The bridge to possible

Building Ethernet AI Fabrics with Silicon One

Designs, Vision, and Challenges

Ramesh Sivakolundu
Scott Carter

BRKNWT-2407

cisco Live!

#CiscoLive

Cisco Webex App

https://ciscolive.ciscoevents.com/ciscolivebot/#BRKNWT-2407

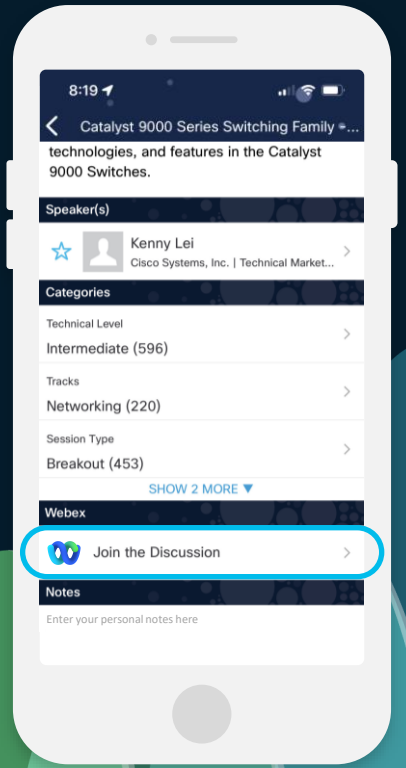
Questions?

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

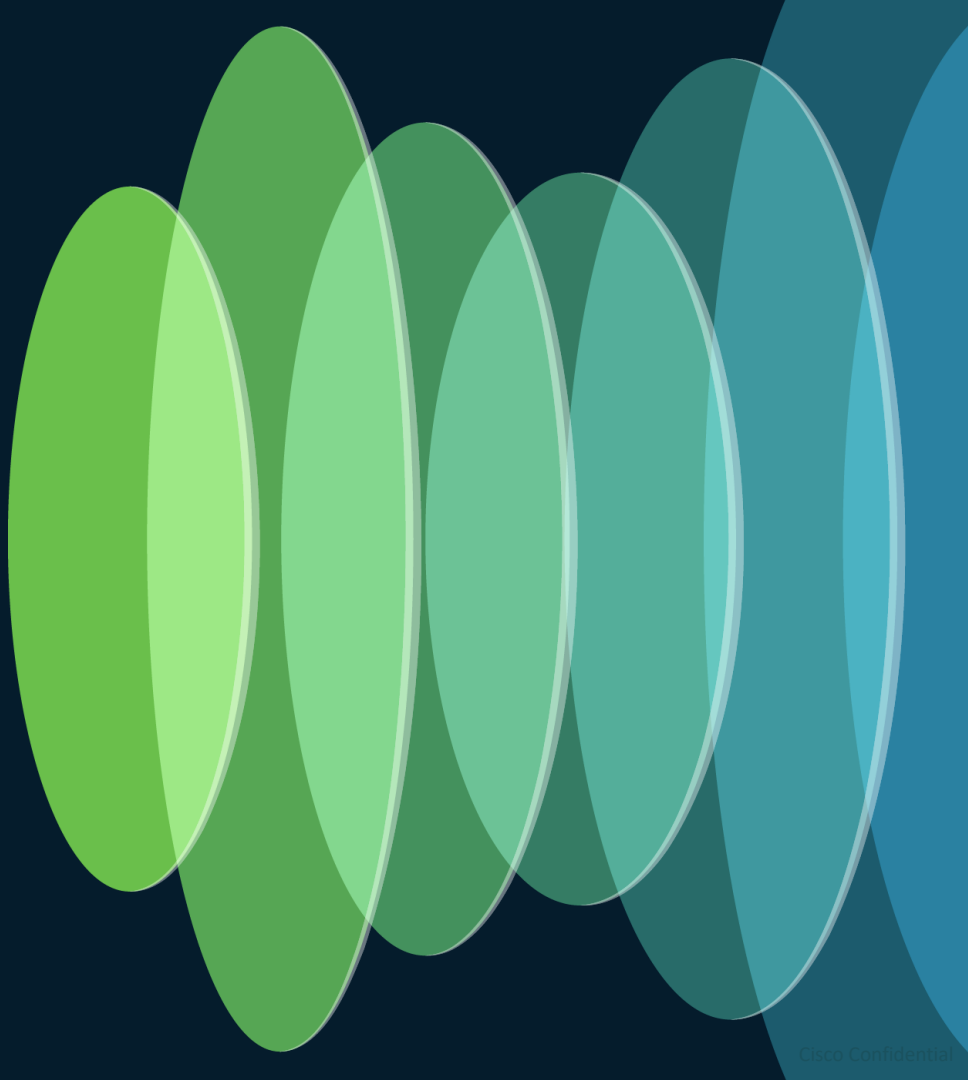
Webex spaces will be moderated by the speaker until June 7, 2024.



Agenda

- Brief State of AI Networking
- AI Workflows and the Training Network Bottleneck
- Parallelism and its Impact on Cluster Design
- Today’s AI infrastructure Options
- Building AI Infrastructure with Silicon One and Cisco 8000
- Addressing Trends
- Wrap-up and Questions

Brief State of AI Networking



“Every Hyperscaler has active plans to implement Ethernet at scale in their AI networks.”

Kevin Wollenweber

SVP/GM, Cisco Data Center and Provider Connectivity

The Elephant in the Room

Let's ask ourselves some questions about InfiniBand

Performance Considerations

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

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

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

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

Operational Considerations

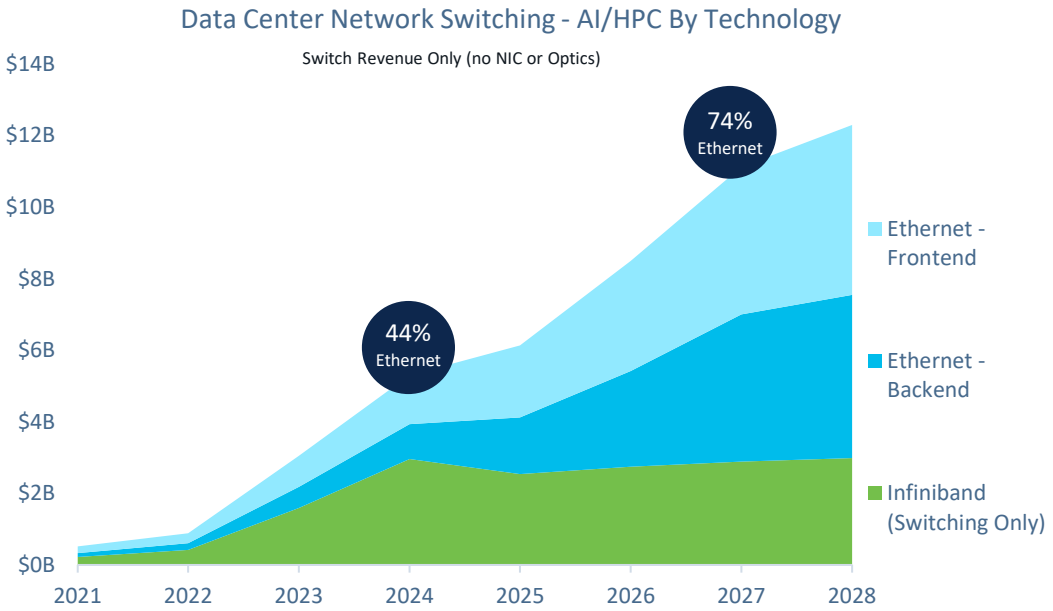
Multi-Vendor Support
How many vendors support the technology?

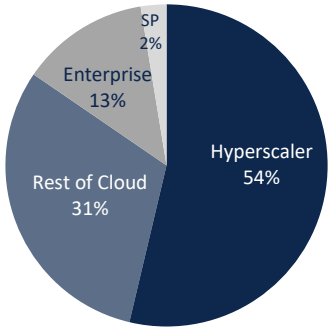
Support for customer-built AI Machines
Is the network flexible to support multiple GPU types?

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

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

Don’t just take our word for it…



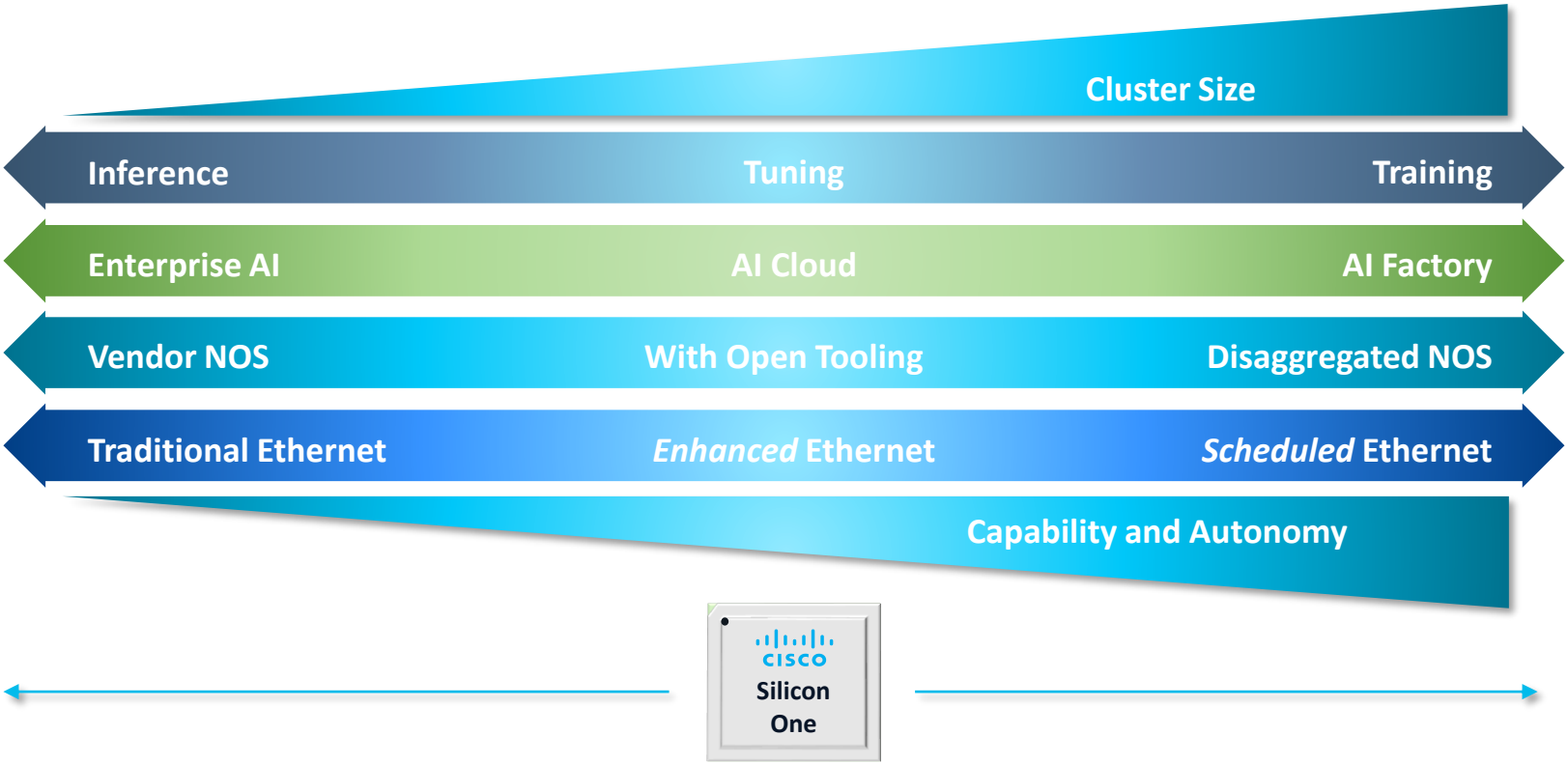


Source: 650 GROUP Market Intelligence Research

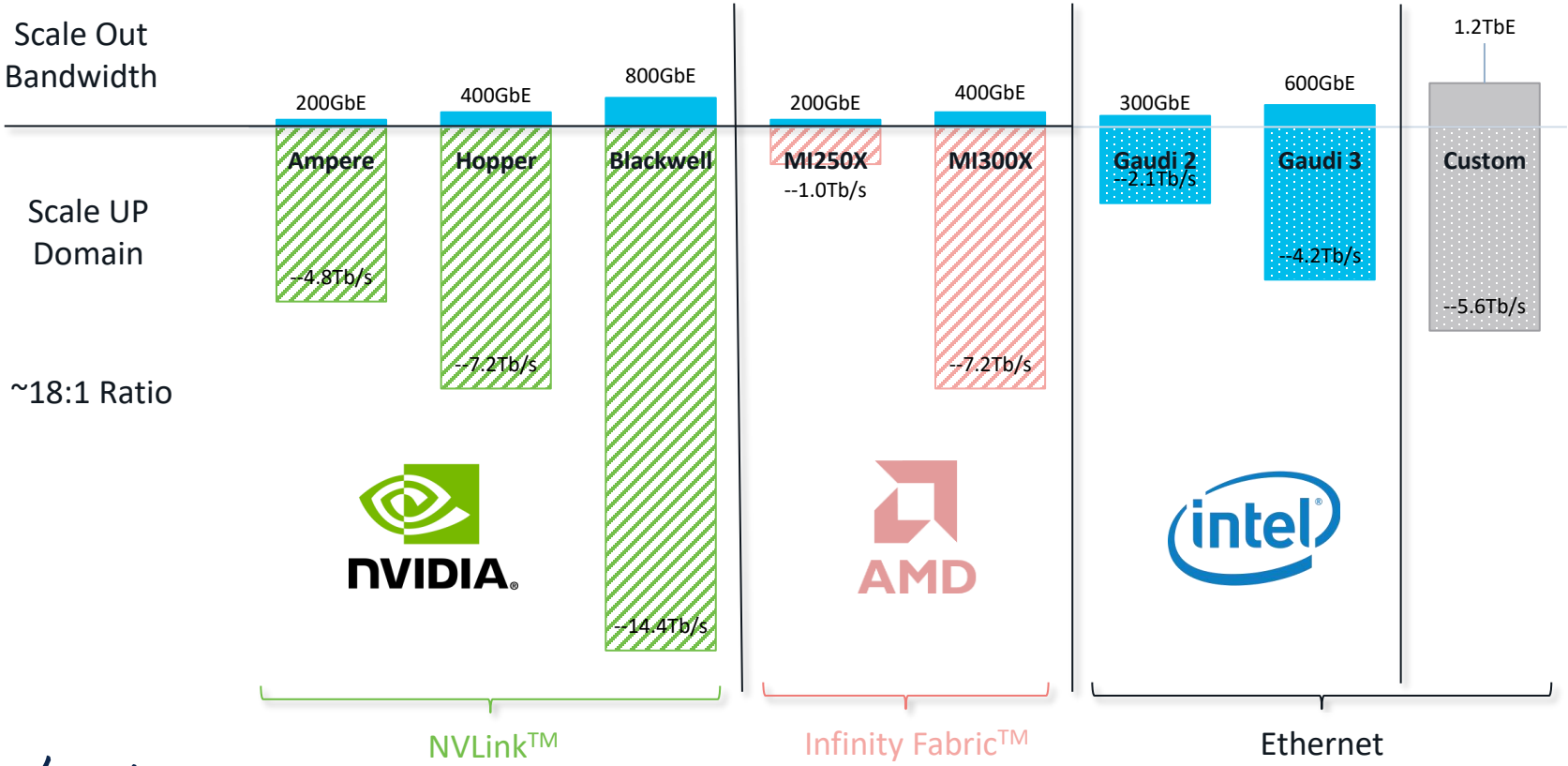
Data Center AI Networking Quarterly Market and Long-Term Forecast Report 4Q23 (March 6th, 2024)

Where does Ethernet fit across AI Landscape?

An Oversimplified Clustering of AI Technology Requirements

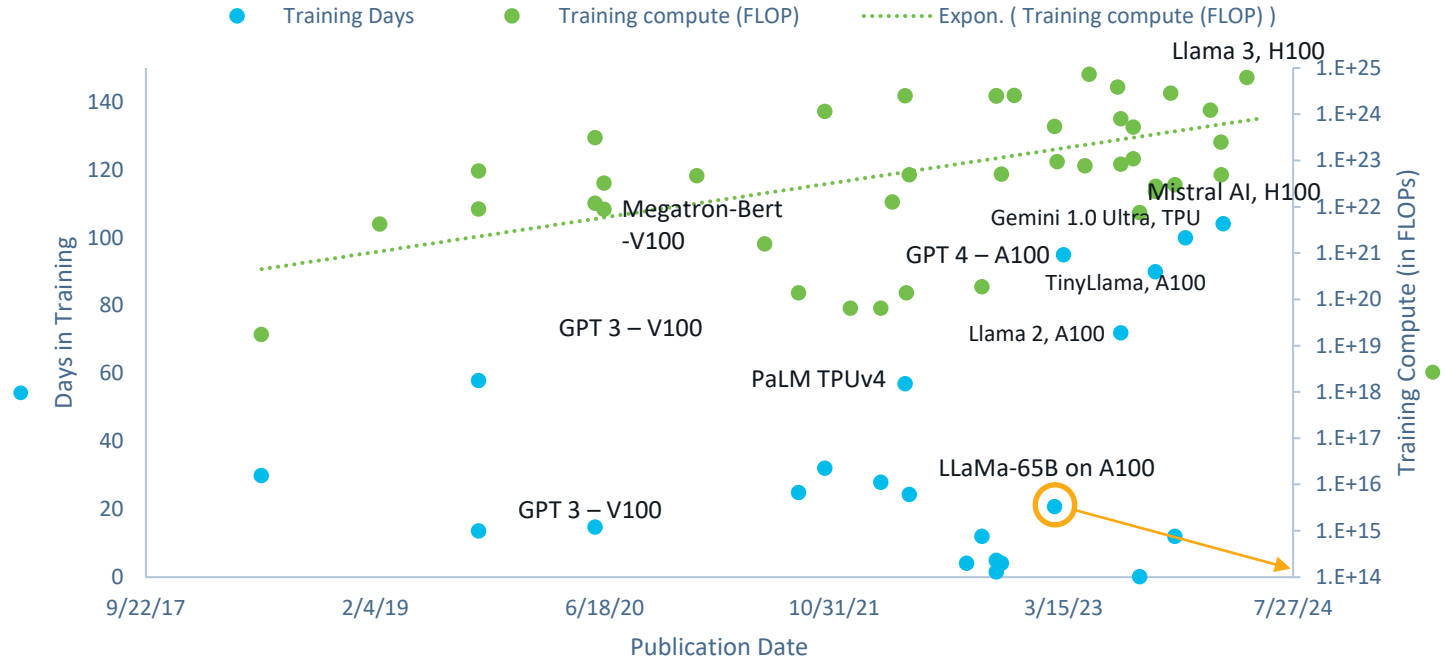


Turns out *Attention isn't* all you need…



Large Model Compute Need is Doubling Every 6 months

Accelerator performance and scale of deployment determines training speed



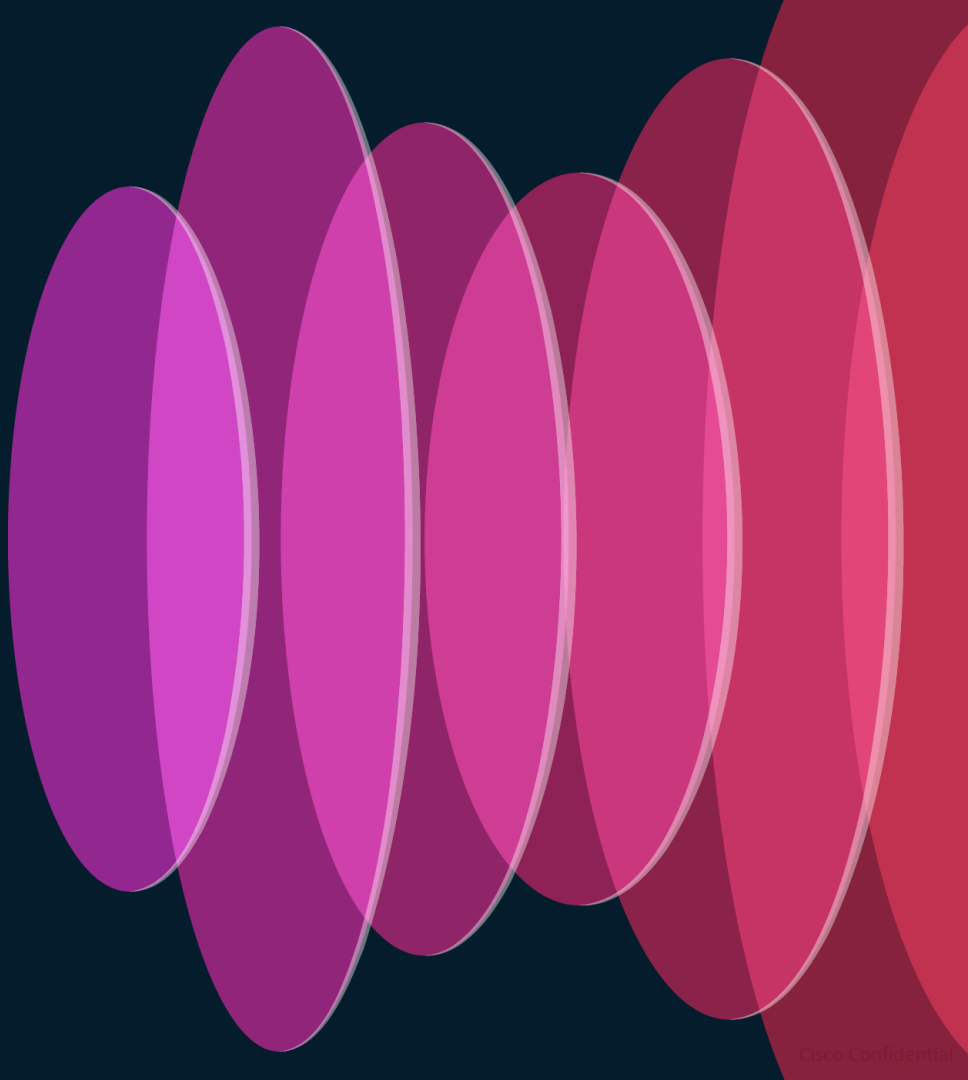
With 32k GPUs, LLaMa-65B can be trained in 1 day!

Source: EpochAI and Networking @Scale 2023 (Meta)

How do we ***optimize*** our ***accelerator investment?***

How do we ***get the network out of the way?***

AI Workflows and the Training Network Bottleneck



LLMs are orders of magnitude more intensive than DLRM

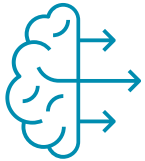
Deep Learning Recommendation Models

Search, Feed ranking. Ads & content recommendation

Inference needs a few Gigaflops for 100ms TTFT

Narrower scope, domain specific

Training: ~100 Gigaflop/ sentence



Large Language Models

Intricacies of human language

Inference needs 10s of Petaflops for 1 sec TTFT

Generate intelligent, creative responses

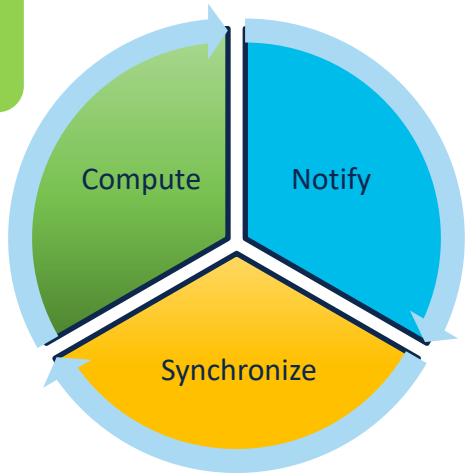
Training : ~1 Petaflop/ sentence

An Improved user experience means *a faster time to first token*, making *distributed inference an imperative*

The AI/ML Workload Cycle

GPU Execute Instructions

High Bandwidth capable GPUs can saturate network links



Send results of computation

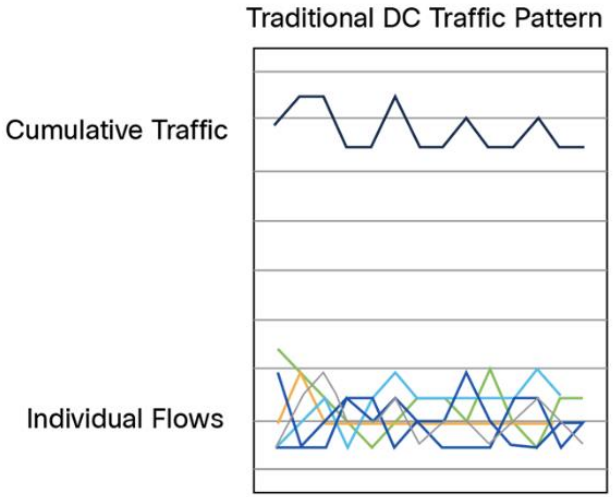
Different collective communication patterns
All Reduce (Aggregate/reduce everyone's data and send to everyone

Wait for all GPUs to complete

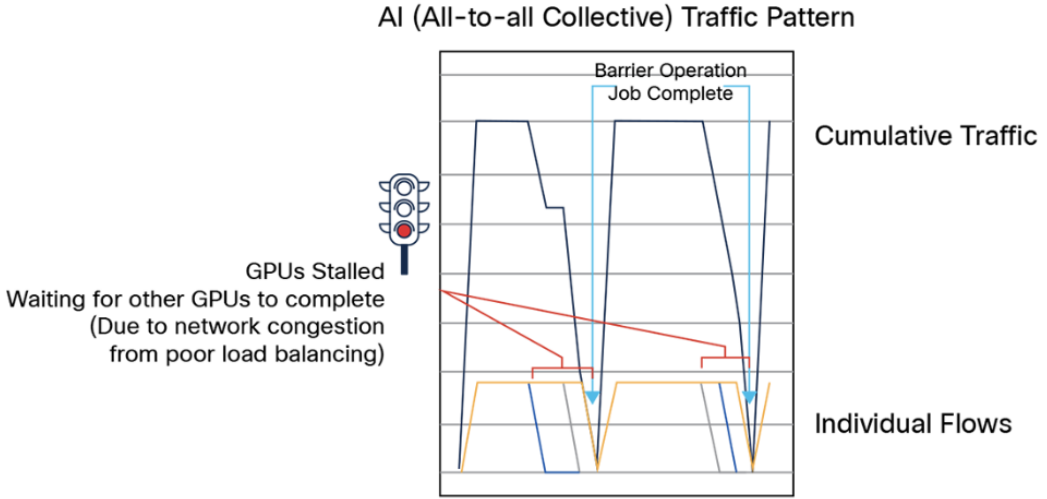
Synchronizes all GPUs
Compute stalls, waiting for the slowest path
Job Completion Time (JCT) influenced by the worst-case tail latency

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

The network can become the bottleneck

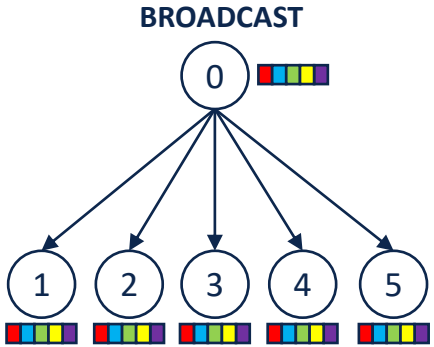


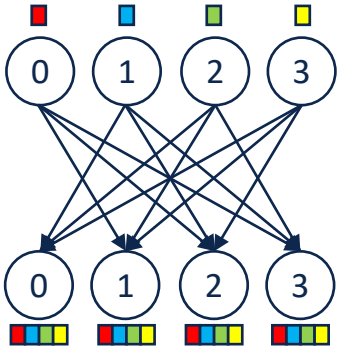
Many asynchronous small BW flows
Chaotic pattern averages out
to consistent load

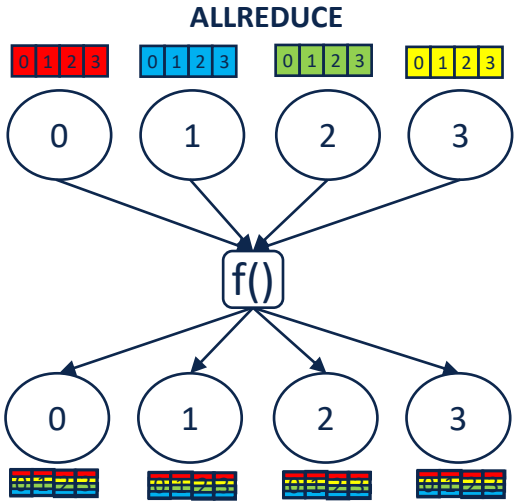


Few synchronous high BW flows
Synchronization magnifies long tail
latency and bad load balancing decisions

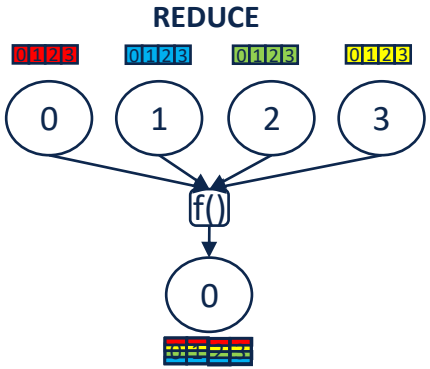
Typical Collective Operations

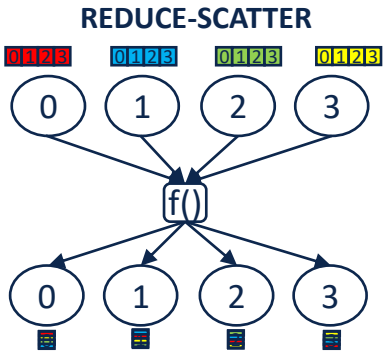




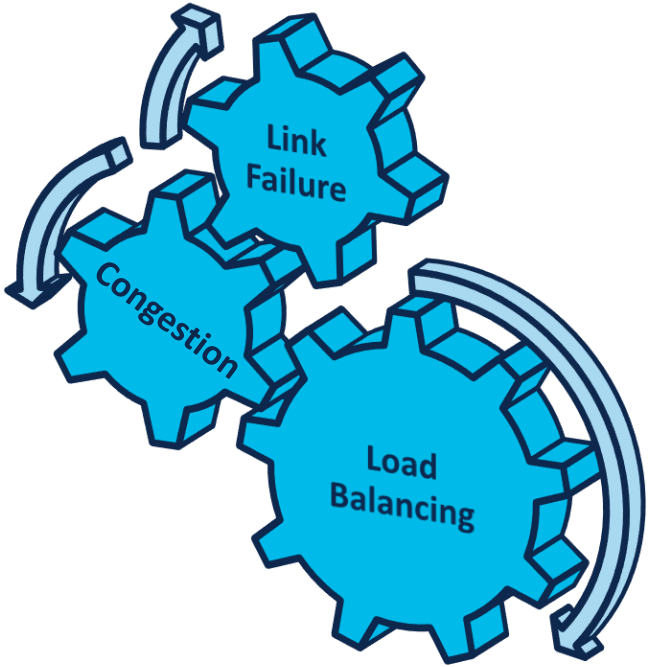


BARRIER is like ALLREDUCE without data, for synchronization only.





Minimizing Job Completion Time is *the AI Challenge*



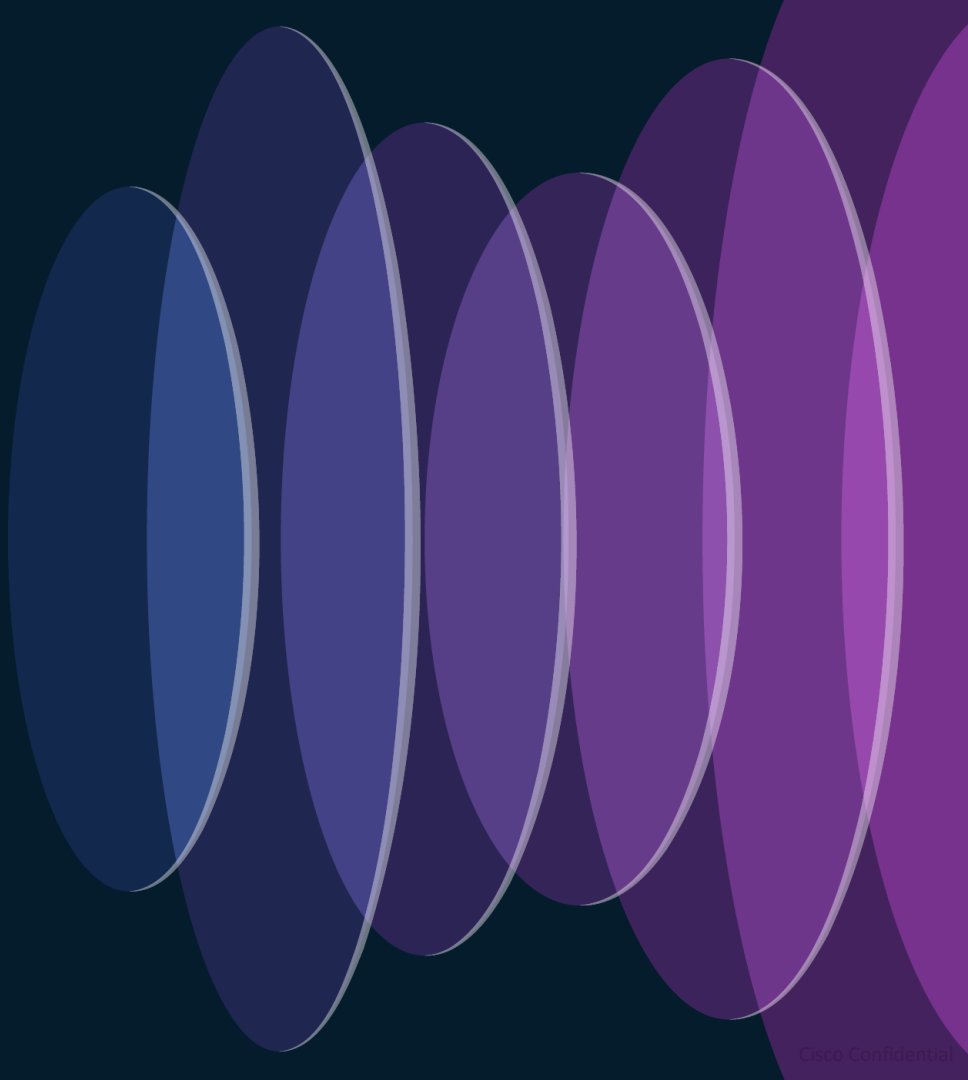
Wrenches in the works

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

AI Ethernet Fabric Options

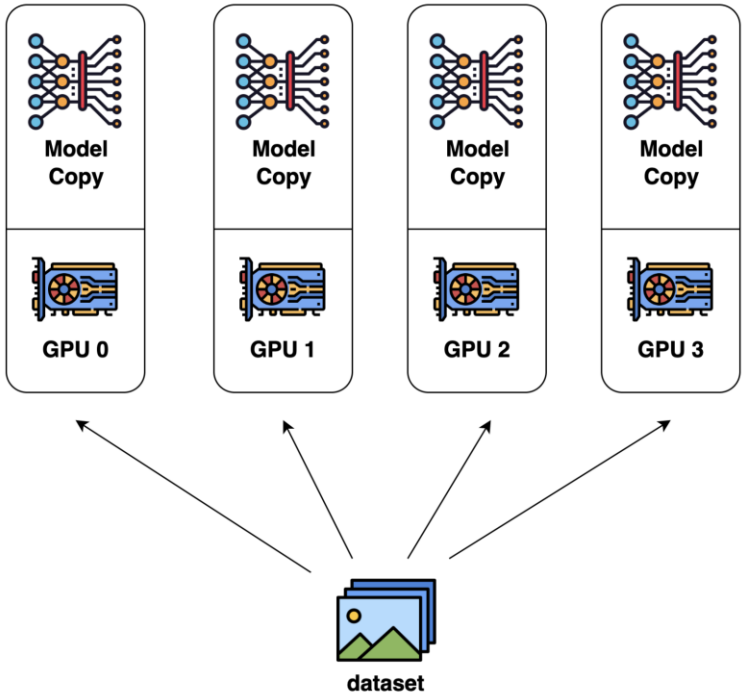
	1 Ethernet	2 *Enhanced* Ethernet		3 *Scheduled* Ethernet
Load Balance	Stateless ECMP	Stateful Flow/Flowlet	Spray & Re-order in SmartNIC	Spray & Re-order in leaf
Fabric Congestion Management	Congestion Reaction with ECN/PFC	Congestion Reaction with congestion score to adjust distribution		Congestion Avoidance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

Parallelism and its impact on Cluster Design



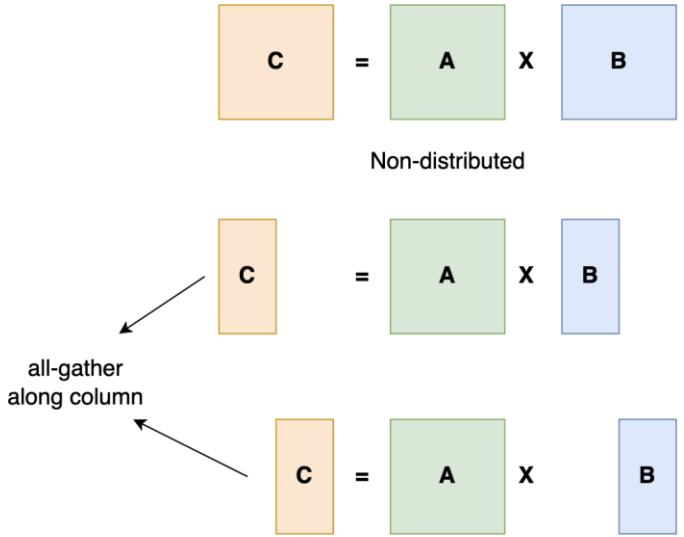
Data Parallelism

- Dataset is divided
- Subset is processed simultaneously by different processors.
- Effective when dealing with vast datasets.



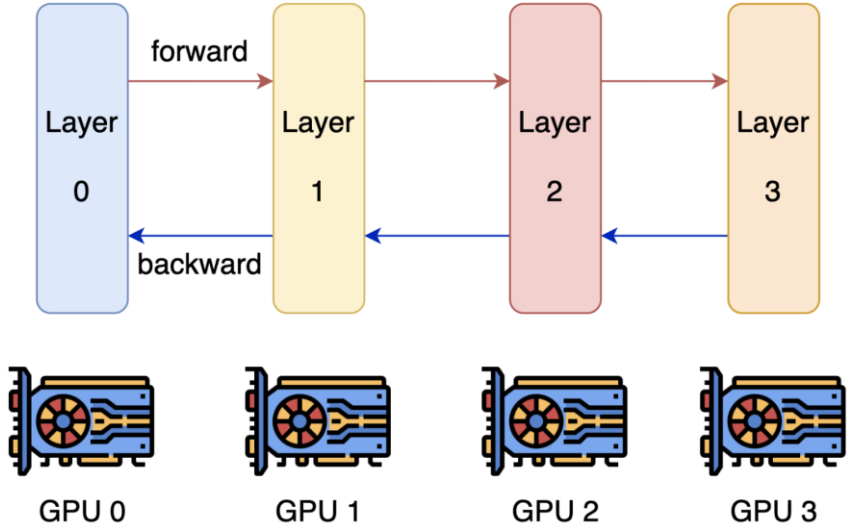
Model (Tensor) Parallelism

- Dividing the neural network into segments
- Each segment is processed by a separate processor.
- Beneficial when dealing with extremely large models

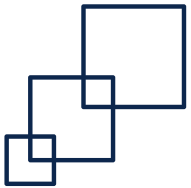


Pipeline Parallelism

- Organizes the workflow into stages
- Each stage is executed by a separate processor.
- Useful in scenarios where there are distinct sequential steps in the model training



2D or 3D Parallelism



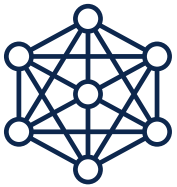
Inference

2D Parallelism - Memory efficiency

- Pipeline parallelism first
- Further each pipelined stage is split using tensor parallelism.
- Extreme partitioning can lead to communication overhead
- Impact compute efficiency.

3D Parallelism – Compute Efficiency

- Leverage data parallelism beyond tensor and pipeline approaches
- Topology aware mapping
- Optimize for intra-node and inter-node communication overheads



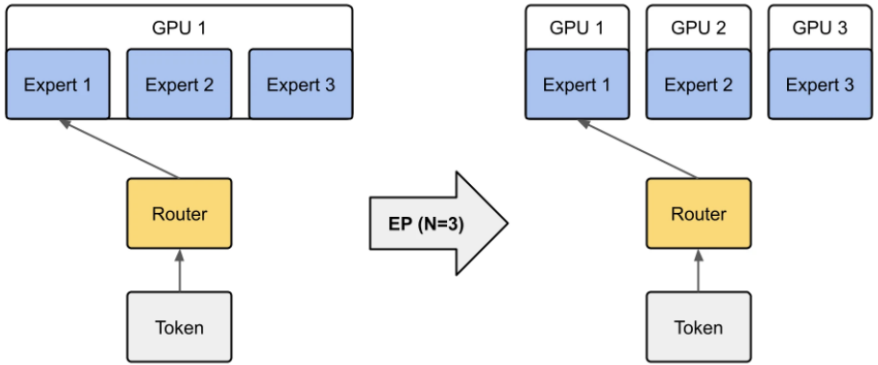
Training

Expert Parallelism

- Model of Experts (MoE)
  - Combines predictions from multiple models
  - Improves overall accuracy

- Experts – Smaller models
  - Trained to perform well in a certa domain.
  - Complex Neural network
  - Simple Decision Tree

- Distributes experts across GPUs



Why Do We Need Parallelism?

Parallelism allows for:

- Efficient Data Transfer
- Cluster Computation Collaboration
- Scalability, Flexibility and Manageability

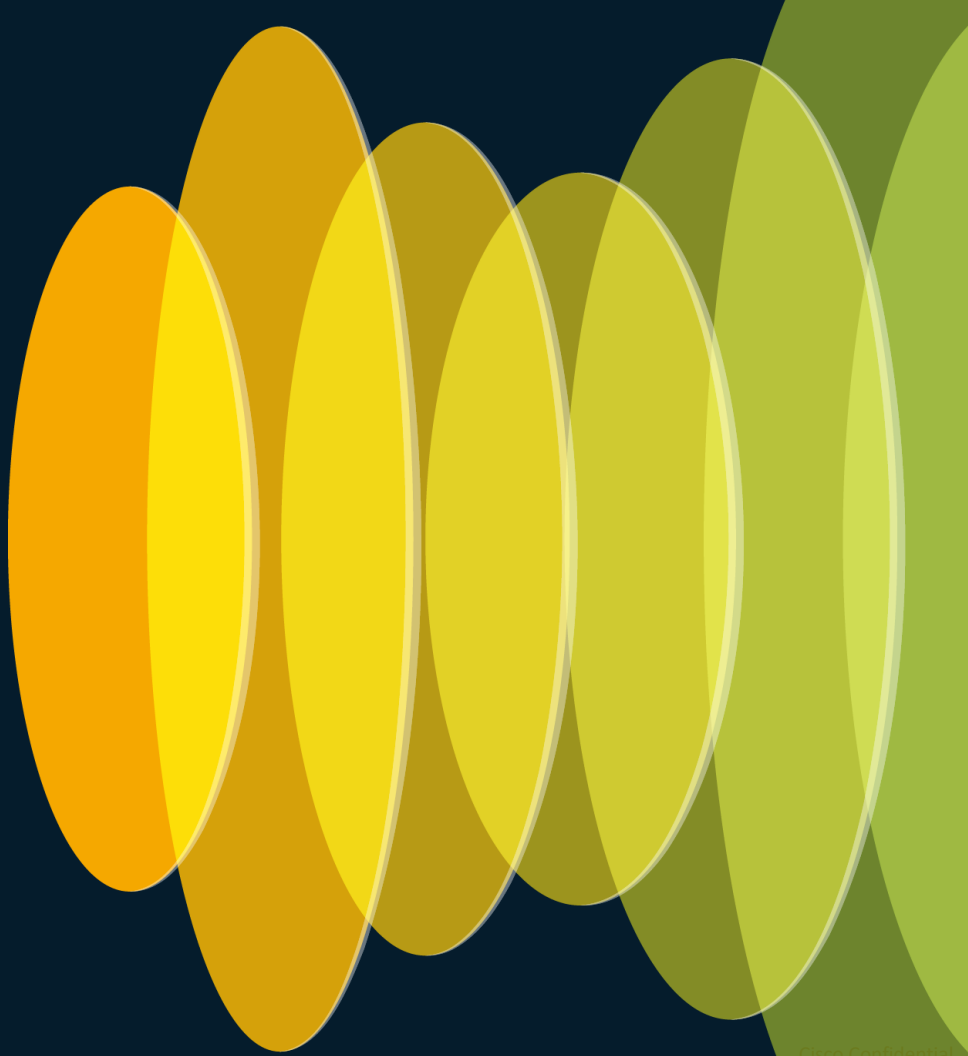
What we can do:

Get the Network Out of the Way

- Efficient utilization of cross-sectional bandwidth
- Enhanced & Scheduled Ethernet offerings
- Build this into class-leading Silicon, Hardware, Software, and offerings

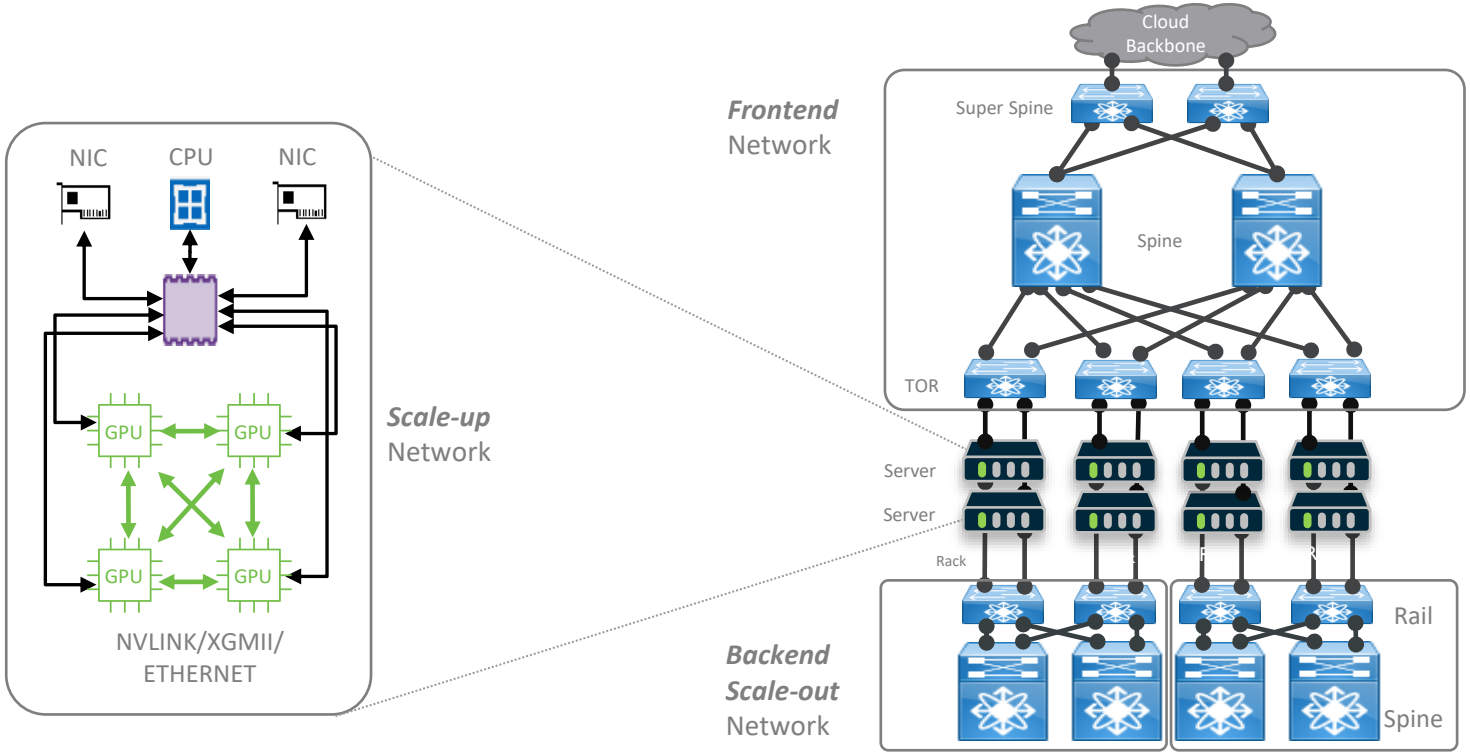
Be Open, Be Extensible, Be Adaptable

Todays AI Infrastructure Options



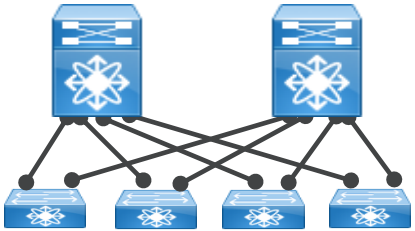
cisco Live!

AI Network Type Fundamentals



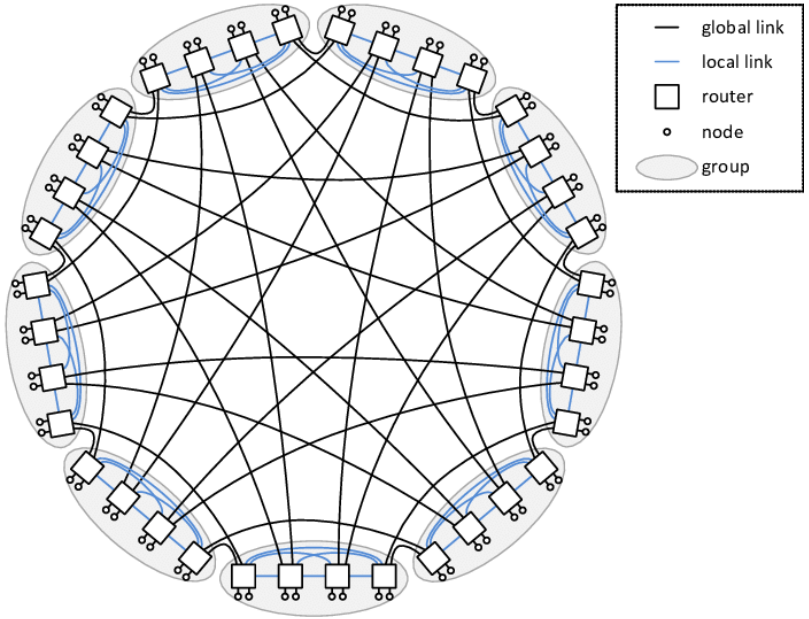
Scale-Out Options: Trusted and True Clos

- Clos Network / Fat Tree:
  - Aka, “Leaf and Spine”
  - Standard Ethernet DC design
  - ”Radix” scales number of switches in a stage (tier) of the network
  - Plenty of redundant paths, can produce consistent number of hops by forcing all traffic to traverse spines
  - Optionally, can locally switch to exploit locality
  - More links required as you scale the network
  - “Fat Tree” is a term used commonly in HPC and can be thought of as a more hierarchal version of a Clos Network



Scale Out Options: The Dragonfly

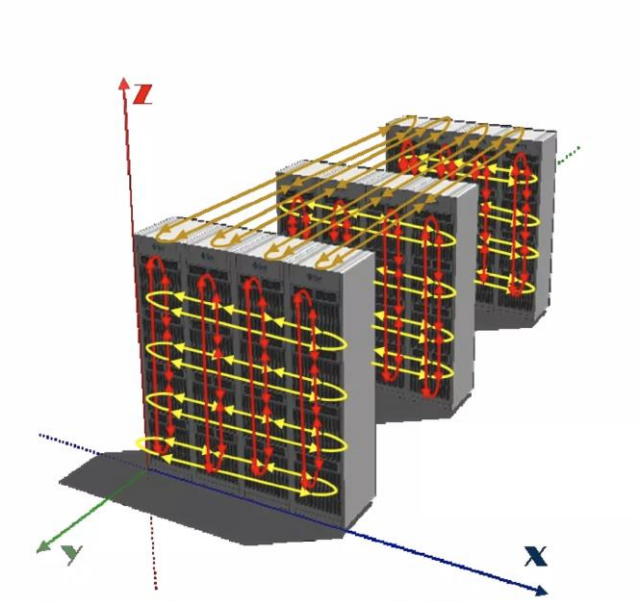
- Dragonfly:
  - Exploits locality within groups
  - End-points are grouped, with each group interconnected to other groups
  - Adaptive routing should utilize non-shortest paths to efficiently use inter-group links
  - Elements of a ring topology (useful for ML applications where workload passes between neighbors)
  - Common InfiniBand topology for HPC
  - Dragonfly+ scales to higher number of hosts (36 port radix allows 26,000 end-points in Dragonfly vs 105,000 in Dragonfly+)1
  -



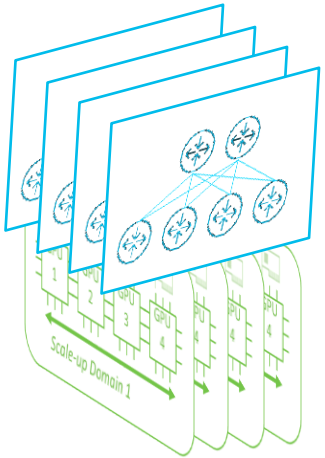
1 – “Dragonfly+: Low Cost Topology for Scaling Datacenters” https://ieeexplore.ieee.org/document/7885210

Scale Out Options: The Torus Mesh

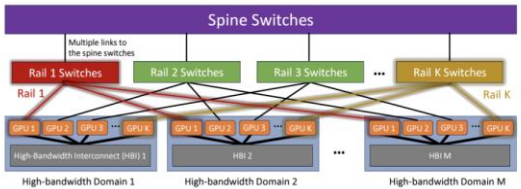
- 3D Mesh/Torus
  - Doesn’t require high radix switching and can use short distance cables
  - End-points are grouped, but topology can be blocking
  - Application/network needs to be topology aware



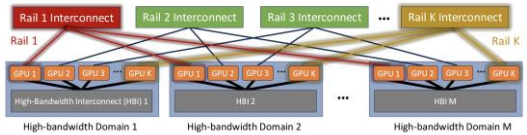
The ”Rails” Approach



- A **rail** is the set of all GPUs with the same index (or rank) between multiple scale-up domains (servers), interconnected with a rail switch.
- Each rail is connected by a **dedicated but separate Clos network**.
- The GPU is interconnected between the Scale-up and Scale-out networks:
  - NVLink/XGMII/GbE high-bandwidth but short-range interconnection.
  - A conventional RDMA-enabled NIC.
- Rails allows for greater scaling, but is dependent on the communication library to coordinate between domains.



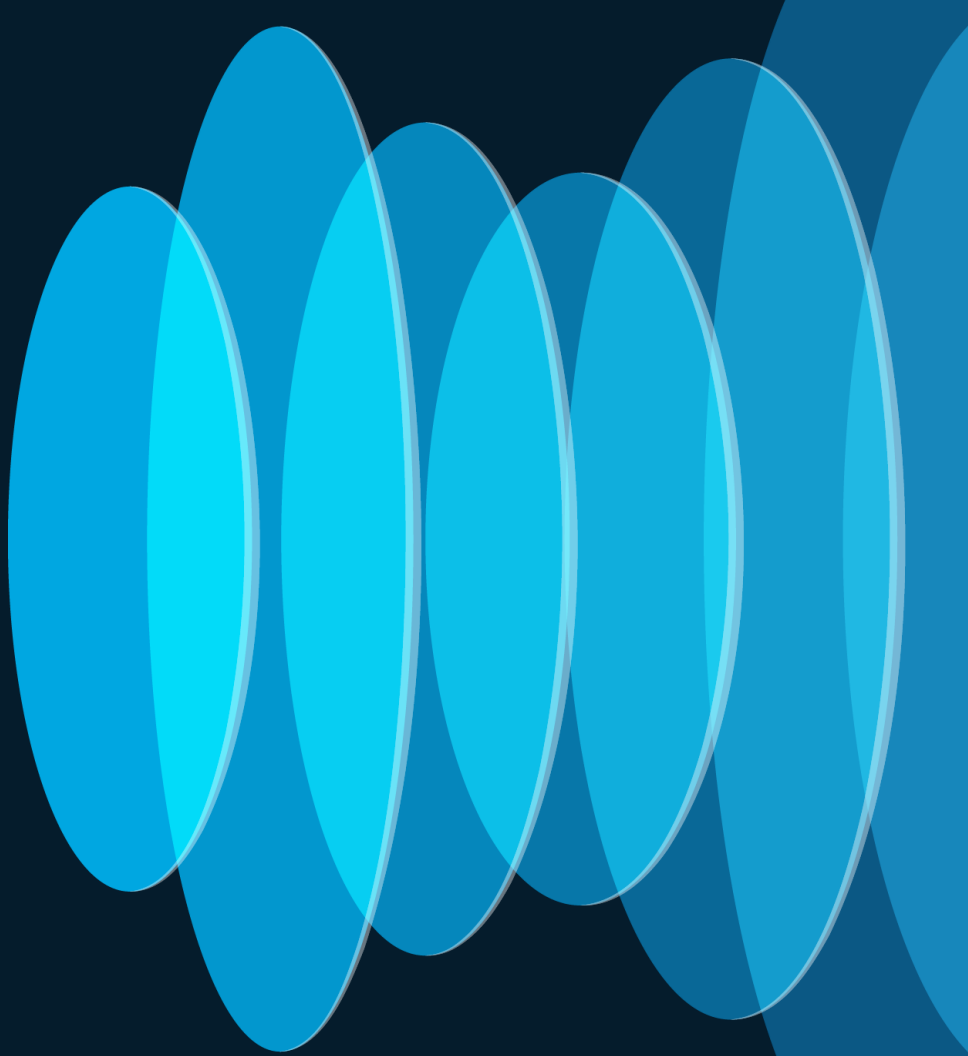
RAIL Optimized



RAIL Only

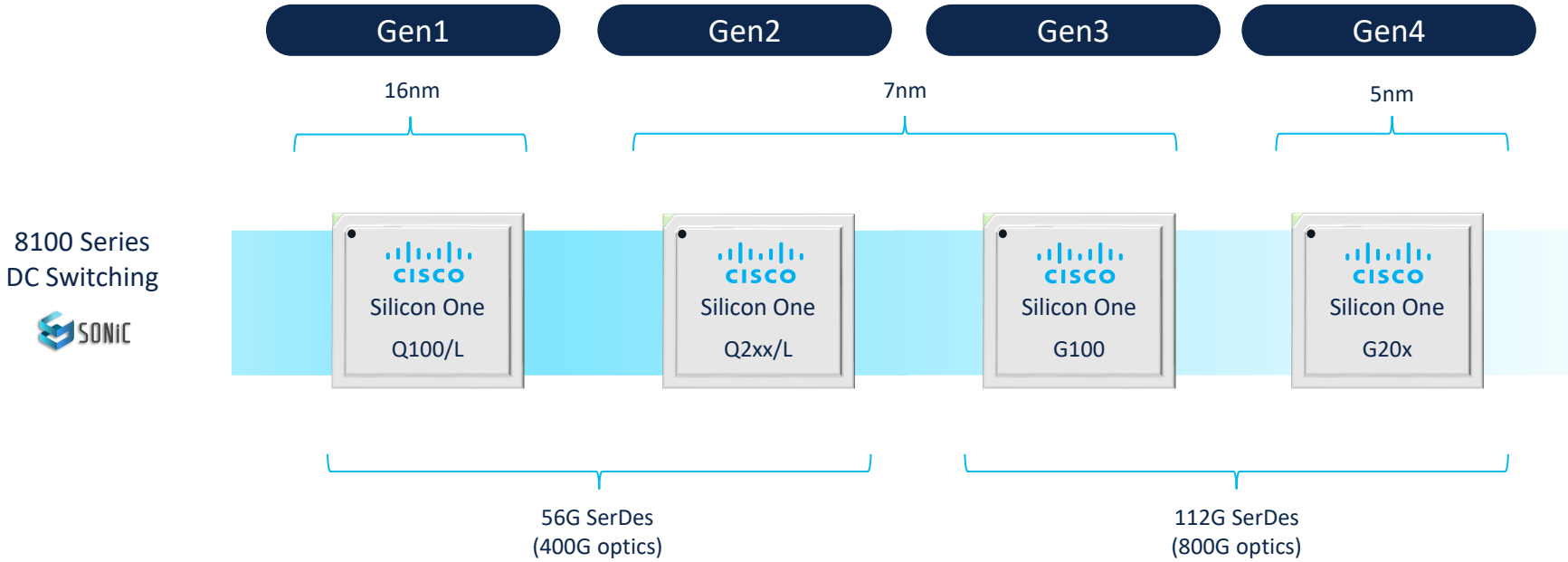
Source: [How to Build Low-cost Networks for Large Language Models]

Building AI Infrastructure with Silicon One and Cisco 8000



cisco Live!

Cisco Silicon One



Cisco Silicon One G200

Uniquely efficient and Optimized for AI/ML

One architecture
A simpler and easier network to maintain

High Performance
2x higher performance than G100

Sustainability via technology
2x more power efficient than G100

Ultra-low latency
2x Lower Latency than G100

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

Fully shared packet buffer
Optimal fairness, burst performance, JCT

51.2 Tbps

AI Primed

Advanced 112 Gbps SerDes
Cisco designed next-generation ADC SerDes
Support for Optics, 4-meter DAC, LDO and CPO

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

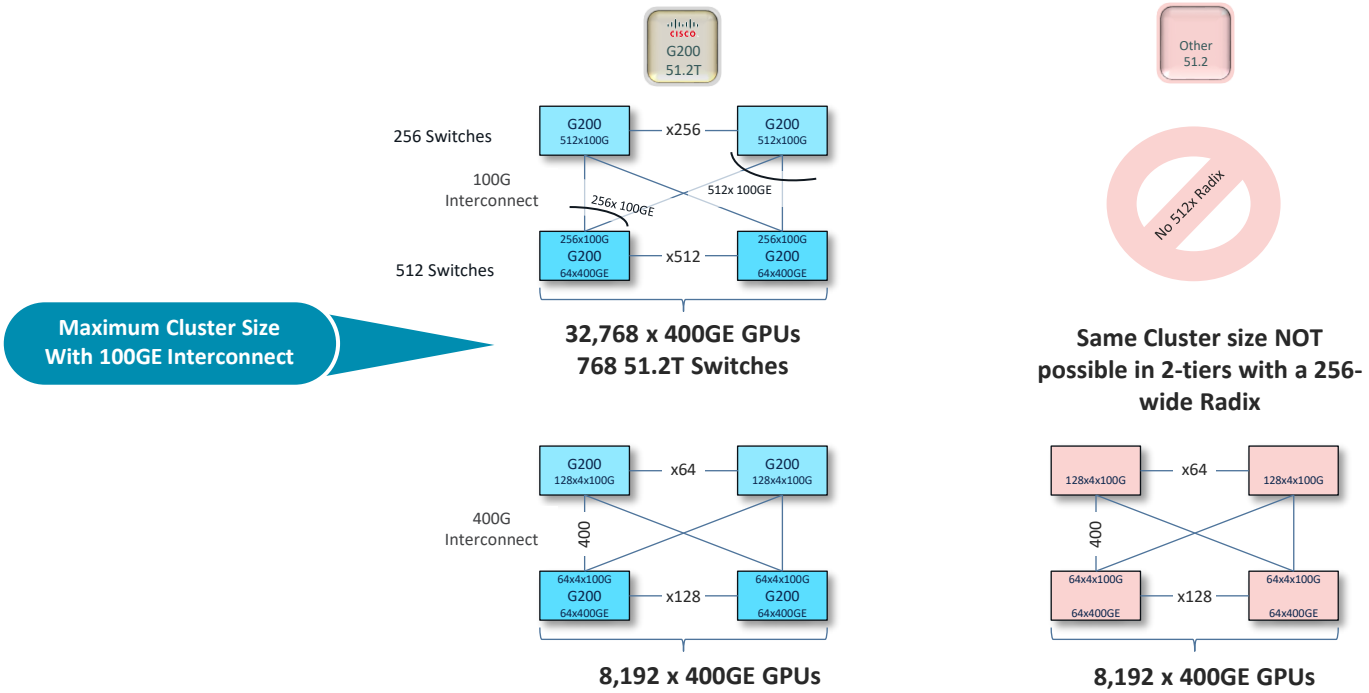
Link failure avoidance
HW based traffic link failure redistribution optimizes real-world large-scale deployments

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

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

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

Why a full 512-wide Radix Matters for 2-Tier Clusters

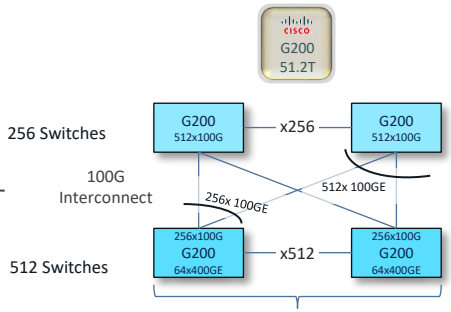


Why a full 512-wide Radix Matters for 2-Tier Clusters

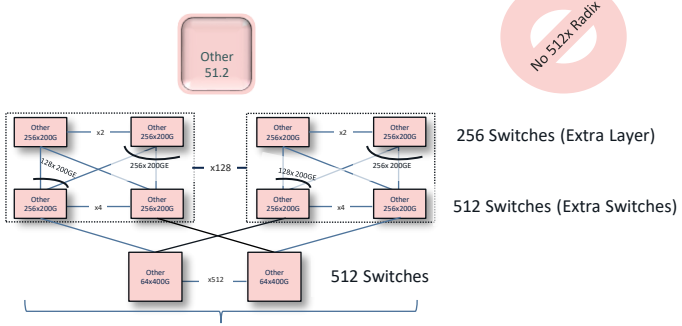
- 50% Less Optics
- 40% Less Switches
- 33% Less Networking Layers

- Up to 1 Mega Watt (MW) Less Power

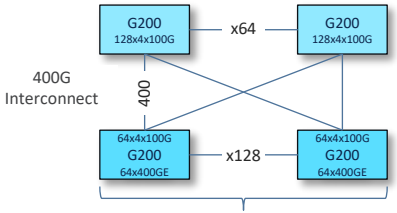
Maximum Cluster Size With 100GE Interconnect



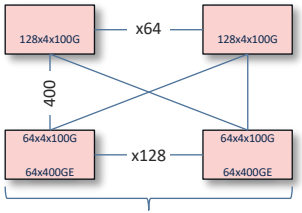
32,768 x 400GE GPUs
768 51.2T Switches



32,768 x 400GE GPUs
1,280 51.2T Switches



8,192 x 400GE GPUs



8,192 x 400GE GPUs

8122-64EH-O

AI Primed

SONiC

64 x 800G

Shipping Q3 2024

cisco Silicon One G200

Hardware Summary
Single 51.2T G200 ASIC (5nm) 256 MB SRAM packet buffer
Quad Core x86 CPU 32 GB DRAM
RS-232 Console, 1GbE Management, 1XUSB2.0, 2XQSFP28 Telemetry PIE (Punt Inject Engine) ports
4 Fans, 1+1 PSU Redundancy Port side intake airflow
2kW AC & 3kW AC & DC PSUs
(H) 3.45 x (W) 17.3 x (D) 24.7 in. (H) 8.76 x (W) 44.0 x (D) 62.7 cm 37 lbs (16.7 kg)

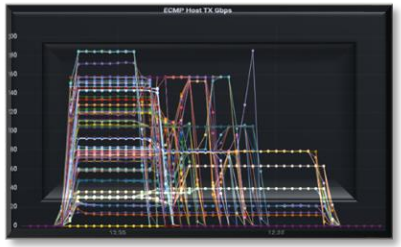
- 51.2T G200 Optimized for high-radix DC and AI applications
- ETC 800 GbE support
- Lowest power consumption per-bit
- 512 x 100G – full 512 interface network radix
- Exceptional Serdes performance for powering LPO optics

Perfect Load Balancing and Congestion Control Exists Today…

…*inside* distributed systems by Cisco's perfectly load-balanced, congestion-free fabrics.

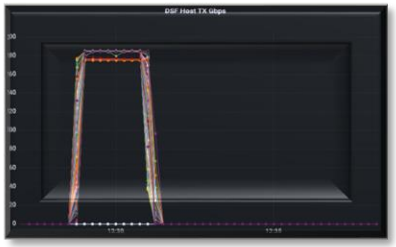
Todays hardware-based output-queued crediting mechanism used in our class-leading modular chassis can be applied at the network level with ***Cisco Scheduled Ethernet***, removing the network barrier to job completion time

Traditional ECMP Ethernet Traffic Pattern by Host



Tail Latency = Idle GPUs

Scheduled Ethernet Enabled Traffic by Host



Synchronizing Begins Sooner
Job Completes 2x Faster!

AI Primed

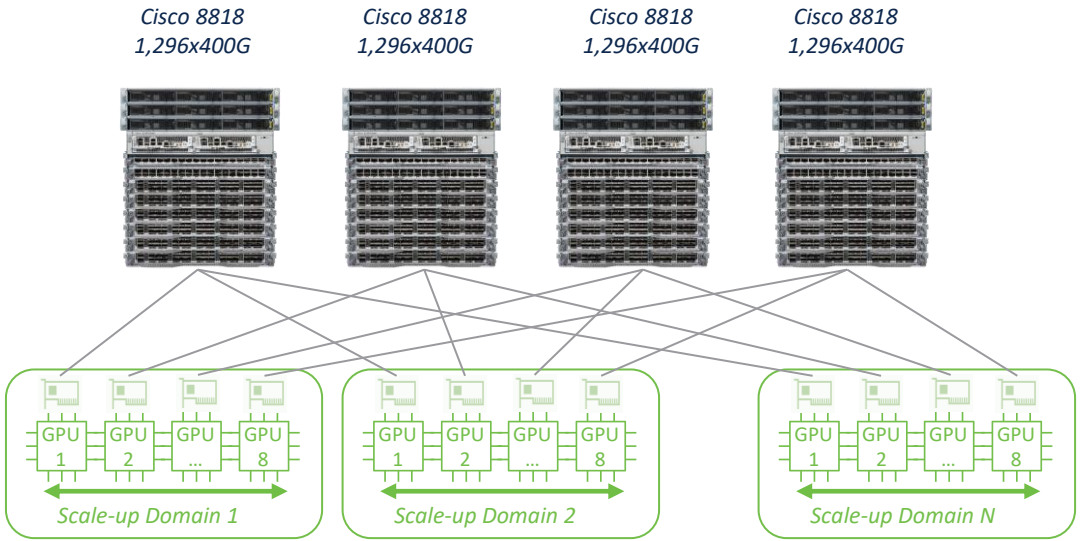
Register Right now for BRKNWT-3406 for a deeper dive into Scheduled Ethernet!

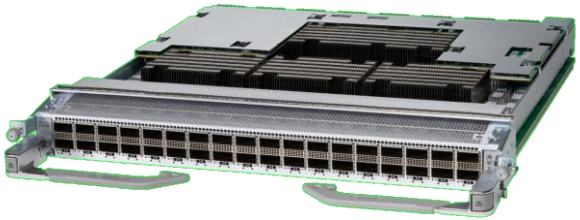
Scaling Out With Modular

All the benefits of Scheduled Ethernet on Rails!

36 x 800G
Available Now

cisco Silicon One P100



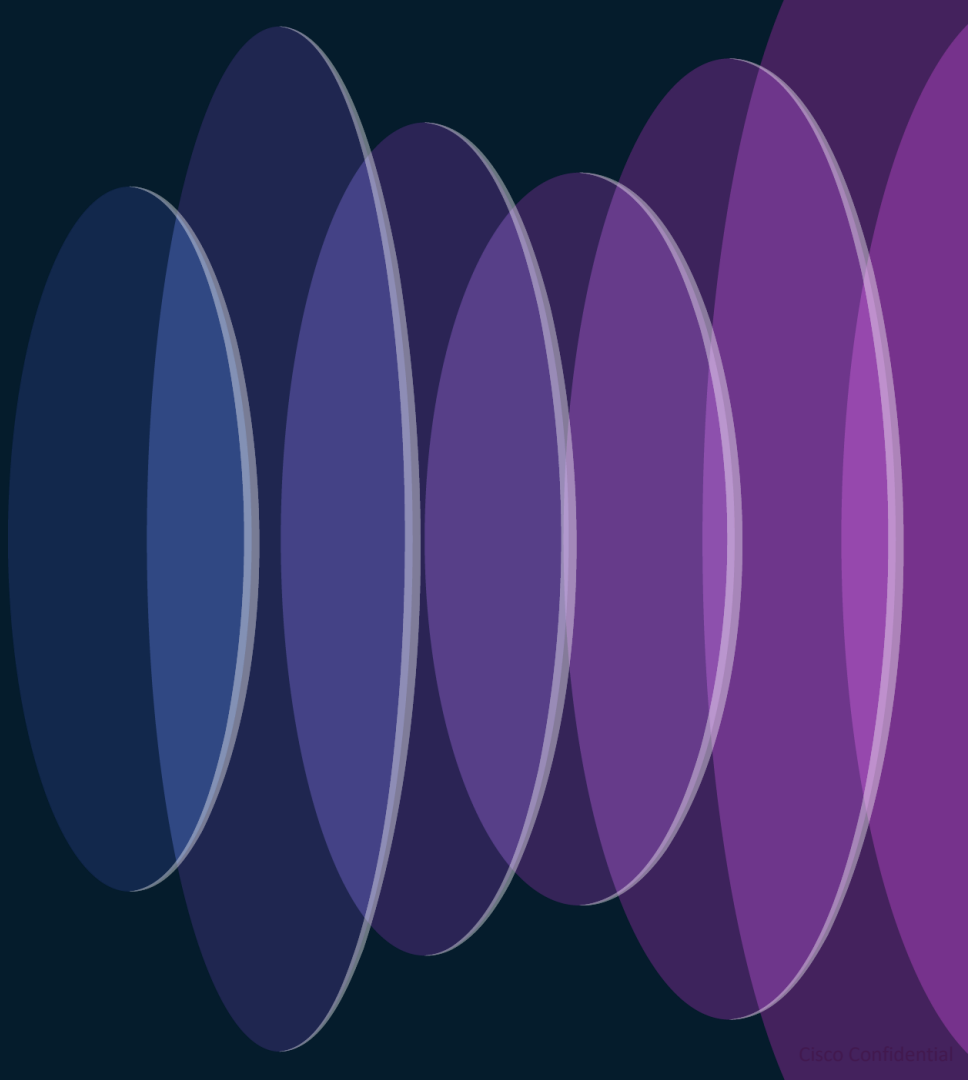


88-LC1-36EH-C

4x 19.2T P100 ASIC (7nm)
4x 72 MB SRAM / 4x 8 GB HBM packet buffer

10,368 x 400GE GPUs addressed with only 8x 8818's
Less optics, less cabling, Less Complexity and complete redundancy

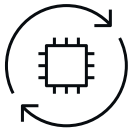
Addressing Trends

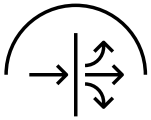


RoCEv2 as a Scale-out Transport Protocol

Good, but can be improved upon

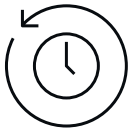
PFC Requires Intense Buffering





Victim flows, congestion trees, PFC storms and deadlocks

Go-back-N retransmission



Congestion control relies on pause and timeout

ULTRA ETHERNET VISION

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

THE NEW ERA NEEDS A NEW NETWORK

As **performant** as a supercomputing interconnect

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

As **scalable** as a cloud data center

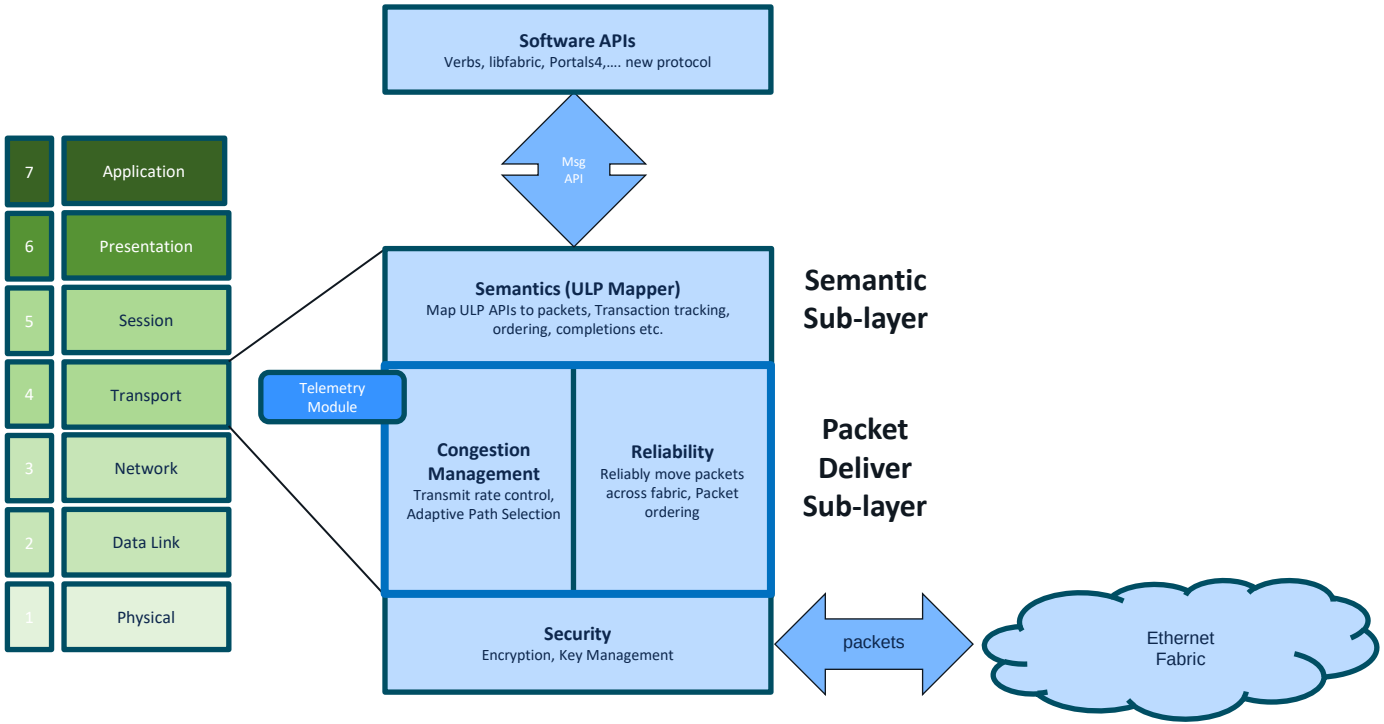
The UEC Seeks to Bring Open Standards to AI Networks

Open specifications, APIs, source code for optimal performance of AI and HPC workloads at scale.

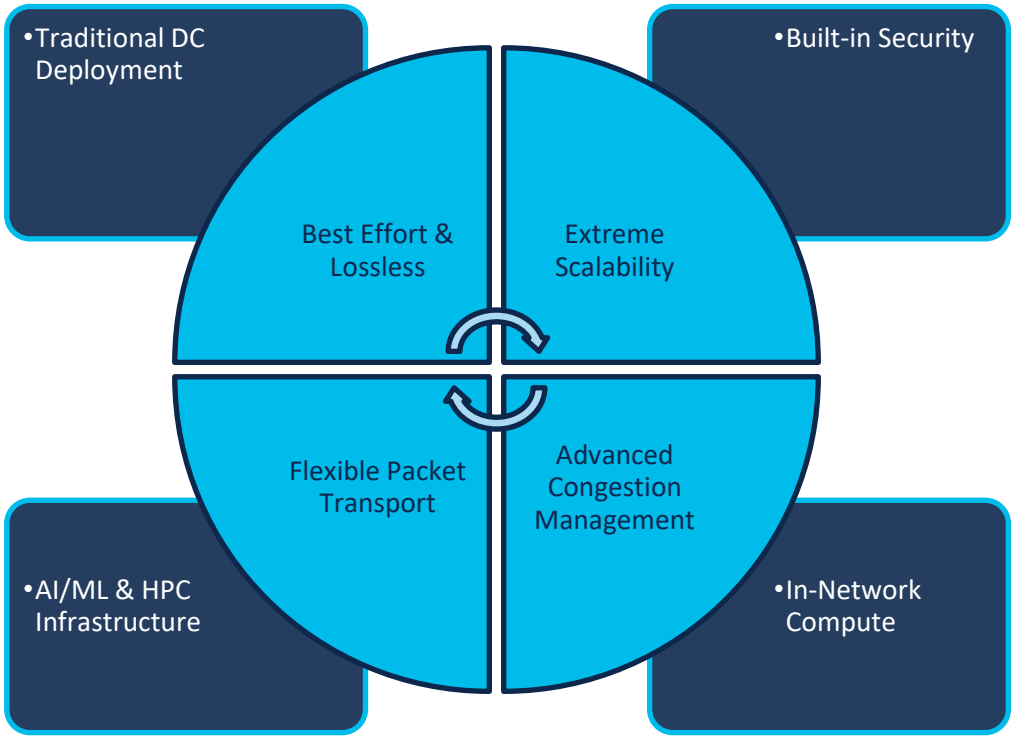
AMD ARISTA BROADCOM CISCO EVIDEN an atos business

Hewlett Packard Enterprise intel Meta Microsoft ORACLE

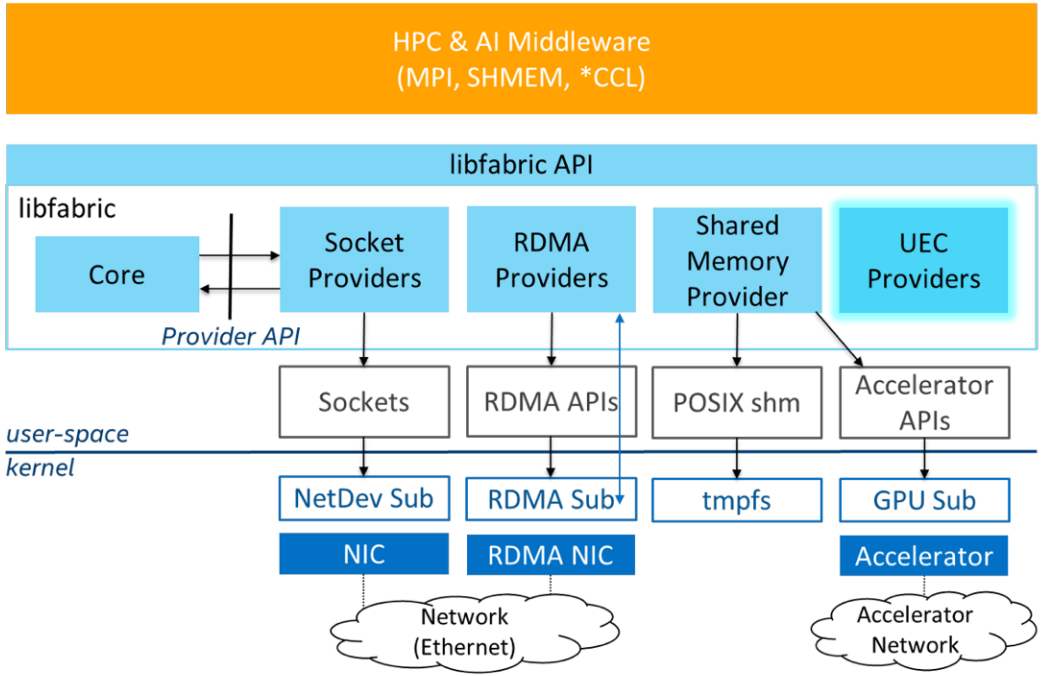
UE – Transport Layer



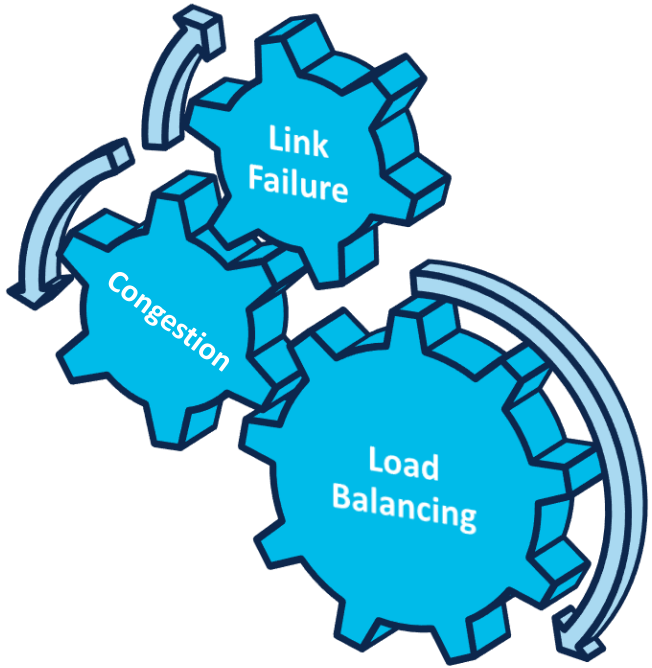
UET Key Features + Design Goals



UE – Software – Libfabric



Revisiting *the AI Infrastructure Challenge*



Wrenches in the works

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

Greasing the skids

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

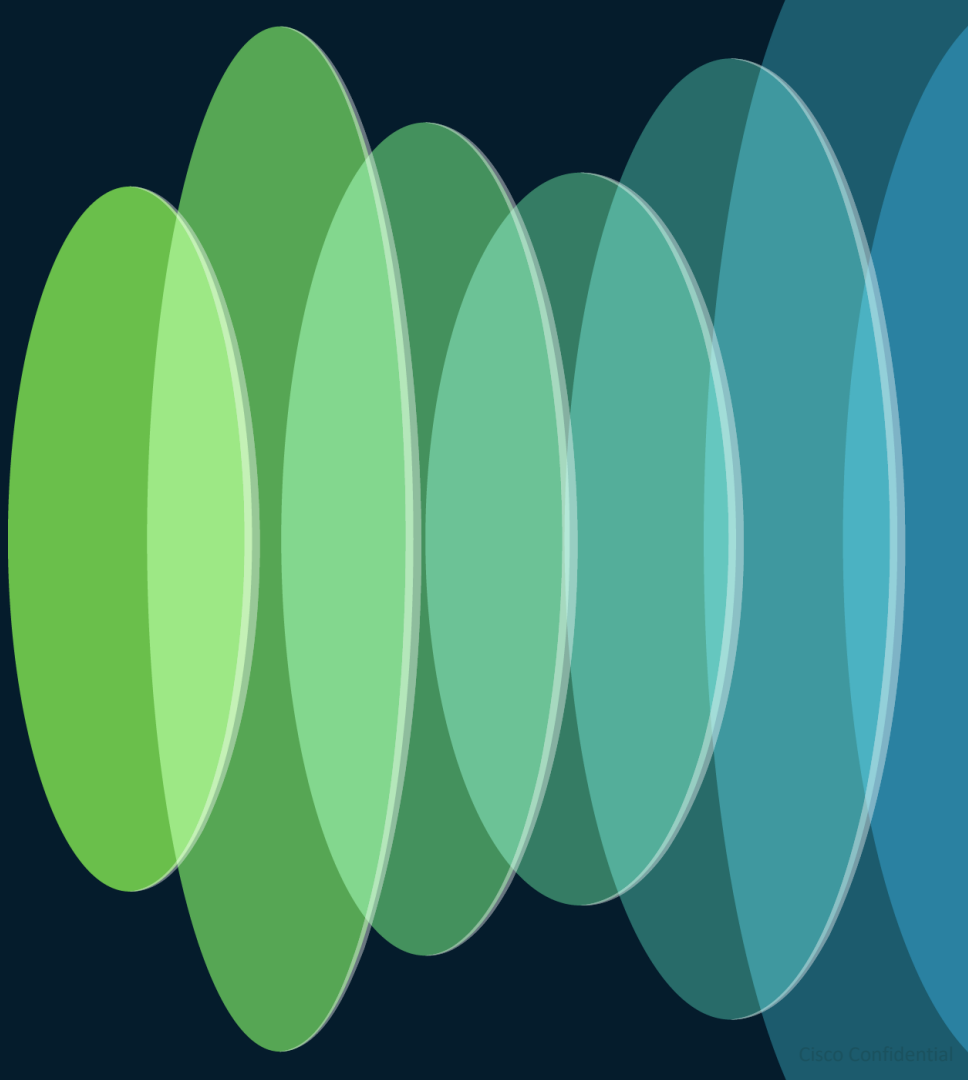
AI Ethernet Fabric Options

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

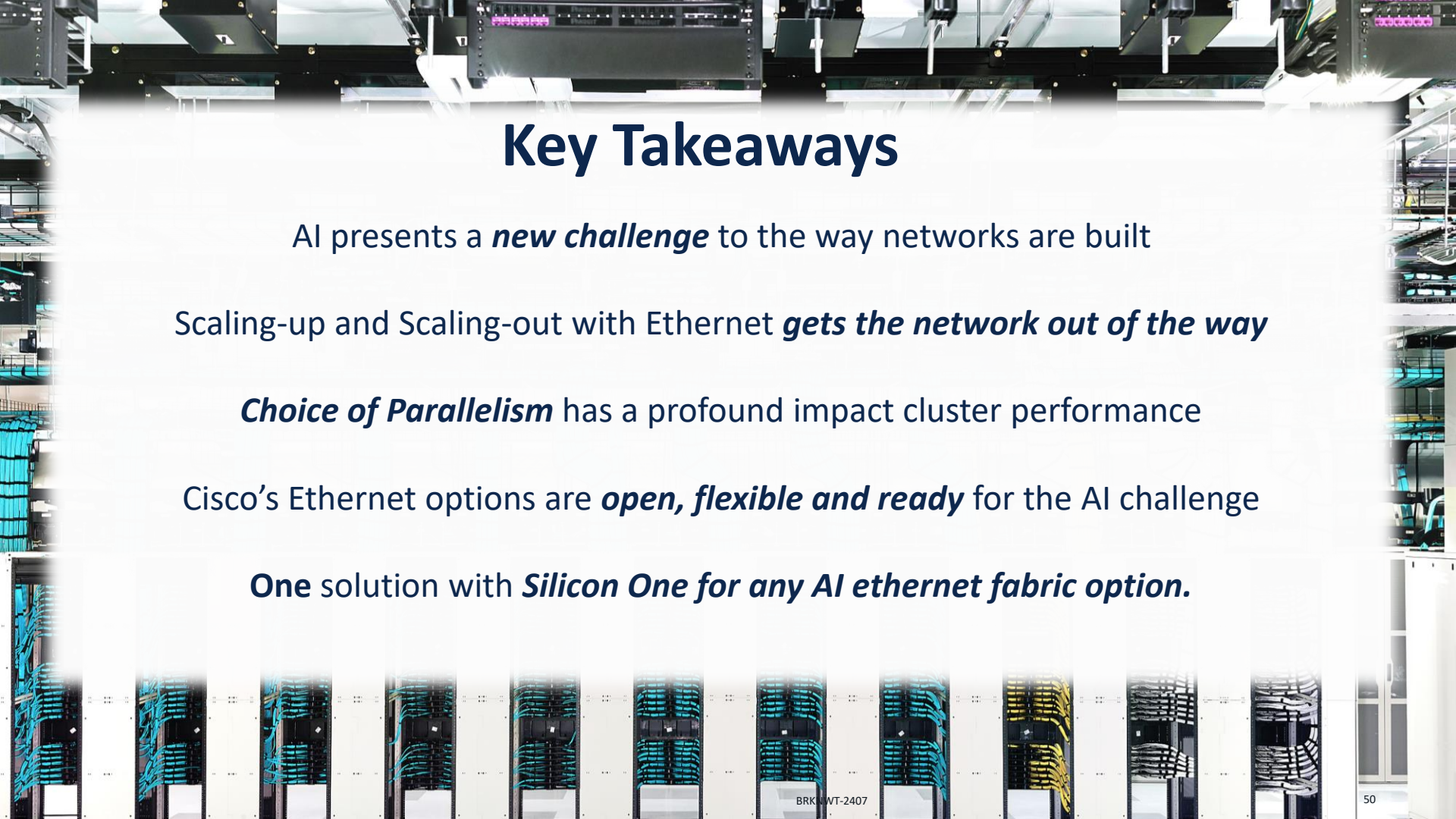
Effectiveness IS dependent on Traffic Characteristics

Effectiveness IS NOT dependent on Traffic Characteristics

Wrap-up and Questions



cisco Live!



Key Takeaways

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

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

Choice of Parallelism has a profound impact cluster performance

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

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

Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2025.

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

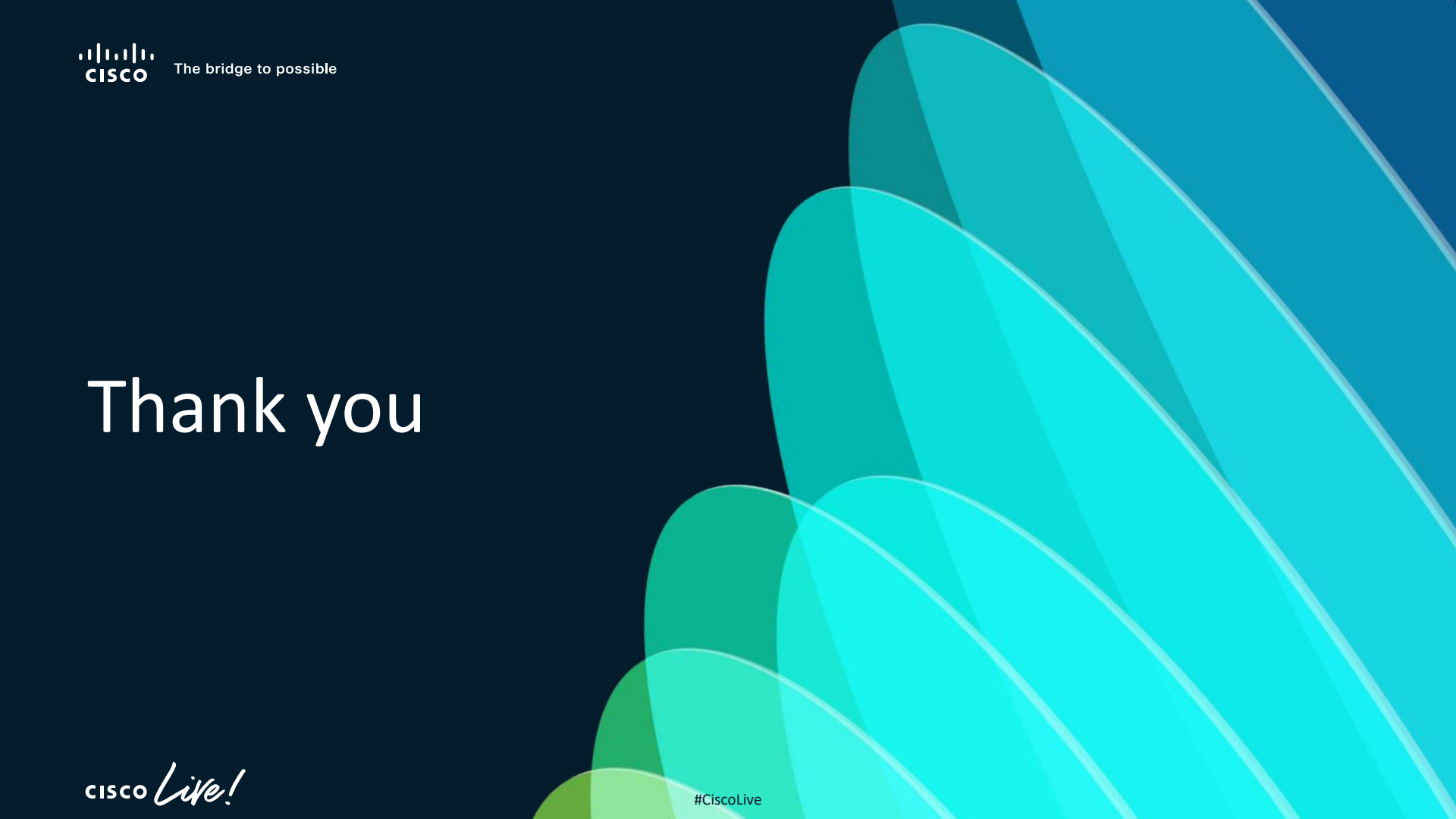
Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live mobile app.

Continue your education

- Visit the 8122 in the Product Zone!
- See the G200 in the AI Hub
- Book your one-on-one
  Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library
  for more sessions at www.CiscoLive.com/on-demand

Contact us at:
Ramesh Sivakolundu (sramesh@cisco.com)
Scott Carter (scocarte@cisco.com)


The bridge to possible
Thank you
cisco Live!
#CiscoLive