The Networking in the ‘AI’ Era

Davide Bassani

Systems Engineer

davide@arista.com

Agenda

1. Who is Arista Networks?
2. Deep Learning Networking Requirements
3. Networking Requirements for Inter-GPU Traffic
4. Architectures for AI Network Fabrics
5. A Quick Glance to the Future of AI Ethernet
6. Conclusions

1 di 6

WHO IS ARISTA NETWORKS?

Introducing Arista Networks

28% Data Center Ethernet Mkt Share
40/50% 100G/400G Speed Mkt Share

30+ Million Ports Shipped

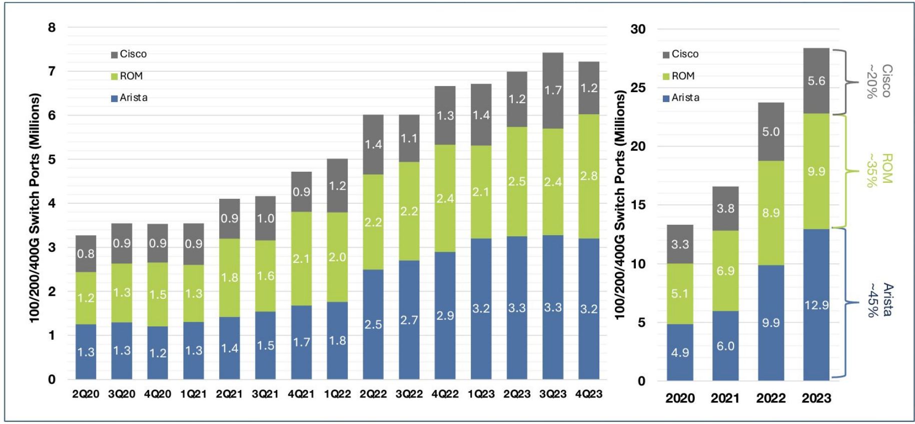
9000+ Customers

54 New Products (last year)

1 Operating System

Arista's Market Leadership in 100G/200G/400G

Data Center High Speed Ethernet Port Analysis



Source: Crehan Ethernet Switch Data Center Total Vendor Tables – 4Q'23

Introducing Arista Networks

$5.9B
2023 REVENUE

22.2%
5-YEAR CAGR THROUGH FY'23

62.2%
2023 GROSS MARGIN

44.1%
LTM OPERATING MARGIN

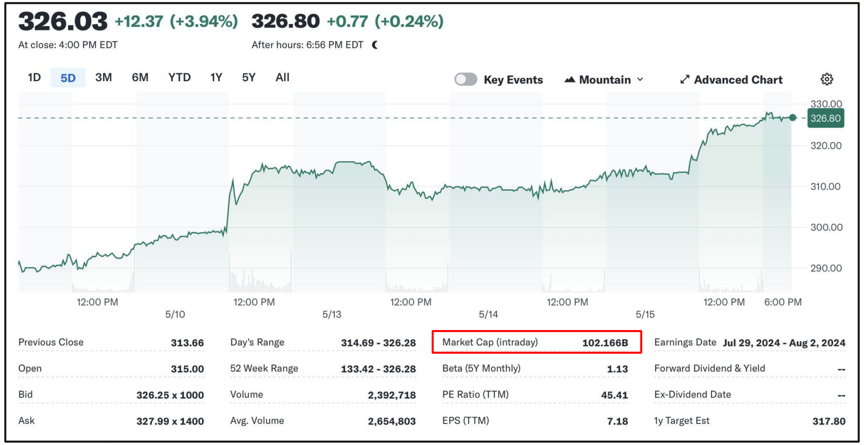
IPO 2014
June 6th

S&P 500
Added in 2018

5 Year TAM
$37B 2023 to $60B 2027

NYSE

Ticker: ANET



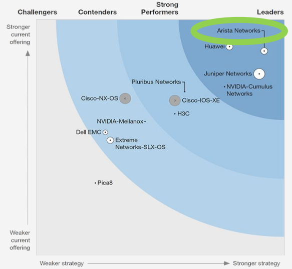
326.03 +12.37 (+3.94%) **326.80** +0.77 (+0.24%)

At close: 4:00 PM EDT After hours: 6:56 PM EDT

Previous Close	313.66	Day's Range	314.69 - 326.28	Market Cap (intraday)	102.166B	Earnings Date	Jul 29, 2024 - Aug 2, 2024
Open	315.00	52 Week Range	133.42 - 326.28	Beta (5Y Monthly)	1.13	Forward Dividend & Yield	--
Bid	326.25 x 1000	Volume	2,392,718	PE Ratio (TTM)	45.41	Ex-Dividend Date	--
Ask	327.99 x 1400	Avg. Volume	2,654,803	EPS (TTM)	7.18	1y Target Est	317.80

Industry Recognition

Arista Networks recognized as a leader with top score in strategy category in The Forrester Wave™: Open Programmable Switches for Business wide SDN, Q3 2020



Acquisitions

mojo
WiFi Solutions
August 2018

METAMAKO
Ultra Low Latency Switch
September 2018

big switch networks
Network monitoring and SDN pioneer
February 2020

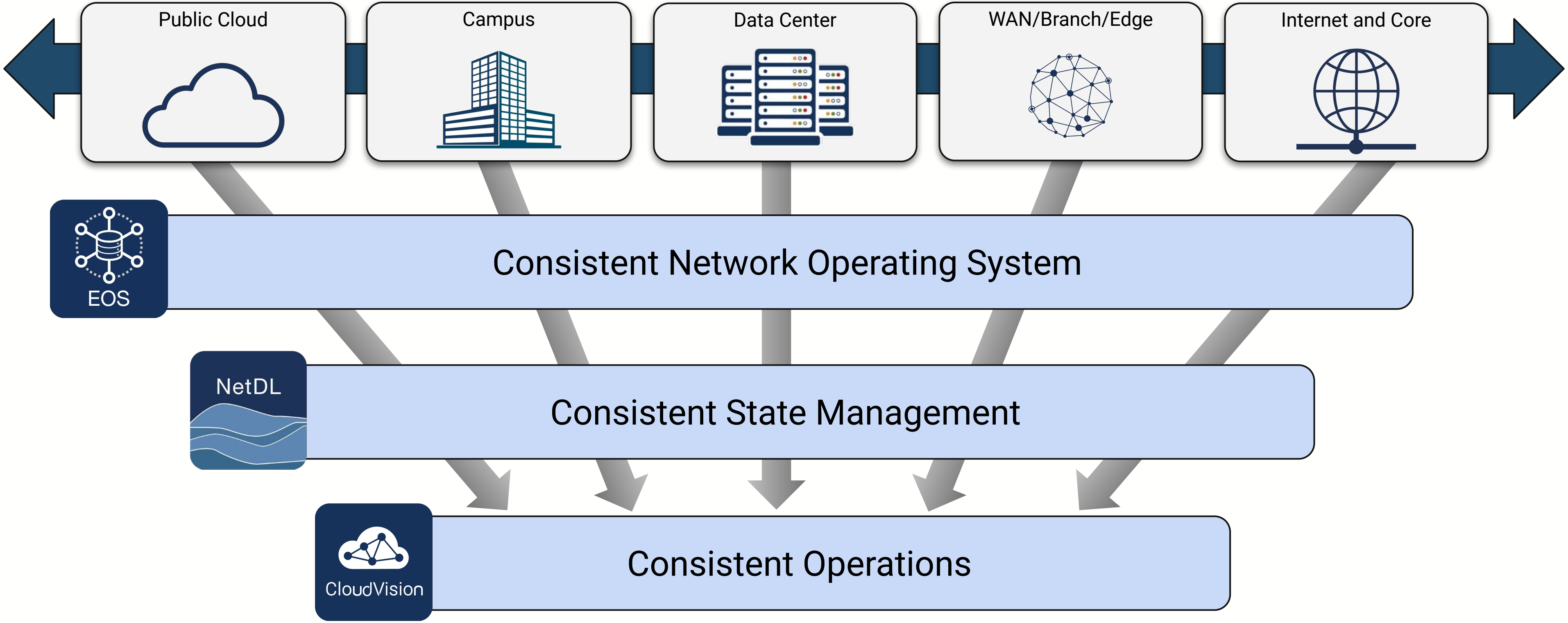
AWAKE
Network Detection & Response
September 2020

untangle
Edge Threat Management
March 2022

Pluribus NETWORKS
Software Defined Networking
August 2022

ARISTA

Software Driven Cloud Networking



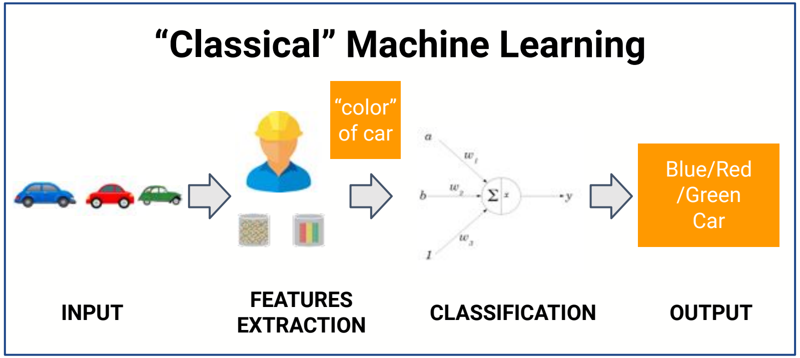
Consistent Engineering and Operations Across All Network Domains

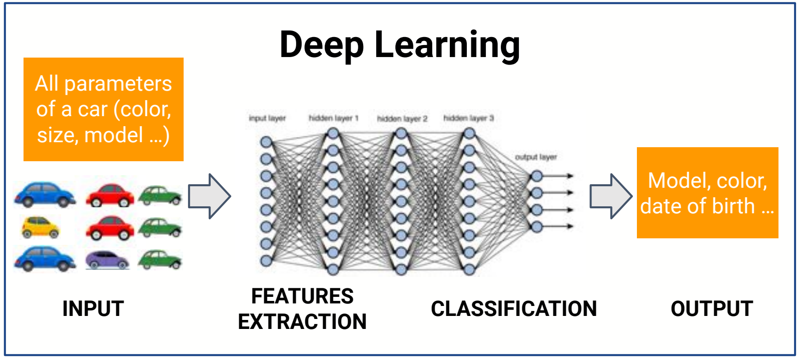
ARISTA

2 di 6

DEEP LEARNING NETWORKING REQUIREMENTS

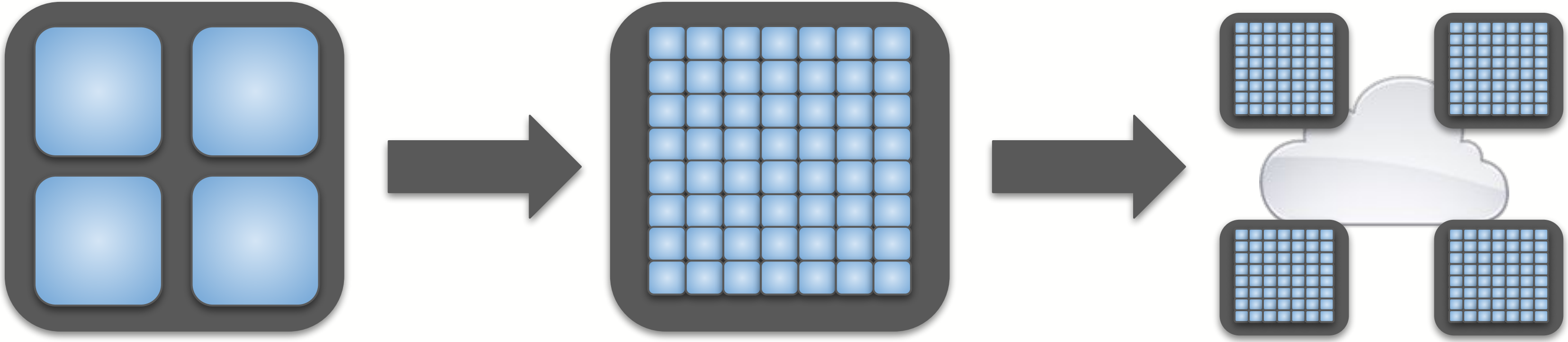
Machine Learning vs Deep Learning





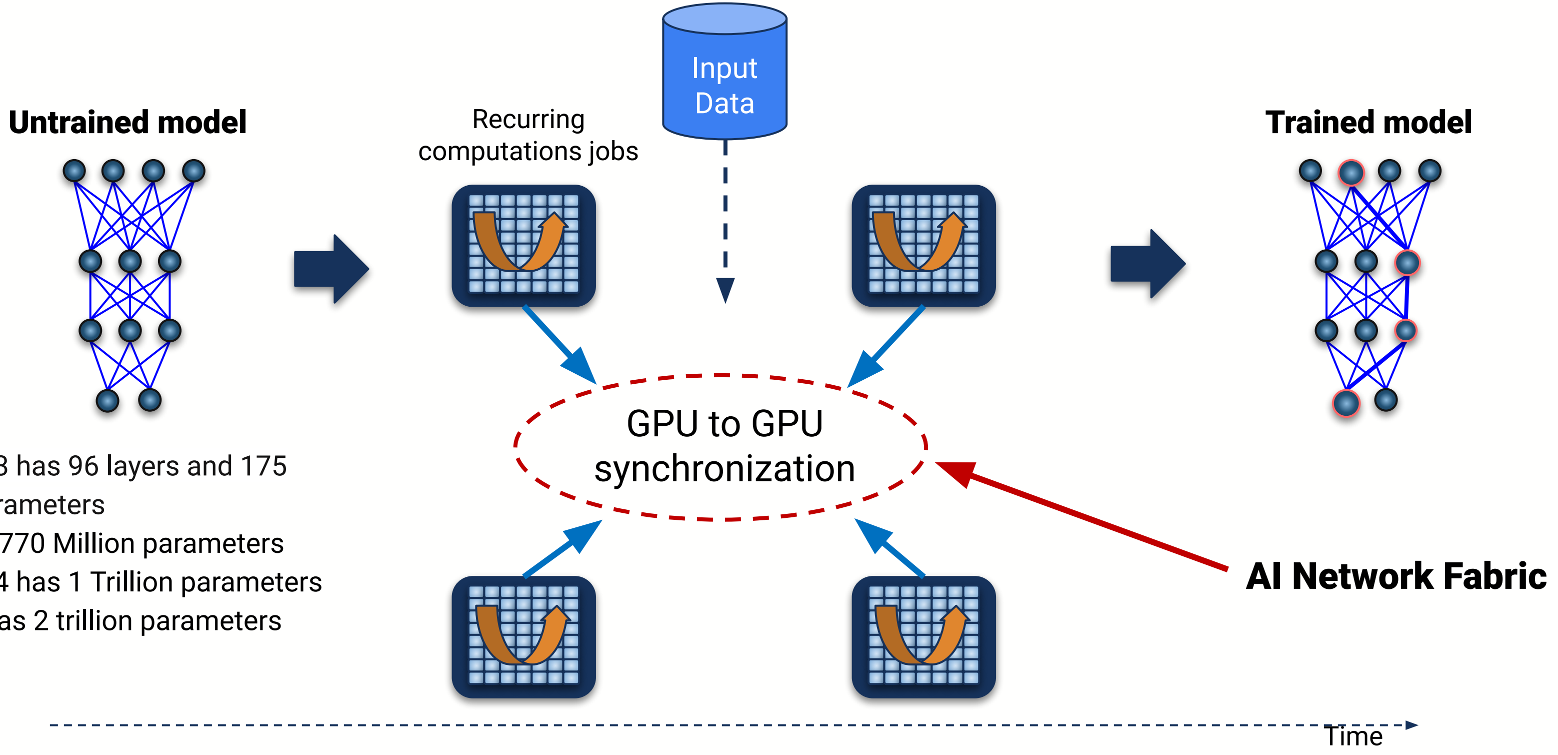
	Machine Learning	Deep Learning
# Input Values	Low	High
Feature Extraction	Manual	Automatic
Classification	Simple	Complex
Output	Raw	Granular
Data Exchanged	"Low"	Huge

The Usual Suspects to Manage AI



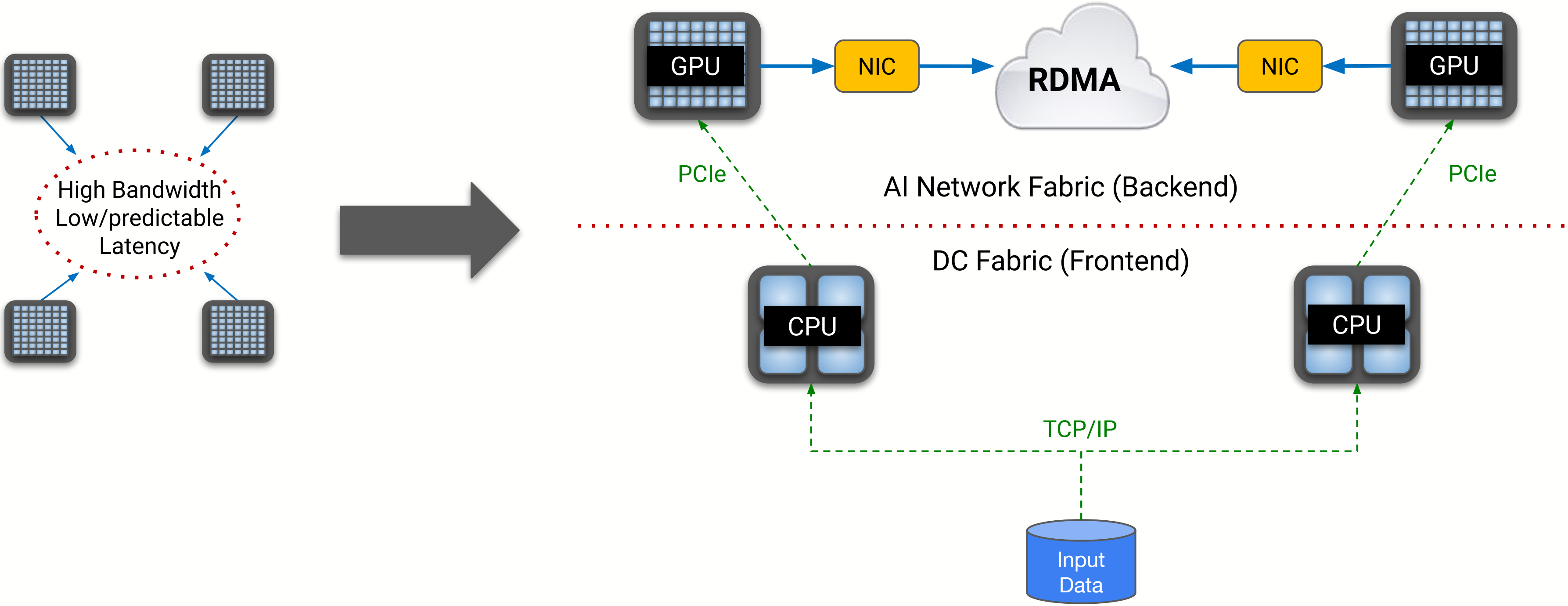
CPU	GPU	GPU Clusters
1 – 16 cores	100s – 1000s cores	100s – x0k GPUs
Optimized for serial tasks	Optimized for parallel tasks	Scale-out GPU Network for AI Workloads

Distributed Training a Complex Model



- ➔ ChatGPT3 has 96 layers and 175 billion parameters
- ➔ Bard has 770 Million parameters
- ➔ ChatGPT4 has 1 Trillion parameters
- ➔ Llama2 has 2 trillion parameters

GPU to GPU Interconnect



High Bandwidth
Low/predictable
Latency

GPU

NIC

RDMA

NIC

GPU

PCIe

AI Network Fabric (Backend)

PCIe

DC Fabric (Frontend)

CPU

CPU

TCP/IP

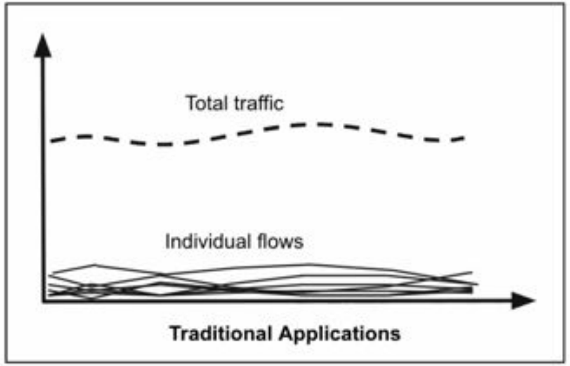
Input
Data

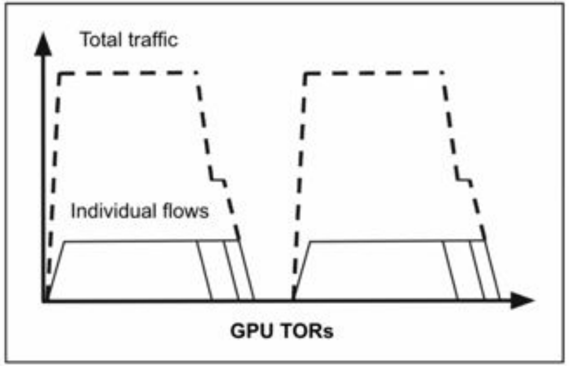
3 di 6

NETWORKING REQUIREMENTS FOR INTER-GPU TRAFFIC

What's Different About AI Workloads?

- AI workloads are using "collective" communications for parallel computation
- Development frameworks are using RDMA approach to bypass kernel for more efficiency
- Main traffic characteristics:
  - Tight time synchronisation between bursty traffic flows
  - Small number of large sized flows (<10 per nic)
  - Very low level of entropy
  - Short periodic burst of network activity followed by high computation processes
- Highly susceptible to collision
- "Slowest" member drives performance!



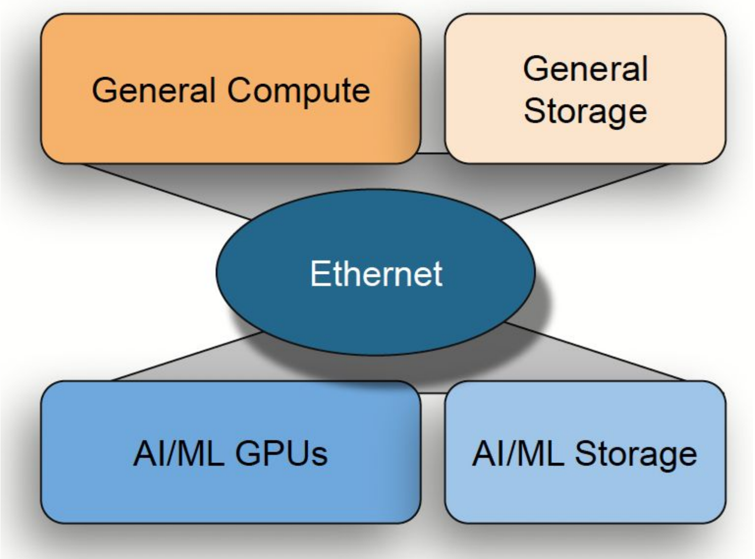


AI training workload are highly coordinated and highly sensitive to delay variations

Networking for AI: What Do You Need?

- A **fast, lossless** network
  - For many forms of communication

    ALL-REDUCE, BROADCAST, ALL-TO-ALL
  - Graceful handling of large/bursty synchronized flows
  - Fast and reliable transfer from host to network (RDMA)
- A network with **consistent latency**
  - Tail latency (high percentile latency) is likely to impact job completion time significantly
- A network **without collision**
  - Distribute equally low-entropy flows along all physical paths
- **Visibility** and **telemetry**
  - To identify bottlenecks in the network or application



AI Workloads require a dedicated high performance lossless network

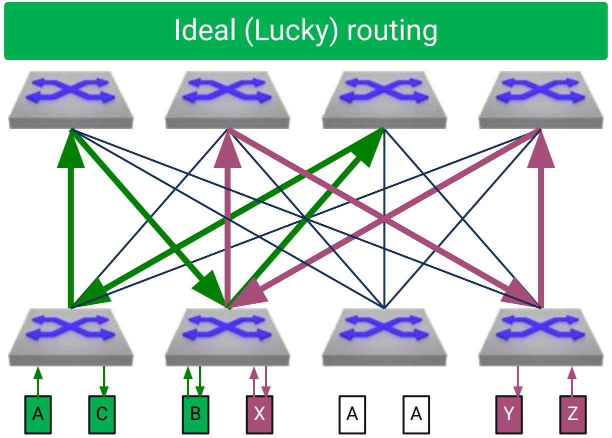
RDMA over Converged Ethernet (RoCE)

- Network protocol that allows RDMA over an Ethernet network
- The second version (RoCEv2) enhances the protocol with UDP/IP header
  - Operations on routed ethernet networks: ubiquitous in large datacenters
  - IP QoS: DSCP or alternatively COS/VLAN PRI - Priority Flow Control (PFC)
  - IP congestion control: the Explicit Congestion Notification (ECN) signal

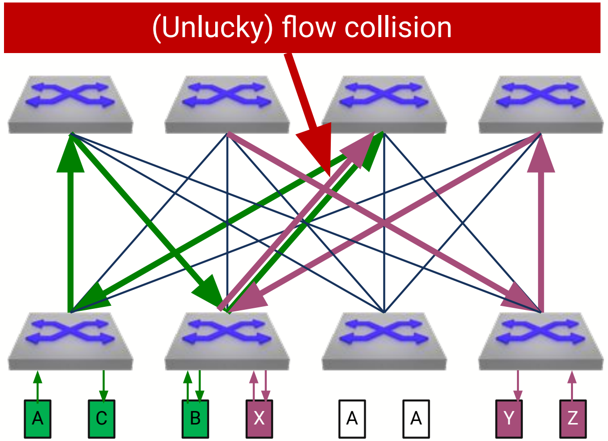
https://www.arista.com/assets/data/pdf/Broadcom-RoCE-Deployment-Guide.pdf

RoCEv2 enables Ethernet infrastructure to behave like a fast, quick and reliable lossless fabric

Flow Collision and Traffic Polarization



Ideal (Lucky) routing



(Unlucky) flow collision

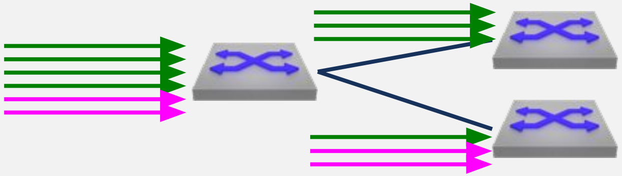
- Load balancing in IP routing is based on “ECMP”
  - Basically it is a Hash of fields in packet header (see next slide)
- But AI clusters don’t drive a significant distribution of parameters
  - Low level of entropy
- Large flows could be polarized on the same links and produce unwanted side effects

How to Avoid Collision / Traffic polarization

- **ECMP hashing**: limited efficiency, especially with Ring (few flows with a lot of data)

- **LB Numbered**: LB Number assigned on each ingress interface so that all traffic arriving on a specific interface is effectively mapped to an egress interface between TOR & Spine (Stitching)

- **Dynamic Load Balancing**: Smart flow distribution based on link utilization
  - ECMP optimization available on selected Arista platforms
  - Flows are allocated to new links based on current utilization, significantly increasing hash performance/efficiency
  - Continuous reevaluation of best links with flows rebalancing



- **Cell based fabric***: Capability to spray a flow among multiple links

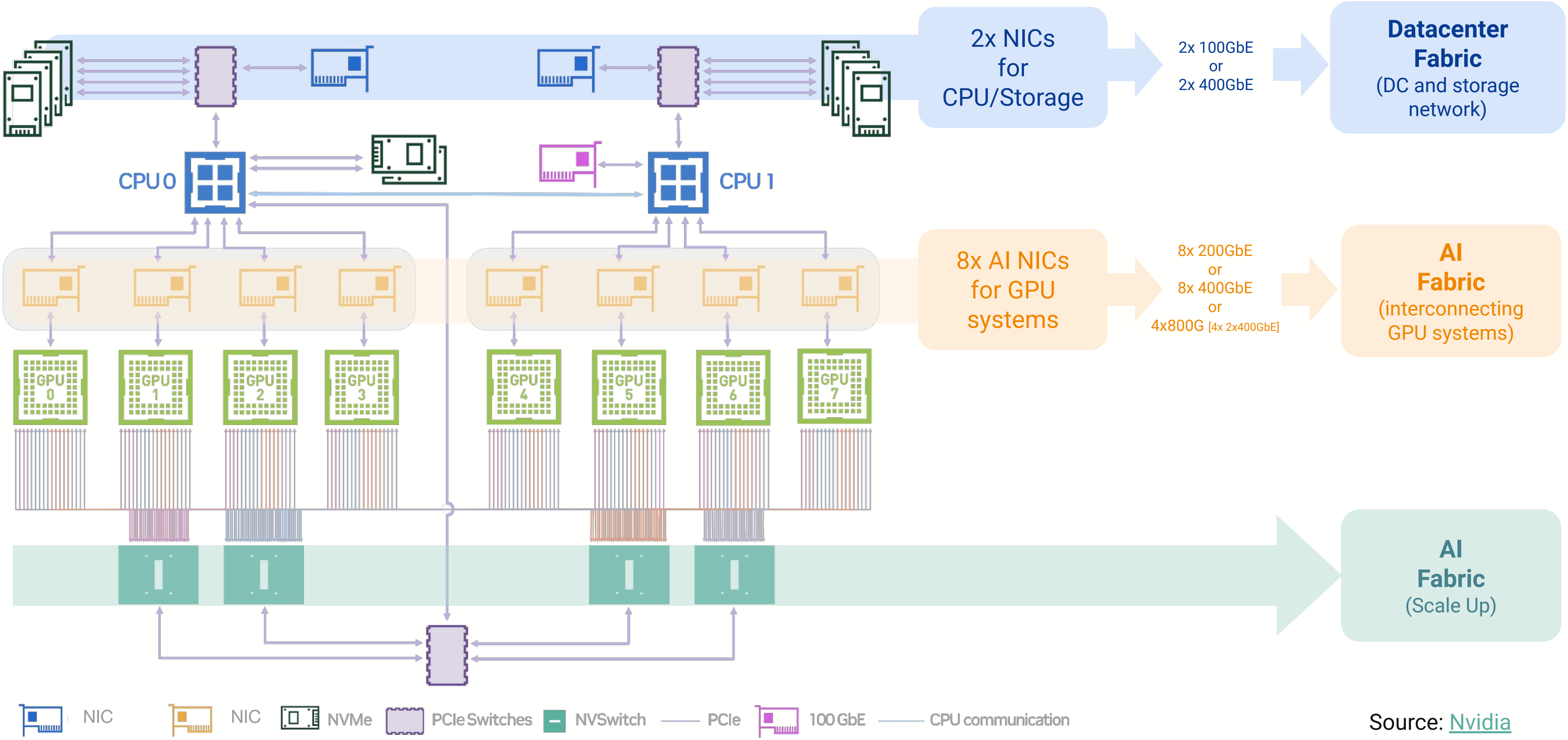
Efficiency

* **coming soon**

4 di 6

ARCHITECTURES FOR AI NETWORK FABRICS

Building Blocks of a Typical GPU System



2x NICs for CPU/Storage

2x 100GbE or 2x 400GbE

Datacenter Fabric (DC and storage network)

8x AI NICs for GPU systems

8x 200GbE or 8x 400GbE or 4x800G [4x 2x400GbE]

AI Fabric (interconnecting GPU systems)

AI Fabric (Scale Up)

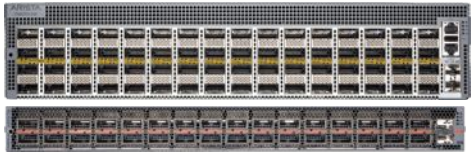
Source: Nvidia

Arista Platforms for AI Networking

7060DX5/6 - 7388X5

Low Latency, Fixed Systems, Single-chip

7388X5 – 64x 400G



7060DX5 – 64 x 400G / 32x 800G

7060DX6 – 64 x 800G

25.6/51.2 Tbps Systems

7800R3

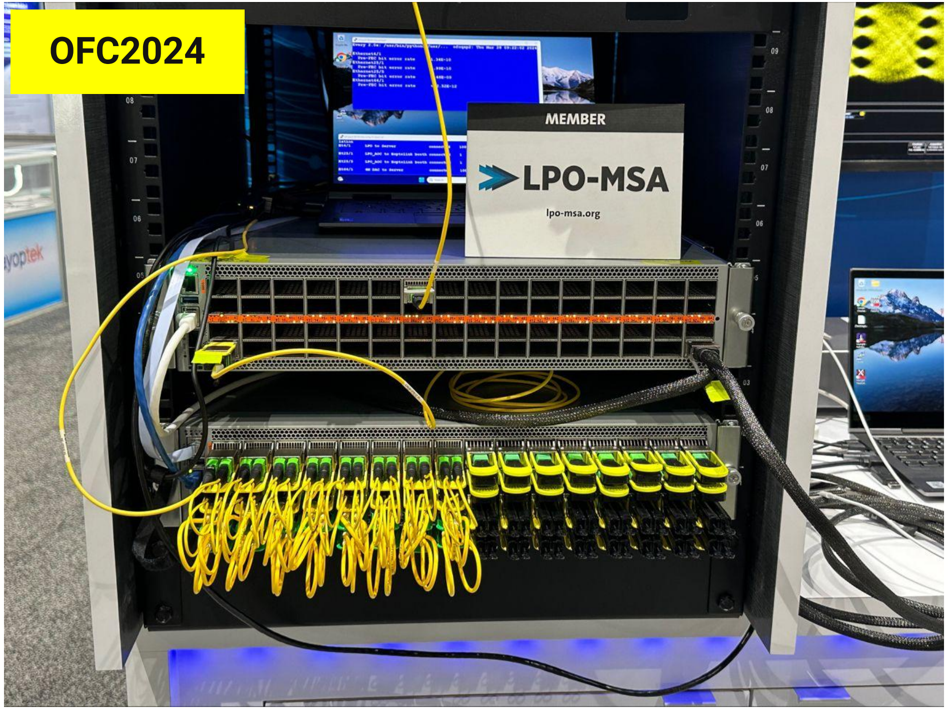
High Density, VOQ,
Multi-chips with optimal cell-based internal fabric

7800R Series 4 to 16 Slots

Up to 460.8Tbps Systems

Next Generation 7060X6-64PE Series

51.2T Throughput: 64 Ports 800G QSFP-DD800 or OSFP800


OFC2024

- Optimized for AI/ML workloads and Hyperscaler
- 51.2 Tbps System with a single chip
- 5nm Process – Lower Power
- 165MB Buffer
- Consistent Tomahawk Architecture
- Comprehensive Instrumentation and Traffic Management

AI Fabric Architectures

AI Fabric

Key requirements

400GbE access ports

No oversubscription

Optimized flow distribution

Lossless

Advanced Telemetry

Key Variables

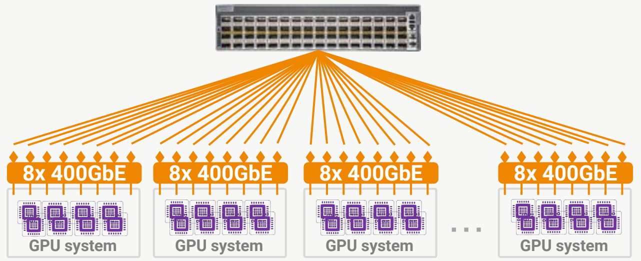
Total # of AI NIC ports

AI NIC SerDes Speed

Rack physical layout and fiber plant

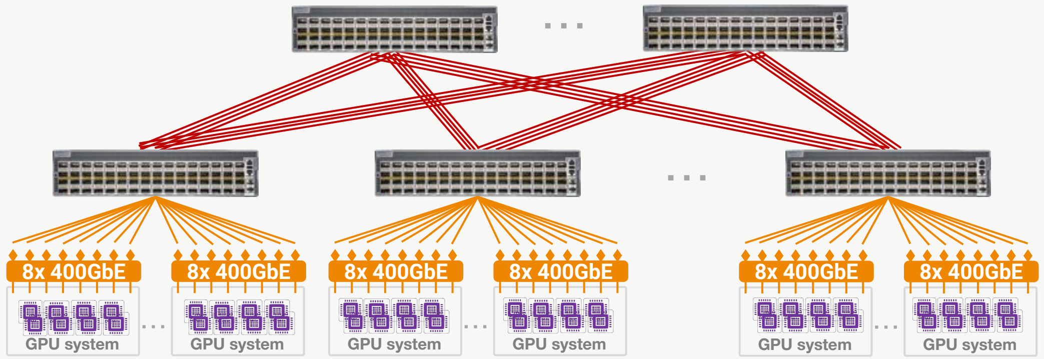
Cost

Single-tiered AI Fabric



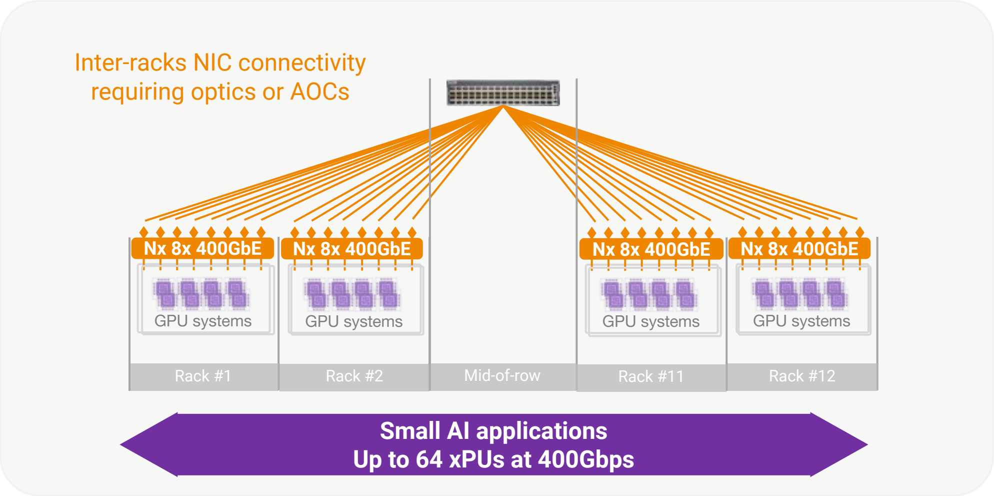
Small and Moderate AI applications (10s and 100s of xPUs)

Multi-tiered AI Fabric



Large AI applications (1000s of xPUs)

Single Tiered AI Fabric - 7060DX5

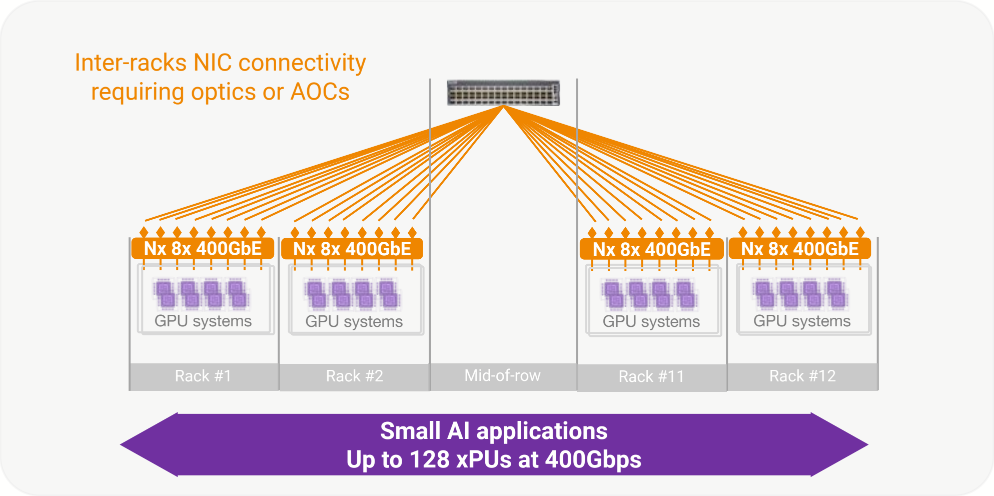


- Fixed Configuration Switch
  - 64x 400G
  - 32x 800G
- No flow collisions
  - Single-asic line-rate forwarding
- ECN and/or PFC to handle incasts
  - Low buffers - Requires tuning
- Single-homed systems
- If a GPU fails, the whole job fails

7060DX5-64S
64 port 400G QSFP-DD

7060DX5-64E & 7060PX5-64E
32 port 800G* (2x400G) OSFP or QSFP-DD

Single Tiered AI Fabric - 7060DX6

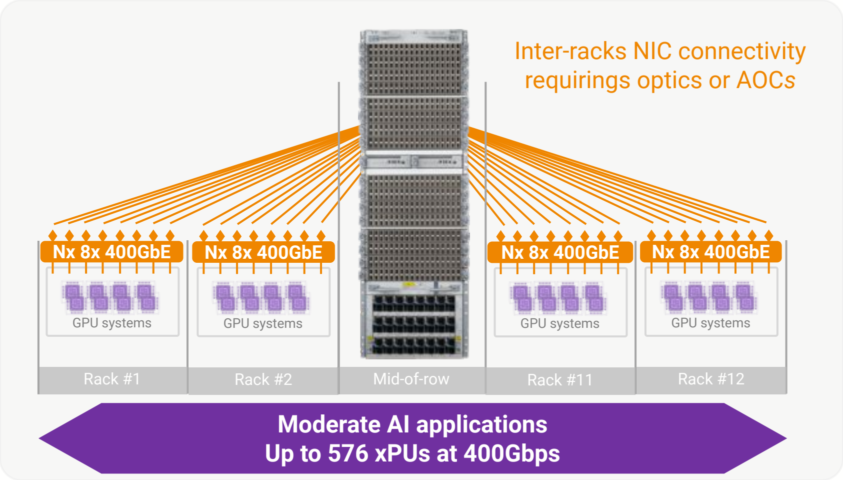


7060DX6-64PE

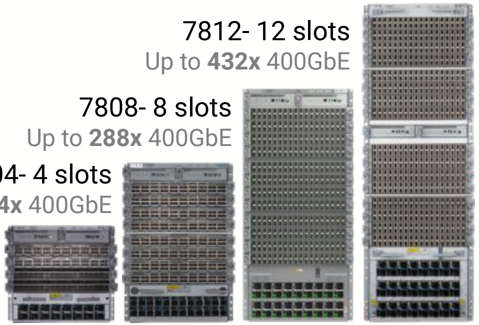
64 port 800G (2x400G) OSFP

- Fixed Configuration Switch
  - 64x 800G
- No flow collisions
  - Single-asic line-rate forwarding
- ECN and/or PFC to handle incasts
  - Low buffers - Requires tuning
- Single-homed systems
- If a GPU fails, the whole job fails

Single Tiered AI Fabric - 7800R3


Inter-racks NIC connectivity requirings optics or AOCs

Moderate AI applications
Up to 576 xPUs at 400Gbps

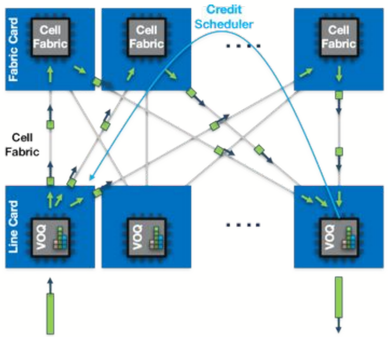

7816 - 16 slots
Up to **576x** 400GbE

7812- 12 slots
Up to **432x** 400GbE

7808- 8 slots
Up to **288x** 400GbE

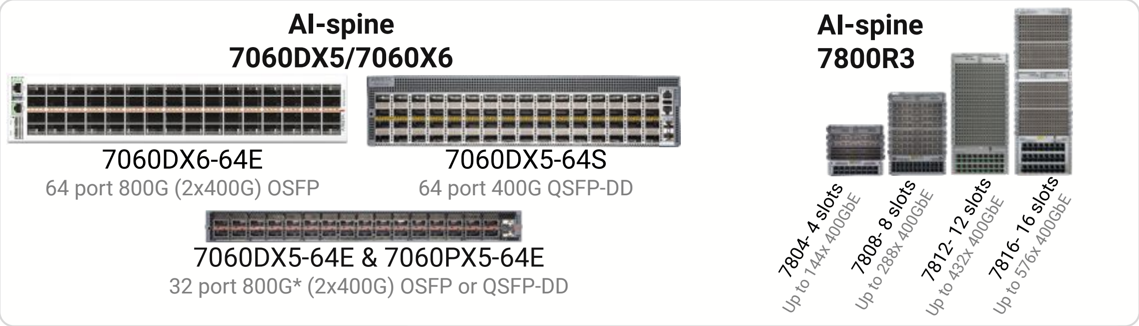
7804- 4 slots
Up to **144x** 400GbE

- **7800R3 Modular chassis offering high port density**
  4, 8, 12 or 16 slots / 36x 400GbE Linecards
- **Non-blocking distributed forwarding**
  Leaf (Linecards) & Spine (Fabrics) in a single chassis
- **No flow collisions between line card and fabric**
  Scheduled Cell-based Fabric
  Built in overprovisioning between line card and fabric
  100% Fair and Efficient Load Balancing within the chassis



- **High Availability**
  Fabric, fan, power supply, sup redundancy
- **ECN and/or PFC to handle incasts**
  Deep buffers - Requires minimal tuning

Multi-Tiered AI Fabric



AI-spine
7060DX5/7060X6

7060DX6-64E
64 port 800G (2x400G) OSFP

7060DX5-64S
64 port 400G QSFP-DD

7060DX5-64E & 7060PX5-64E
32 port 800G* (2x400G) OSFP or QSFP-DD

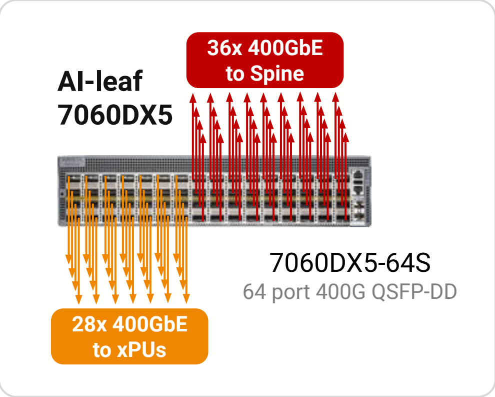
AI-spine
7800R3

7804- 4 slots
Up to 144x 400GbE

7808- 8 slots
Up to 288x 400GbE

7812- 12 slots
Up to 432x 400GbE

7816- 16 slots
Up to 576x 400GbE

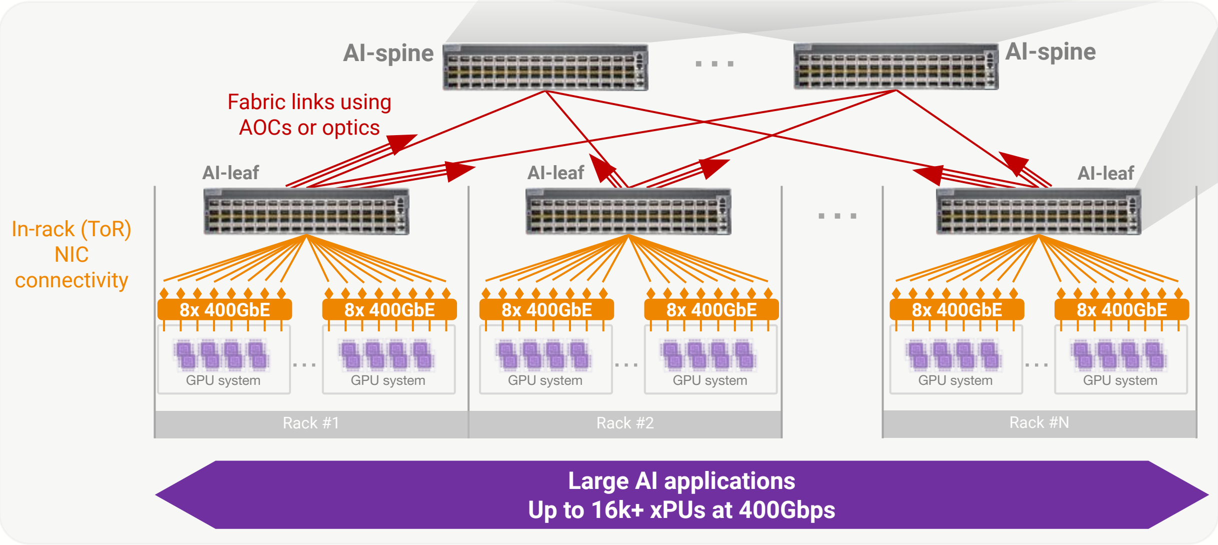


AI-leaf
7060DX5

36x 400GbE to Spine

28x 400GbE to xPUs

7060DX5-64S
64 port 400G QSFP-DD



Large AI applications
Up to 16k+ xPUs at 400Gbps

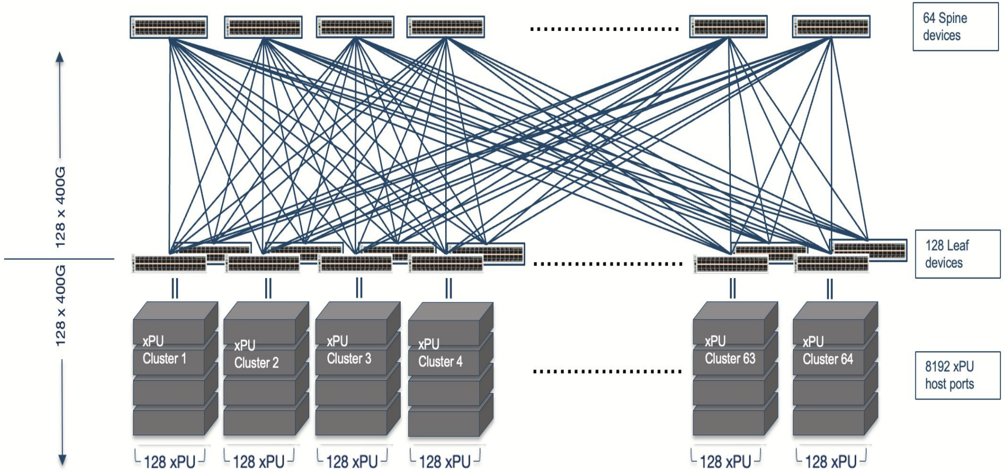
- **High Potential for Flow collision**
  Mitigated with Optimal load-balancing (DLB, Source LB)
- **Uplinks over-provisioning on AI-leaf (1:1.2)**
  No oversubscription in all circumstances
  Address per-flow ECMP imbalance
  Address Link failures
- **ECN and/or PFC to handle incasts**
  Low buffers on AI-leaf - Requires tuning

Example: 8K GPU Cluster w/ 7060X6-64PE

Enterprise AI Focus

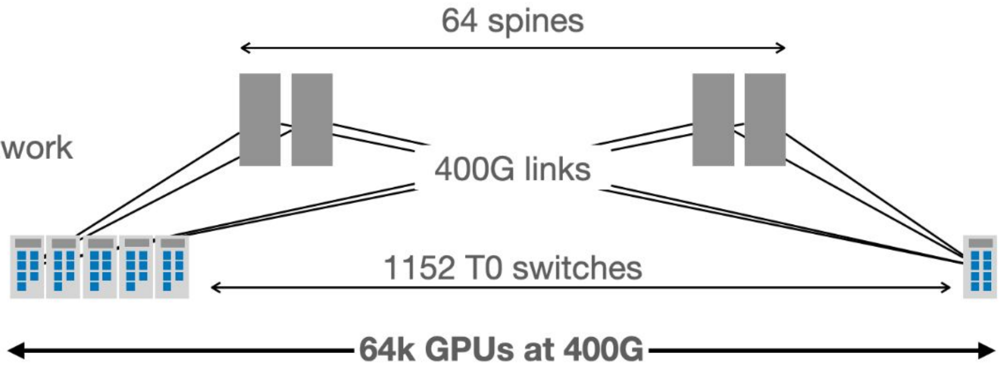
8192 GPUs is ~$200M -> large enterprise cluster

- **Cluster Details**
  - 1024 node GPU cluster
  - 8x 400G NICs per chassis
  - 8192x 400G GPU host ports
- **Design Details**
  - Leaf# 128
  - Spine# 64



Example: 64K GPU Cluster w/ 7800+7060X6

Example 64k GPU network

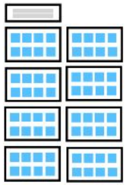


7816 Spine - 460T

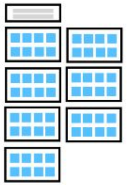
7060 Leaf - 51T
GPUs

- 2 tiers of switches
- 64k 400G GPUs with 400G Leaf-Spine links
- 4032 racks

With 400G GPUs



64 GPUs per Leaf
72k GPUs



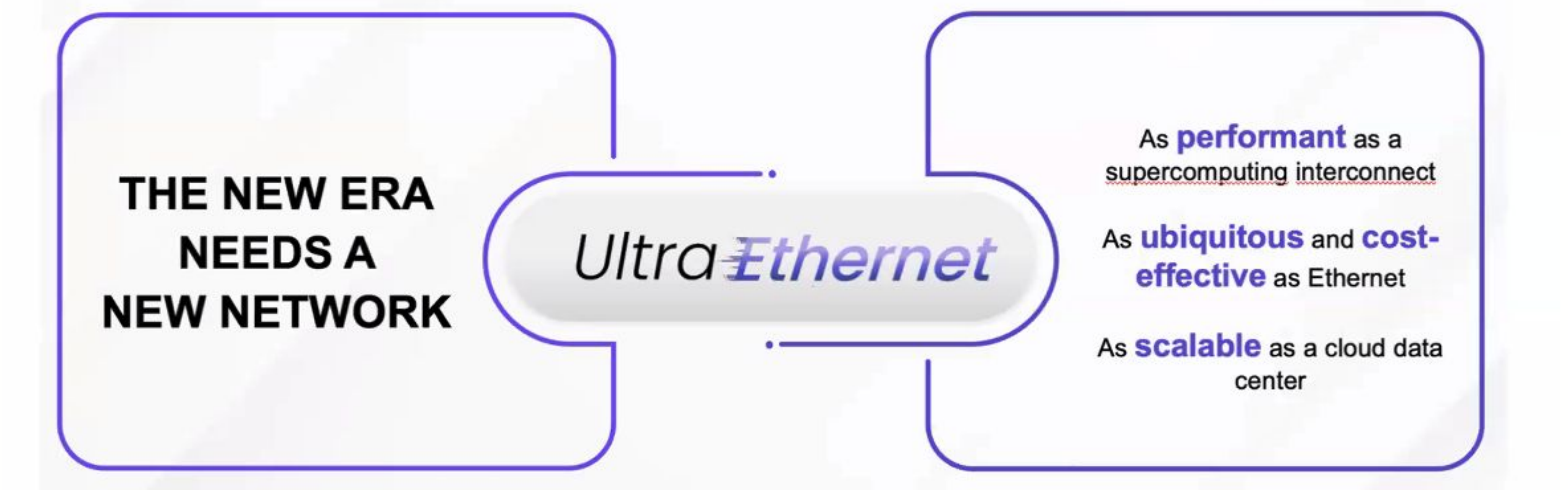
56 GPUs per Leaf
64k GPUs

5 di 6

A QUICK GLANCE TO THE FUTURE OF AI ETHERNET

Ultra Ethernet Vision

Deliver an Ethernet based open, interoperable, high performance, full-communications stack architecture to meet the growing network demands for 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

Source: https://ultraethernet.org/

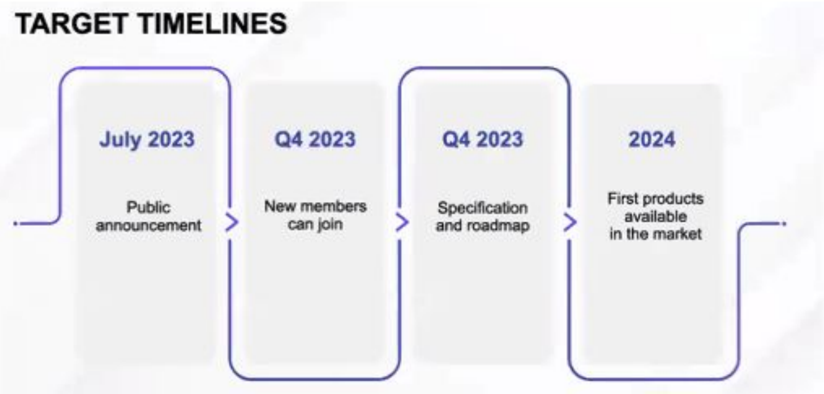
UEC Summary

Mission: IMPROVE Ethernet for AI and HPC

- Started fall 2022, Arista joined Feb 2023
- Steering members: AMD. Arista, Broadcom, Cisco, Eviden, HPE, Intel, Meta, Microsoft, Oracle
- Public launch: July 19th 2023

UEC Deliverables

- July 19th ‘23 - Website and white-paper outlining the problem and the plan
- November ‘23 - Specifications
- Key Specifications is a transport protocol that:
  - enables **packet spraying network** for high utilization
  - supports **out-of-order** packet delivery
  - provides efficient **congestion control**


TARGET TIMELINES

July 2023
Public announcement

Q4 2023
New members can join

Q4 2023
Specification and roadmap

2024
First products available in the market

UEC Deliverables – cont’d

- March ‘24 - Specifications
- UEC progresses towards **v1.0 Set of Specifications**
  - UEC Stack
  - Multi-path packet spraying
  - Flexible ordering
  - “State of the art”, easily configured congestion control mechanisms
  - End-to-end telemetry
  - Multiple transport delivery services
  - Switch offload (i.e., In-Network Collectives)
  - Security as a first-class citizen co-designed with the transport
  - Ethernet Link and Physical layer enhancements (optional)

UEC Transport - Key Properties

- Scales to **1,000,000 Nodes**
- **Packet-Level multi-pathing** for very high network utilization
- **AI-Optimized, configuration-free congestion control**
  - **Incast Management** to address fan-in at the last hop
  - **Rate Control** to ramp quickly to wire rate without impacting existing flows
- Support for **out-of-order packet delivery** with in-order messaging completion
- **Low tail latency**

Highest infrastructure utilization at ultra high scale, without tuning

RDMA Network Technologies Comparison

Feature	InfiniBand	Ethernet/RoCEv2	Ultra Ethernet
Primary RDMA Interface	IB Verbs	IB Verbs	libfabric
Scalable Control Plane	🔴	🟢	🟢
Multi-Path Packet Spraying	🔴	🟡	🟢
Flow Control	Credit-based	PFC/ECN	Dynamic Multi-path
Scheduled Fabric	🔴	🟡	🟢
E2E Drop Recovery	🟡	🟡	🟢
Transport Encryption	🔴	🔴	🟢
Multi-Vendor Ecosystem	🔴	🟢	🟢

ARISTA

6 di 6

CONCLUSIONS

Takeaways

- No Multihoming - if one GPU job fails they all fail
- GPUs are explotentailly more expensive than the switch
  - Do not make savings on the network
- Traffic is very spiky, 0 to 400G in milliseconds, then back again
  - Real time observability
- Packet delivery and load balancing (elephant flows) are critical:
  - Any data loss, job may have to start again - expensive
  - Any slow down, the entire job slows down - expensive
  - Dynamic Load Balancing or cell forwarding
  - DCB/RoCEv2 for packet delivery

Pillars to run a HPC/AI netowrk

Simple design

IP Standards

Operations

https://www.arista.com/en/solutions/ai-networking

ARISTA

AI ready networks

Thank You

www.arista.com