshake between link partners to reset starting sequence numbers before sending traffic

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**

1



2

Energy-efficient ASICs and optics reduce power per Gbps, aligning with green Data Center narrative

- **800G DPO$ to 800G LPO$ = up to 50% saving**
- **Generic 800G 2xDR4 power consumption is 14.5W while 800G LPO is typical 7.5W, reflecting 48%+ saving**

3

Cut Down TCO: Electrical Cable for Short-Reach Connection

- Direct Attach Copper (DAC)
- Active Electrical Cable (AEC)

ET7502-DAC-xM

	400G DR Optical Transceiver	400G AEC Cable	400G DAC Cable
Max Length	500 meters	7 meters	3 meters
Power consumption	10 watt	5 watt	0.3 watt

Max length and power consumption for 400G connectivity

4

Cut Down TCO: LPO Module + Fiber Cable

- Linear-drive Pluggable Optics (LPO)
- After eliminating some DSP modules
  - Lower power consumption
  - Lower latency
  - Need tuning per model per port

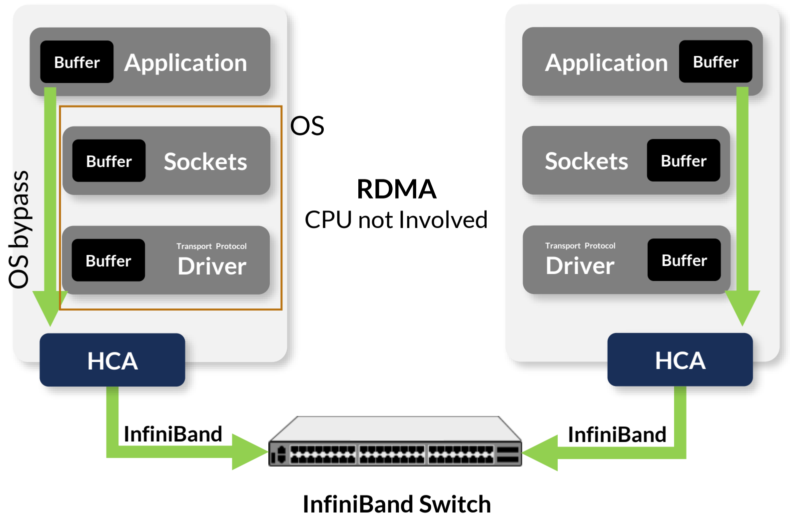
	800G DR Optical Transceiver	800G LPO Module
Latency	50~70 nanosecond	Less than 10 nanosecond
Power consumption	Typical 14.5 watt	Typical 7.5 watt

Latency and power consumption for 800G connectivity

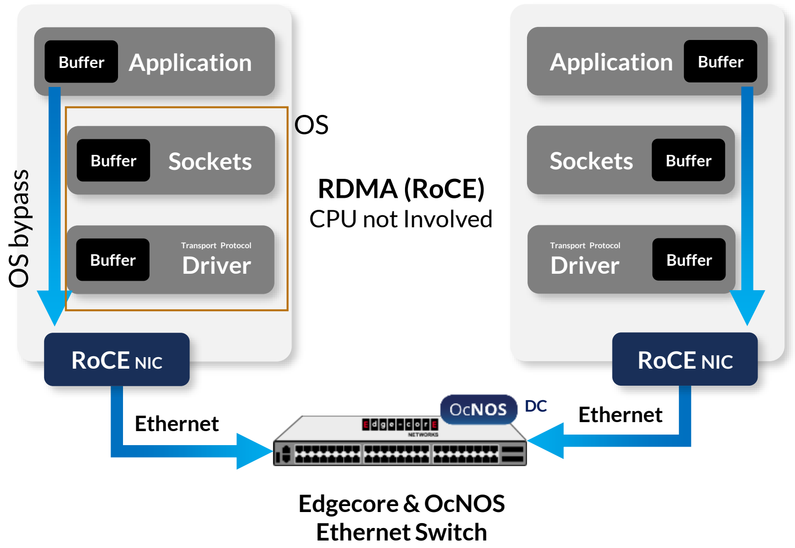
OcNOS AI/ML Solution Customer Benefits

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**

AI/ML Networks with RDMA over
InfiniBand



AI/ML Networks with RDMA over
Converged Ethernet (RoCE)



Choose **Edgecore and OcNOS Ethernet Fabric** for you AI Cluster if you need:

Ubiquity and Interoperability
Seamless integration with existing network infrastructure + modern AI networking stack

Mature Open Ecosystem
Widest and rapidly evolving ecosystem of compatible AI switches and optics with global support

Superior TCO
Ready to deploy open networking solution with perpetual licensing, and leading support pricing

WEBINAR:

OPEN ETHERNET NETWORKING FOR MODERN AI/ML WORKLOADS BUILDING THE AI FACTORY

Q&A

ALAN HUANG
Senior Product Manager
ipinfusion

SUJAY GUPTA
Senior Solutions Manager
Edge-corE NETWORKS

WEBINAR:

OPEN ETHERNET NETWORKING FOR MODERN AI/ML WORKLOADS BUILDING THE AI FACTORY

THANK YOU

ipinfusion™

www.ipinfusion.com

Edge-corE® NETWORKS

www.edge-core.com

Switch – Optics – NIC

SEP. 10 | WEBINAR: **A BLUEPRINT FOR BUILDING THE AI FACTORY**