SNIA DNSF | DATA, NETWORKING, & STORAGE

Ethernet in the Age of AI: Adapting to New Networking Challenges

Live Webinar

November 19, 2024

9:00 am PT / 12:00 pm ET

Today's Presenters

Erik Smith
Co-Chair, SNIA Data, Networking & Storage Forum
Dell Technologies

Raguraman Sundaram
Software Architect
Celestica

SNIA DNSF | DATA, NETWORKING, & STORAGE

The SNIA Community

200
Corporations, universities, startups, and individuals

2,000
Active contributing members

50,000
Worldwide IT end users and professionals

SNIA DNSF | DATA, NETWORKING & STORAGE

What We Do

Drive the awareness and adoption of a broad set of technologies, including:

- Storage Protocols (Block, File, Object)
- Traditional and software-defined storage
- Disaggregated, virtualized and hyperconverged
- AI, including storage and networking considerations
- Edge implementation opportunities and factors
- Storage and networking security
- Acceleration and offloads
- Programming frameworks
- Sustainability

How We Do It

By delivering:

- **Expert webinars and podcasts**
- **White papers**
- **Articles in trade journals**
- **Blogs**
- **Social Media**
- **Presentations at industry events**

www.snia.org/dnsf

SNIA Legal Notice

- The material contained in this presentation is copyrighted by SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
    - Any slide or slides used must be reproduced in their entirety without modification
    - SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

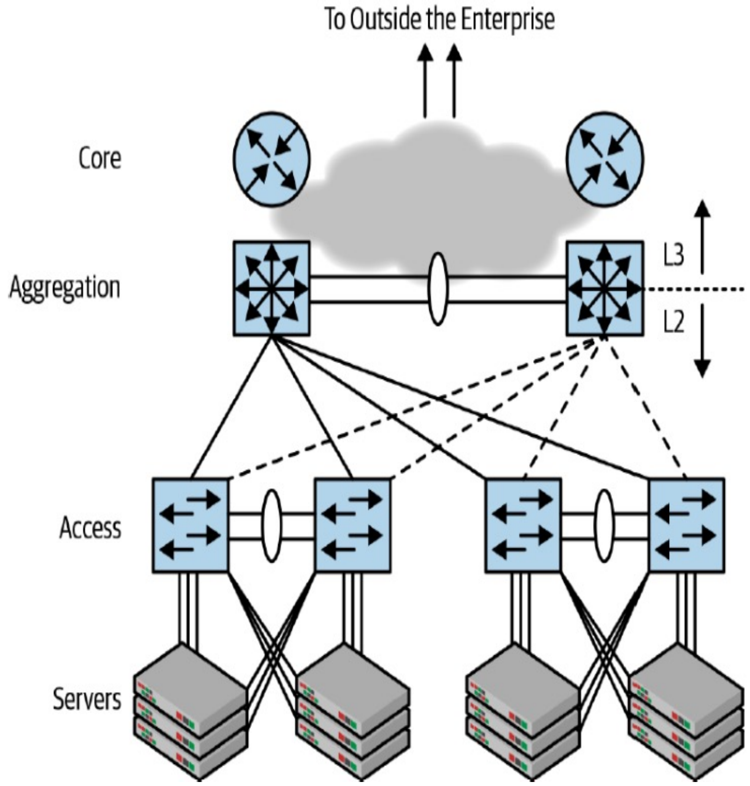
Today’s Agenda

- Overview of Data Center Networks
- LLM GPU Scale and Collective requirements
- Ethernet GPU Fabric Topology
- Ethernet GPU Fabric Requirements
  - Congestion Avoidance
  - Congestion Response

SNIA DNSF | DATA, NETWORKING, & STORAGE

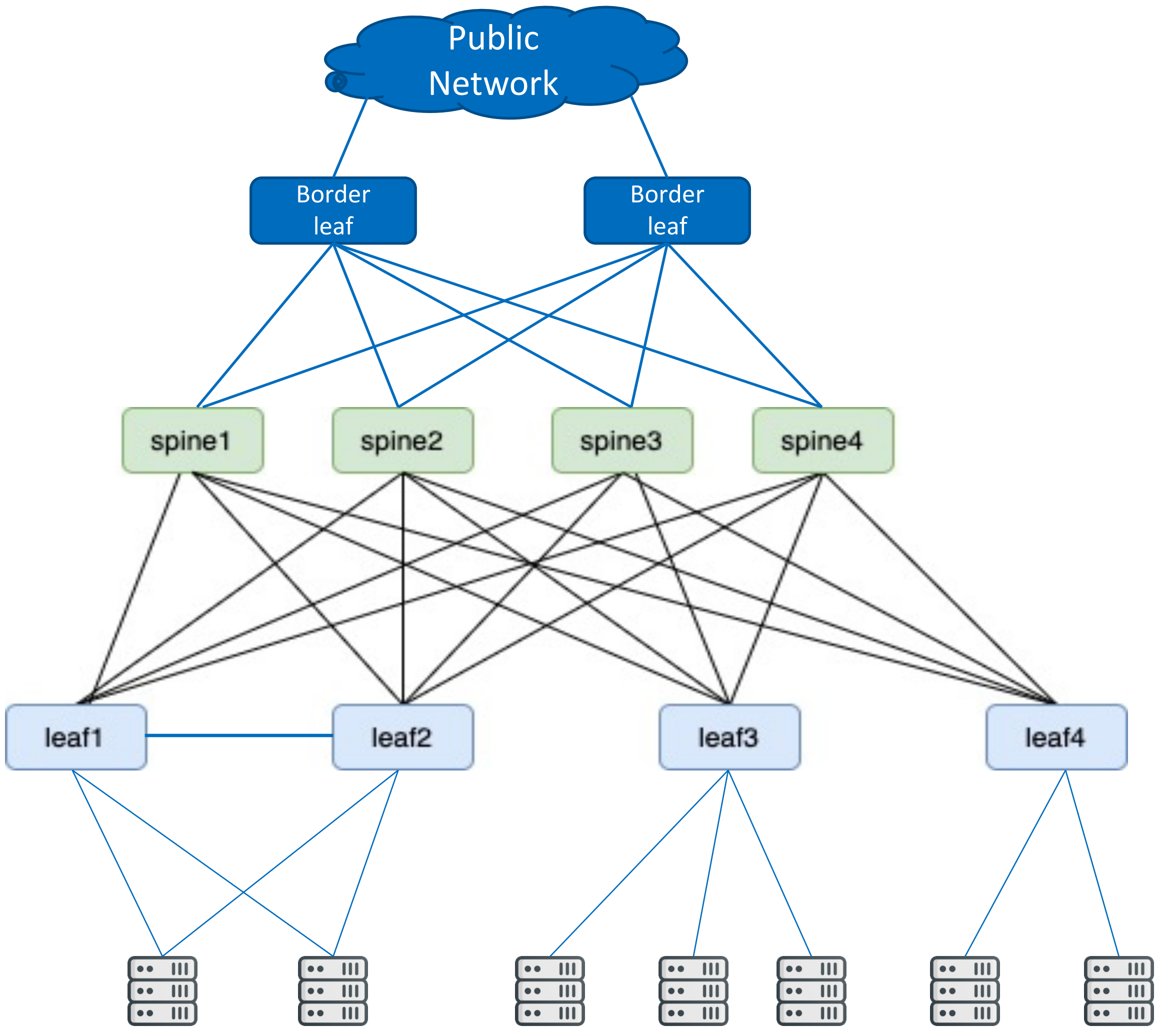
Data Center Networks

- Access, Aggregation and core
- North-South Traffic
- Single Layer-2 domain below aggregation
- Single link failure makes the BW to half
- 2- aggregation switch
- Spanning Tree Protocol
- Vlan cannot span



CLOS Network

- East-West traffic
- High Bandwidth
- Load Balancing - ECMP
- Vxlan
- Multitenancy
- High reliability and fast failure recovery
- Easily scalable

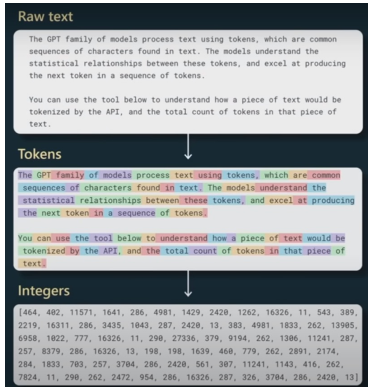


GPU Scale – The LLM Connection

- Large Language Models (LLMs) are type of generative AI trained on vast natural language data using deep learning algorithms
- Most models stick to open data sets for training
- Tokenization translate the raw words into a sequence of integers (tokens).
- A typical data set can contain hundreds of billions to trillions of tokens
- Weights in LLM are learned variables that dictate how the model interprets and generate languages

GPU Scale – The LLM Connection

- Model Math
  - For 175B parameters and 300 B tokens and 6 FLOPS per token per parameter ~ 3.15 x10^23
  - A GPU with ~67 TeraFLOPS per sec it takes 4.7x10^9 seconds
  - To finish the training in 1 month it would take ~1800 GPUs

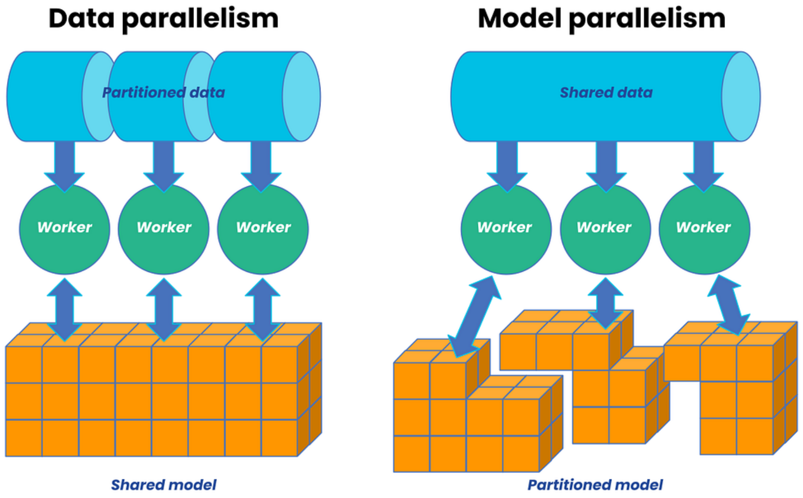


	GPT-3	LLaMA
Sequence length	2k	2k
Parameters	175B	65B
Tokens	300 B	1-1.2 Trillion
Training Time	1 month	21 days

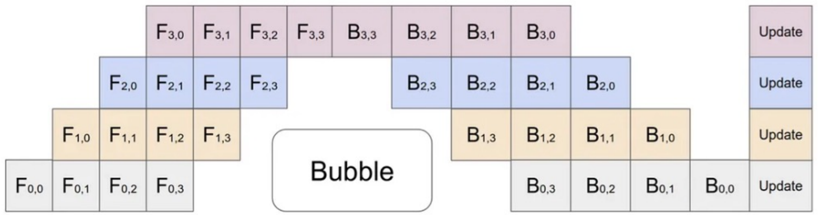
- Model with 175B parameters require greater 1TB of memory
- Storage required checkpoint training state typically around 4TB
- Typical High end GPU has 80 GB of High Bandwidth Memory
- One GPU cannot fit the model parameters or training sets.

SNIA DNSF | DATA, NETWORKING, & STORAGE

Parallelism



(Image source: anyscale.com).



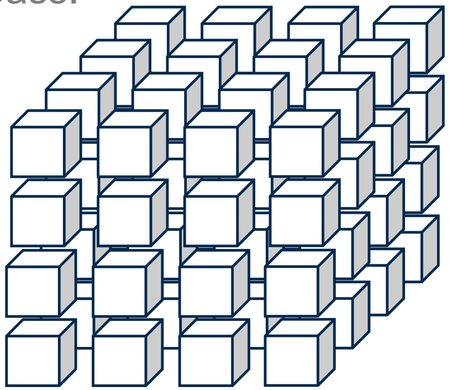
Model + Data Parallelism (GPipe: Huang et al. 2018)

- Data Parallelism
  - Data Set is split into Mini-Batch(s) and each GPU with a copy of the Model make forward pass for predictions and Backward pass for Gradients
  - End of one iteration Gradients are aggregated(averaged)
  - The aggregated gradients are broadcast to all GPUs.
  - Repeat the process till convergence
- Model Parallelism
  - Model is split into several partitions one per GPU.
  - During Forward pass each GPU computes the output and pass on as input to the next GPU
  - Backward Pass each GPU pass on the gradient to the previous GPU in sequence
  - Both Forward/Backward pass create sequential dependency
- Pipeline Parallelism
  - Combines both Data and Model parallelism
  - Dataset is further divided into Micro-batches and GPU works on a micro-batch and its model partition
  - Instead of waiting for backward pass, it starts to process the next micro-batch
  - Increases inter-GPU communication

SNIA DNSF | DATA, NETWORKING, & STORAGE

Parallelism

- Tensor Parallelism
  - Distributes compute intensive parts of the model/pipeline layer by splitting the computation across GPU
  - Relies on nodes actively scattering and gathering interim results
  - Reduces computation/storage requirement per GPUs for the model training/inference
  - Increases inter-GPU computation significantly compared to pipeline parallelism
  - Communication strategy would depend on the computation being split. For instance, a matrix multiplication split by row (column) would require All-Reduce (All-Gather) in the forward pass.



- Regardless of which ever parallelism is used it is evident that the inter GPU communication is significant.
- Any Congestion or Drops in the GPU fabric would result in poor performance
- Latency and Link utilization will also play significant role in performance numbers.

Collectives

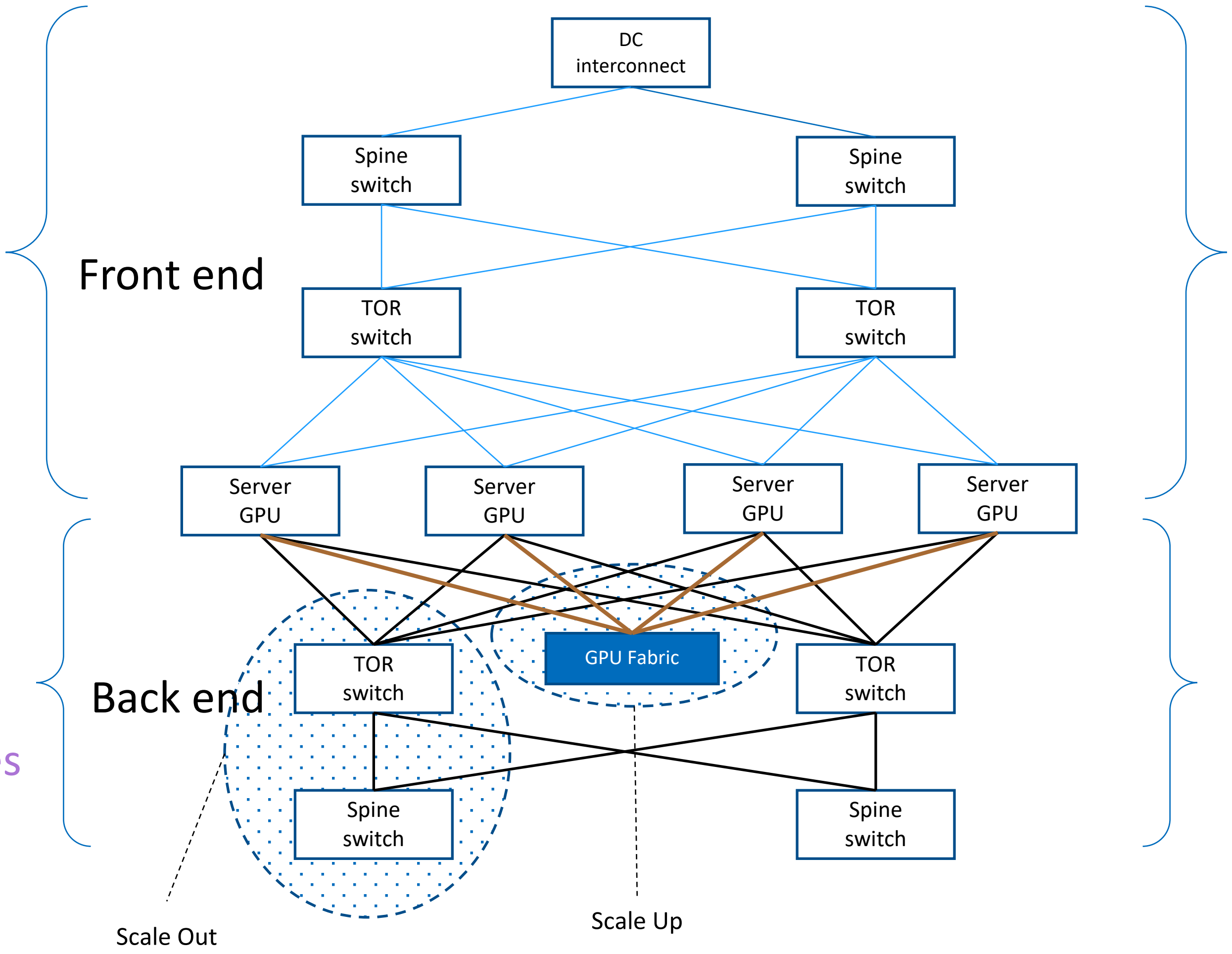
Collectives are set of operations involving communication among group of GPUs to perform a task.

- Reduce
  - Aggregating data from all member and send the result to one member
- All Reduce
  - Aggregating data from all member and send the result to all member
- Scatter
  - Distribute different values from one member to all member
- Reduce Scatter
  - Aggregate data from all member and scatter the results(unique subset of result) to all member
- Broadcast
  - Sending data from one member to all the member in the group
- All Gather
  - Gather all data and distribute it among all members
- AlltoAll
  - Scatter data from all members to all members

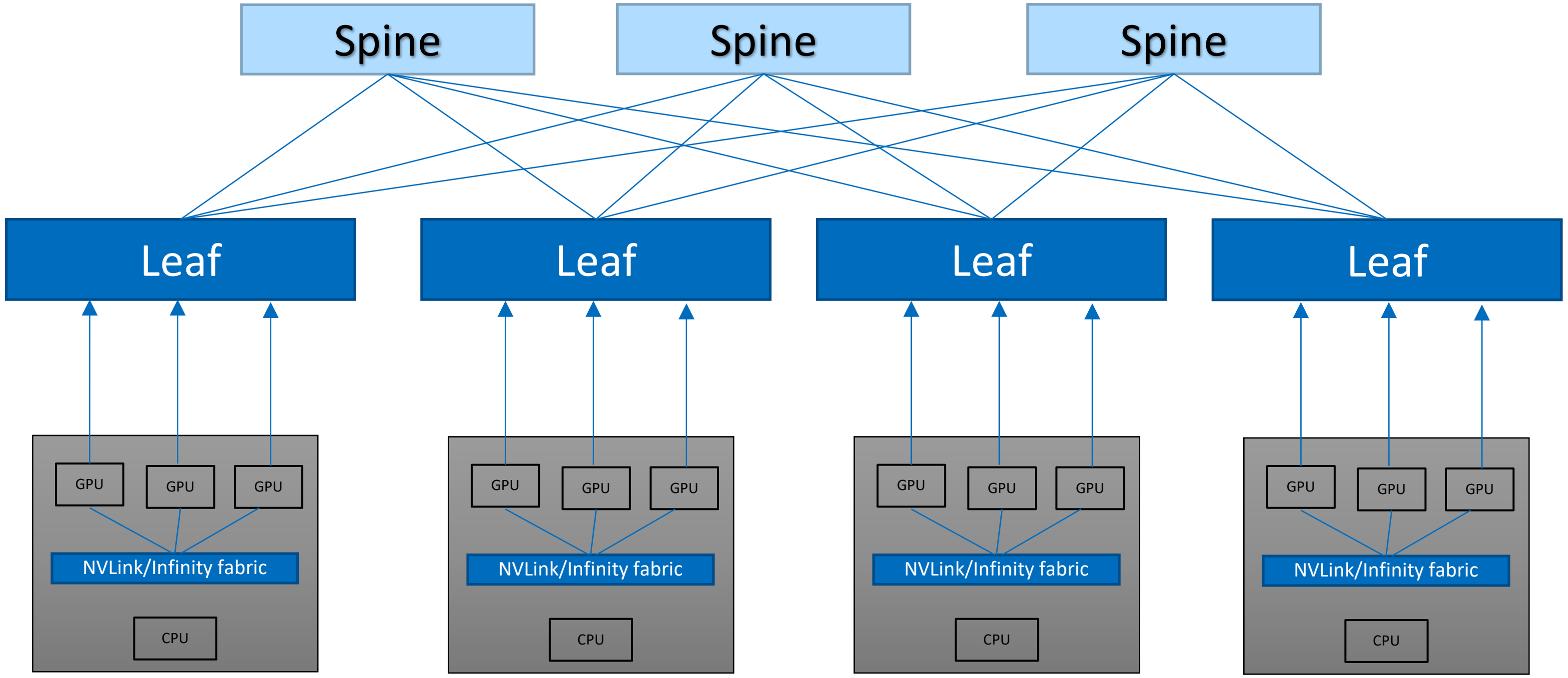
GPU Scale – GPU Fabric Backend Network

- With the GPU scale required for training there is need for an network to connect thousands of GPU
- There are three network in play here. The primary network is called front end network
- The GPU network is Back end network
  - Scale up: Full mesh NVLink/infinity fabric connections between the GPUs of the same server or same rack
  - Scale out: The network connects thousands of GPU in datacenter across racks

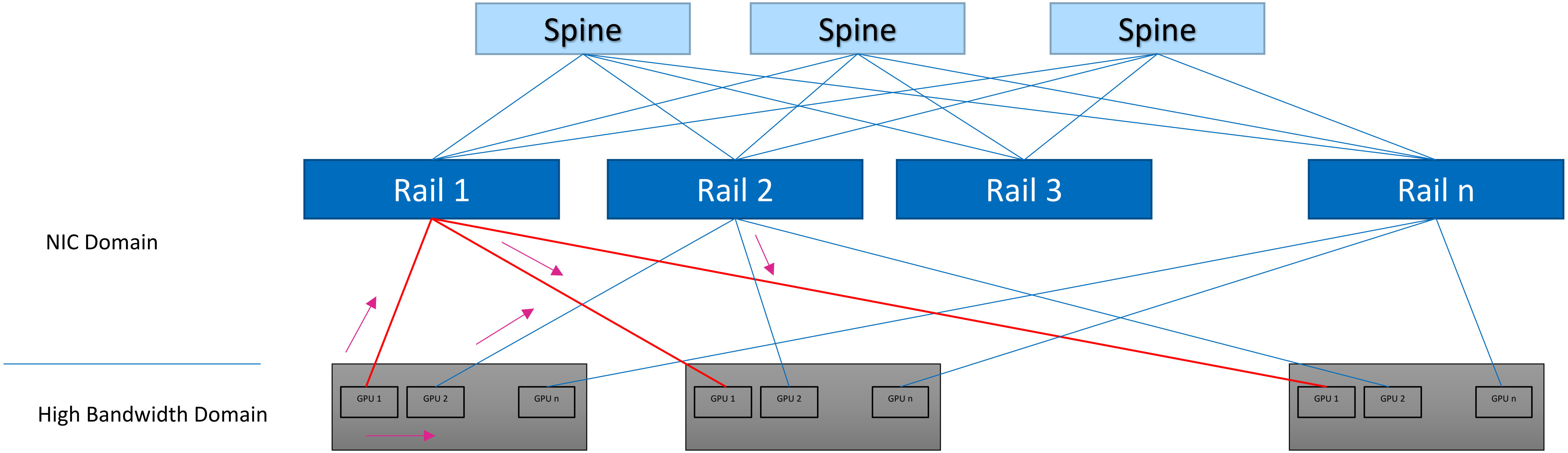
- For Scale Out network Ethernet becomes one of the primary choice
- Because Ethernet is modular, scalable, support high speed, cost effective and works with existing infrastructure
- Ethernet by nature is lossy technology. There are challenges in the area of congestion, latency and utilization etc.,



Scale up and Scale out GPU Fabric

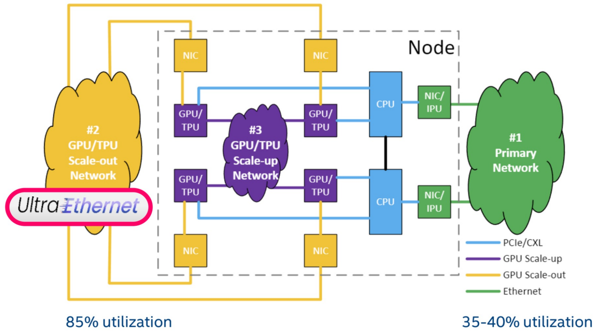


Rail Optimized Topology



AI Networking Characteristics and Challenges

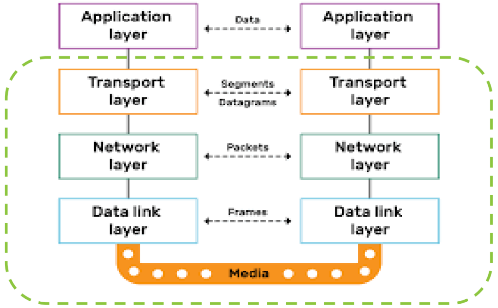
AI/HPC Networks of Interest: Basic Characteristics



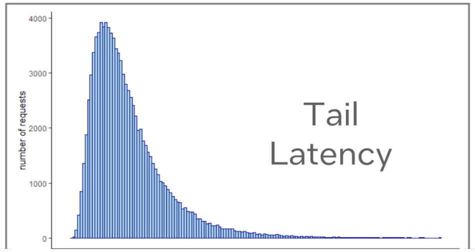
1. **Primary DC network**
   - Used by all 3 deployment models
   - Main network for some HPC At Scale
   - Very large scale: up to 100K-1M Endpoints
   - Distance: >150m ; RTT ~100 uS +; BW/GPU **~10GB/S**
   - Storage attached e.g., over RoCE RDMA
   - Network semantics
2. **GPU/TPU Scale-Out Network**
   - DL/Inference Cluster -10k nodes and ↗
   - Distance: <100m ; RTT <10 uS + ; BW **~100GB/S**
   - Main network for some HPC At Scale
   - Network semantics
3. **GPU/TPU Scale-Up Network**
   - Within a node; small scale e.g., 256 XPU?
   - Distance: ~1m ; RTT ~1 uS +; BW **~1200 GB/S**↗
   - Direct connect and/or switched
   - Memory and Network semantics

- Optimal Link utilization
- Load Balancing – Path aware, Adaptive, Lossless
- Loss Retransmission
- HPCC
- Low latency

AI/HPC Common Requirements



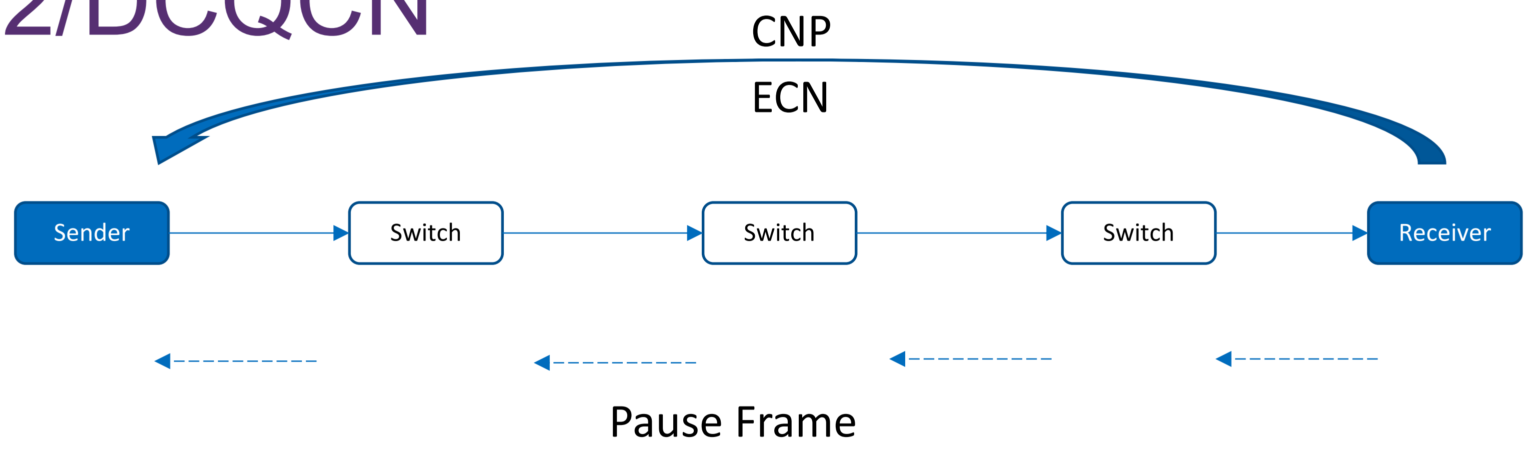
- Transport primitives for
  - Large Scale
  - Multi pathing
  - Relaxed ordering
  - Modernized Congestion Control
  - Optimized RDMA
  - Performance – bandwidth, latency, tail latency, Packets/S
  - High network utilization
  - Stability and Reliability



Key goals: high utilization & low tail latency!

SNIA DNSF | DATA, NETWORKING, & STORAGE

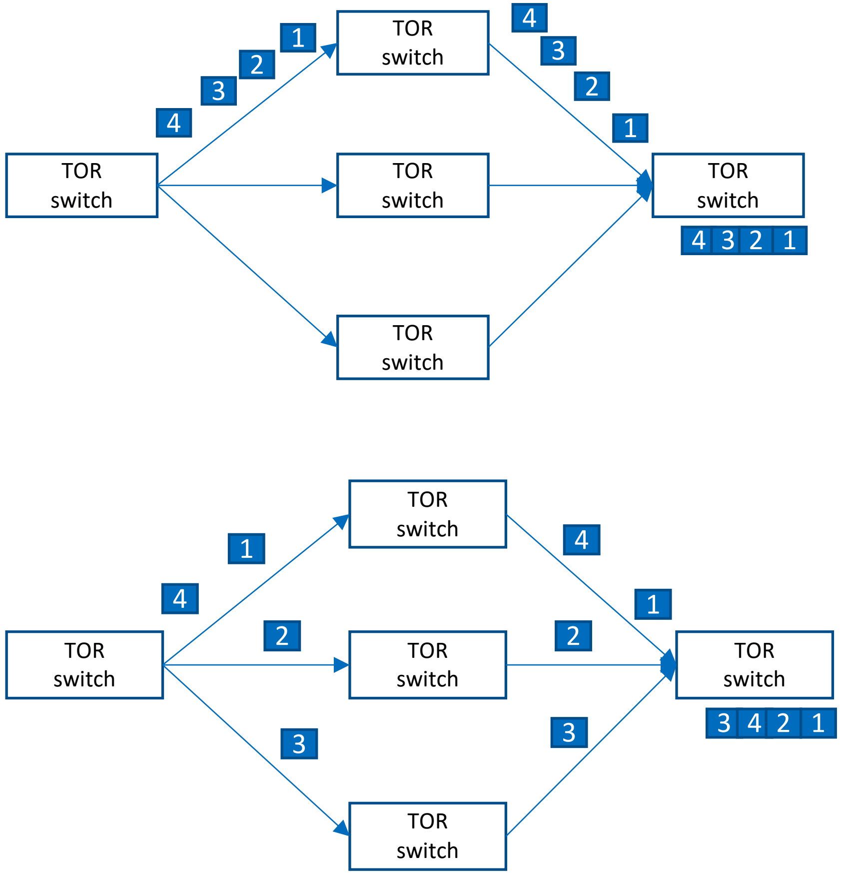
RoCEV2/DCQCN



- RoCEV2 – RDMA over Converged Ethernet (L3)
- Uses DCQCN as the congestion control mechanism which combines PFC and ECN for congestion management in the RDMA networks
- Enable Lossless ethernet traffic
- PFC is port based not per flow based. Non congested flows can be affected by a congested flow on the port
- PFC storm
- ECN is not per flow based.
- Start/stop nature of traffic increases latency

Link Utilization/Packet Spraying

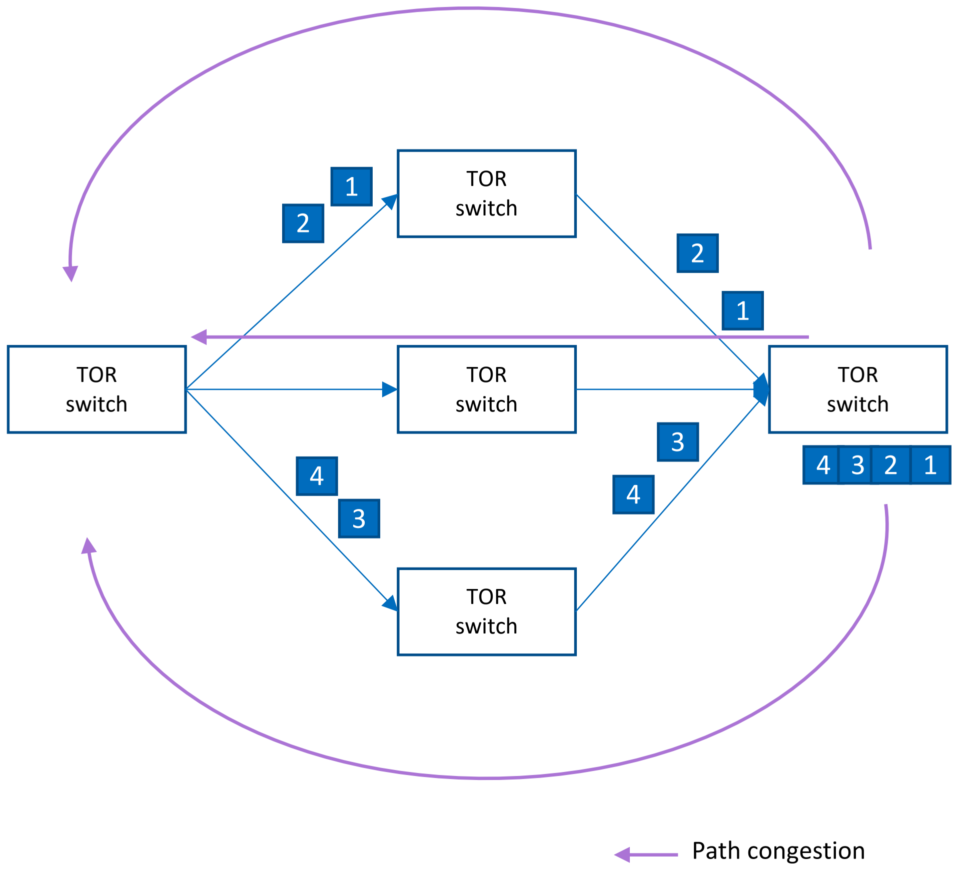
- In traditional ECMP hashing the packet fields are hashed to pick a path, the path for the flow is constant and ensure in-order delivery.
- Under utilizes the paths available to the destination.
- An elephant flow can create congestion on the path, even other paths are available
- Packet spraying is not new to the switches. Sprays the packets from the same flows to available paths
- Receiver will get out of order delivery based on the path length and congestion status.
- Receiver should be capable of arrange the packets in the right order.
- Utilizes the paths in a balanced way better than before
- Spraying can happen from the NICs or the leaf switches
- Need sophisticated hardware for out of order reassembly
- Need careful configuration as it might affect other schemes?



SNIA DNSF | DATA, NETWORKING, & STORAGE

Adaptive Load Balancing/ Path aware CC

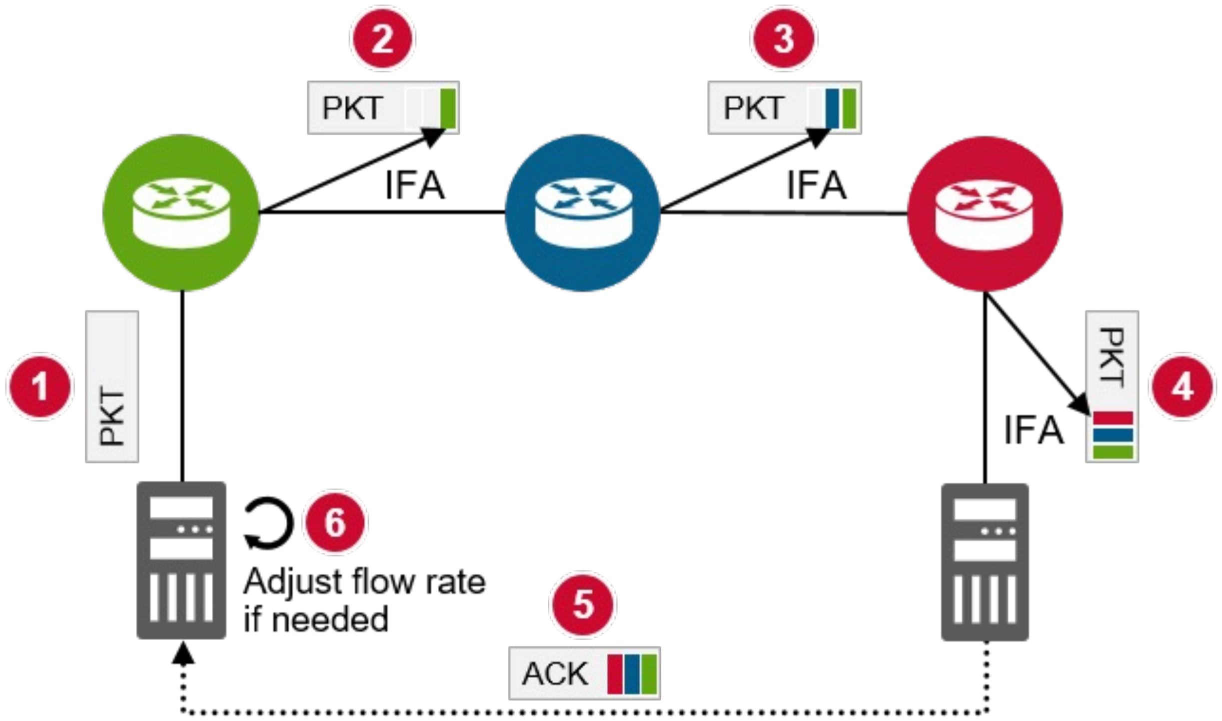
- Real time telemetry about the congested paths
- Dynamically avoid congested paths
- Traditional load balance uses only that nodes queue depths whereas in this case the entire path congestion status is used for load balancing
- During transient congestion the same flow can be load balanced to different path. Out of order delivery to be handled.
- Careful tuning is required as it might cause unwanted out of order packets



SNIA DNSF | DATA, NETWORKING, & STORAGE

INT/CSIG - HPCC

- In band Telemetry (INT)/Congestion signal(CSIG) carry fine-grained network signals for congestion control and traffic management to the end host
- In INT every switch along the path add meta data about the congestion on the node to the packet
- CSIG provides a simple, low-overhead, and extensible packet header mechanism to obtain fixed-length summaries from bottleneck devices along a packet path
- The end host can reflect this congestion information to the source real time to adjust the rate and the window size per flow
- Can help to ramp up higher speeds at faster rate
- IFA/CSIG capable hardware is needed for the entire path. Cost would be an issue



Source : https://www.broadcom.com/blog/high-precision-congestion-control

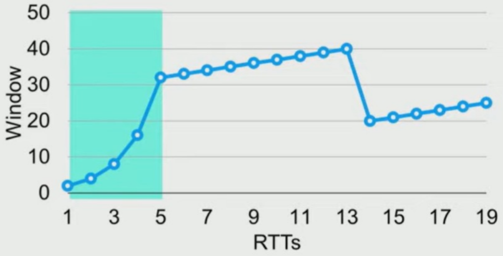
Reference :
https://www.ietf.org/archive/id/draft-ravi-ippm-csig-01.html

Congestion Control

Congestion Control (CC)

...is required, (in addition to optimal load balancing on multiple paths)

- How is UET CC different from TCP?
  - High bandwidth, short RTT (10us)
  - Get to wire rate very quickly
    - 1MB takes 10 usec at 800gbps = 1 RTT
    - No time to wait for TCP slow start
  - And back off quickly when congestion is noticed

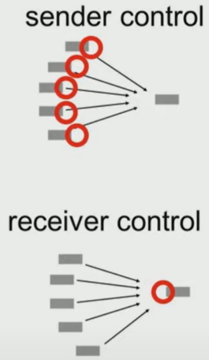


UEC Congest Control is Designed for short RTT, high BW

UET Congestion Control

Two flavors - that can work together

- **Sender-based** (default)
  - fast ramp, fast slowdown
  - uses delay as a measure of congestion
- **Receiver-based** (optional)
  - receiver credit manages receiver buffer
  - simple rate-control for oversubscribed networks

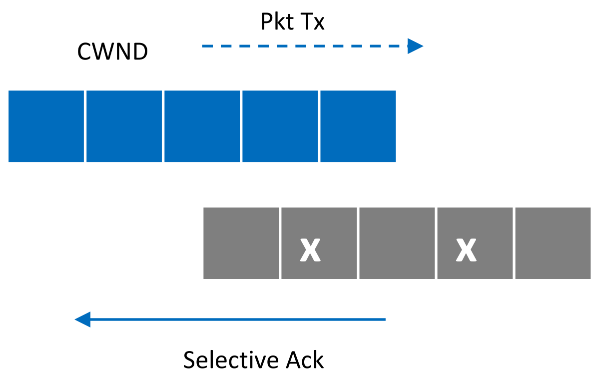


Both are designed for AI and HPC workloads and multi-pathing

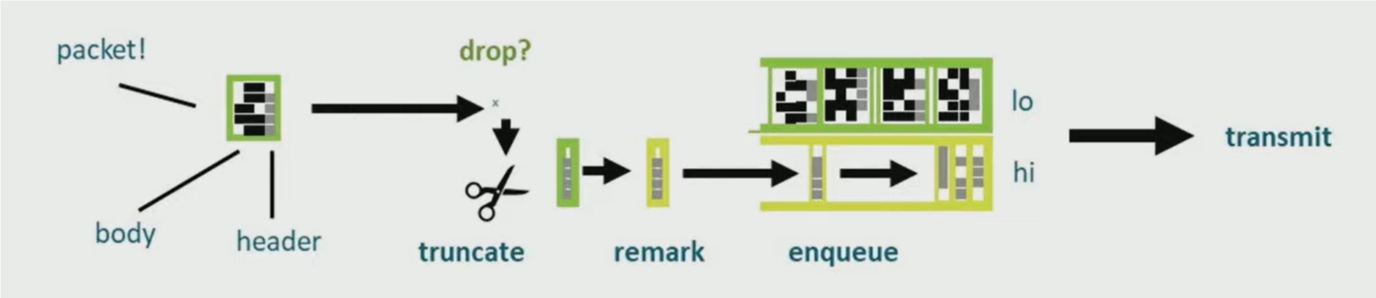
SNIA DNSF | DATA, NETWORKING, & STORAGE

Loss Retransmission

- Enhanced network performance through in-order message delivery and selective acknowledgment (SACK) retransmission.
- Unlike RoCEv2's Go-back-N mechanism, which resends all packets from the point of failure, SACK allows the receiver to identify and retransmit only lost or corrupted packets.
- This targeted approach optimizes bandwidth utilization, reduces latency in packet loss recovery and minimizes redundant data transmission
- Faster job completion times, lower tail latencies, and more efficient bandwidth use



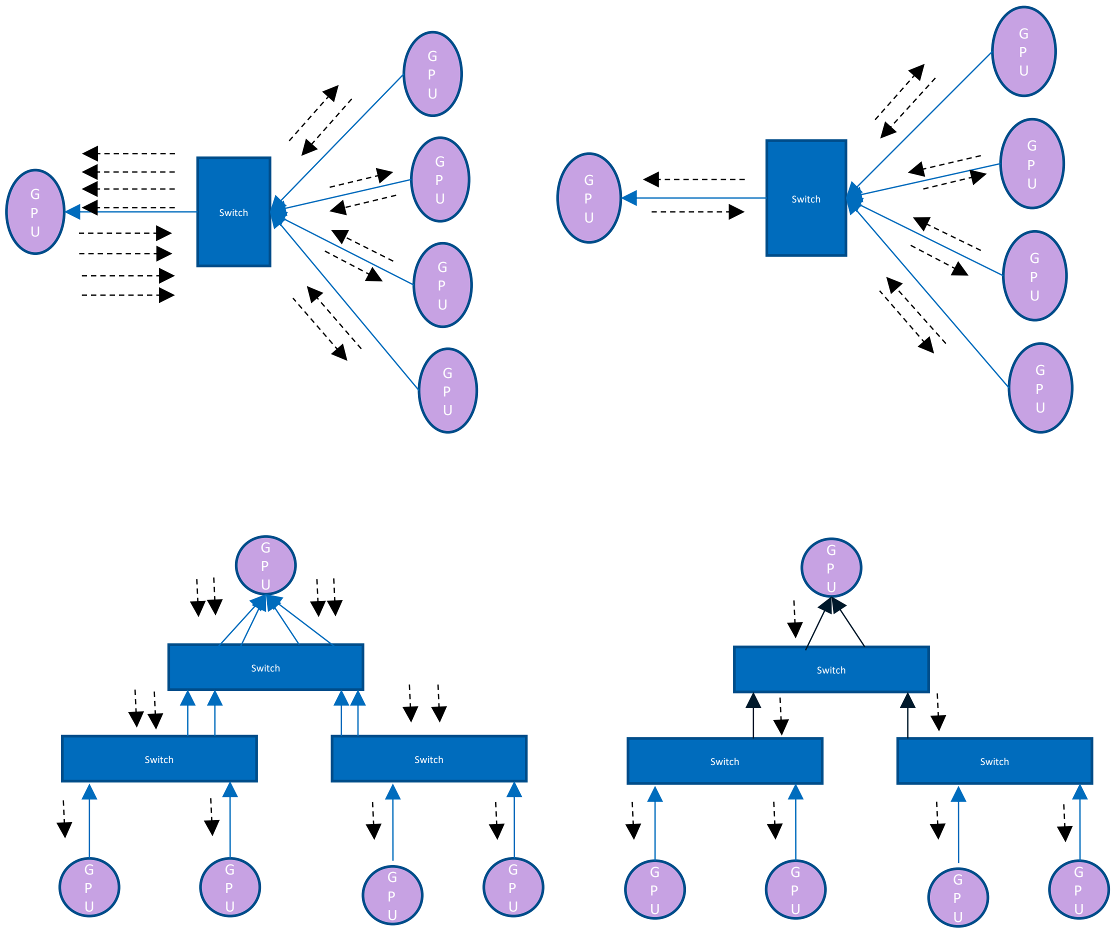
Packet Trimming



- Fast loss detection
- Trim the packet to 64 bytes instead of dropping
- Mark the DSCP to trimmed for the receiver to identify
- Send via the high priority queue
- This would be useful for the receiver to identify the packet loss faster and send SACK to the sender for retransmission

INC – In Network Collectives

- It is a method to offload collective operations to the switches instead of GPUs
- Offloading collectives to network devices reduces traffic bottlenecks and enhance the performance
- Implemented in form of aggregate tree where each node in the tree aggregate the downstream flows and forward the data to the upstream node
- By design avoids in cast congestion during reduce collective operations.



SNIA DNSF | DATA, NETWORKING, & STORAGE

Requirement	UEC Transport	Legacy RDMA	UEC Advantage
Multi-Pathing	Packet spraying	Flow-level multi-pathing	Higher network utilization
Flexible Ordering	Out-of-order packet delivery with in-order message delivery	N/A	Matches application requirements, lower tail latency
AI and HPC Congestion Control	Workload-optimized, configuration free, lower latency, programmable	DCQCN: configuration required, brittle, signaling requires additional round trip	Incast reduction, faster response, future-proofing
In Network Collective	Built-In	NONE	Faster Collective operation, lower latency
Simplified RDMA	Streamlined API, native workload interaction, minimal endpoint state	Based on IBTA Verbs	App-level performance, lower cost implementation
Security	Scalable, 1st class citizen	Not addressed, external to spec	High scale, modern security
Large Scale with Stability and Reliability	Targeting 1M endpoints	Typically, a few thousand simultaneous end points	Current and future-proof scale

Summary

- The size of number of tokens in data set and model parameters in the LLM training require high number of GPUs
- GPU Fabric to scale up to some extent beyond that Scale out network is needed
- Ethernet is one of the primary option for Scale out network because of its proven advantages
- CLOS and Rail-optimized Scale out topology
- The traffic pattern generated because of the AI workloads demand new methods to
  - utilize bandwidth better
  - reduce latency and
  - congestion control
- UEC is working on a UE Transport specification to address the AI workload demands

SNIA DNSF | DATA, NETWORKING, & STORAGE

References

- Large Language Models - The HW connection
- GPU Fabrics for Gen AI Workload
- Rail Optimized Topology - Meta Paper
- SwitchAgg: A Further Step Towards In-Network Computation

UEC

- UEC Introduction
- Networking for AI and HPC, and Ultra Ethernet

Q&A

After this Webinar

- Please rate this webinar and provide us with your feedback
- This webinar and a copy of the slides are available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webinar, including answers to questions we couldn’t get to today, will be posted on our blog at https://sniansfblog.org/
- Follow us on X @SNIADNSF

Thank You