RDMA over Ethernet for Distributed Training at Meta Scale

Presented by Yang Zhou
Tue 10/14

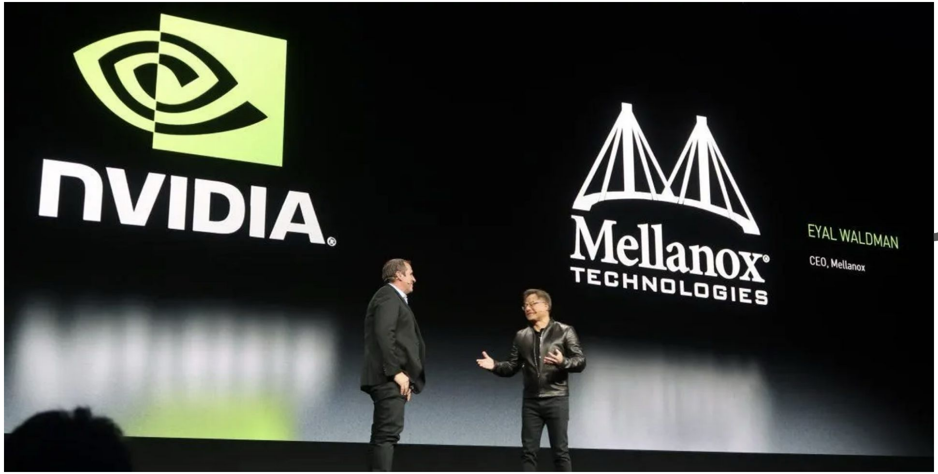
Trend: ML workloads are going distributed

Larger and larger training scale

More disaggregated serving

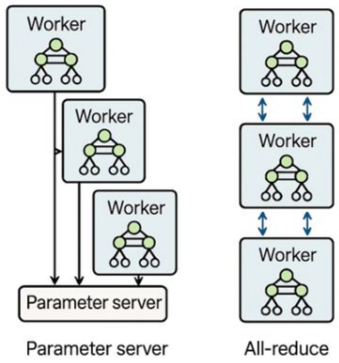
Massive expert parallelism in MoE

… as a result:

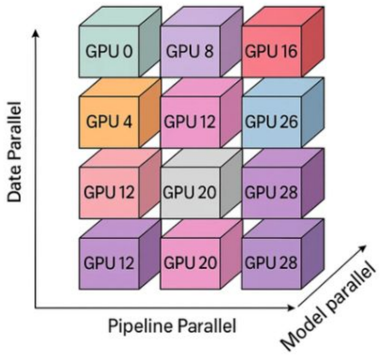


- NVL72, NVL144
- 800Gbps NIC, 102.4Tbps switch
- Co-packaged optics (CPO)

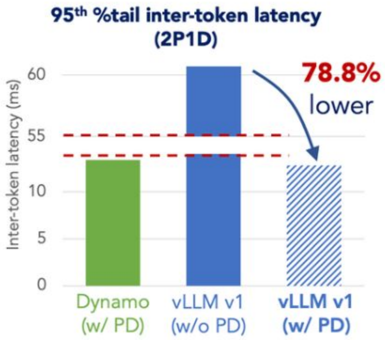
Fast Evolving ML Workloads



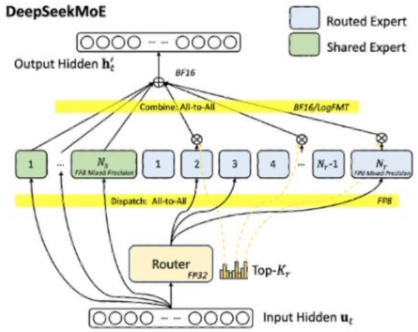
DNN training: parameter servers, allreduce



LLM training: allreduce, allgather, and reduce-scatter



PD disaggregation: P2P transfer



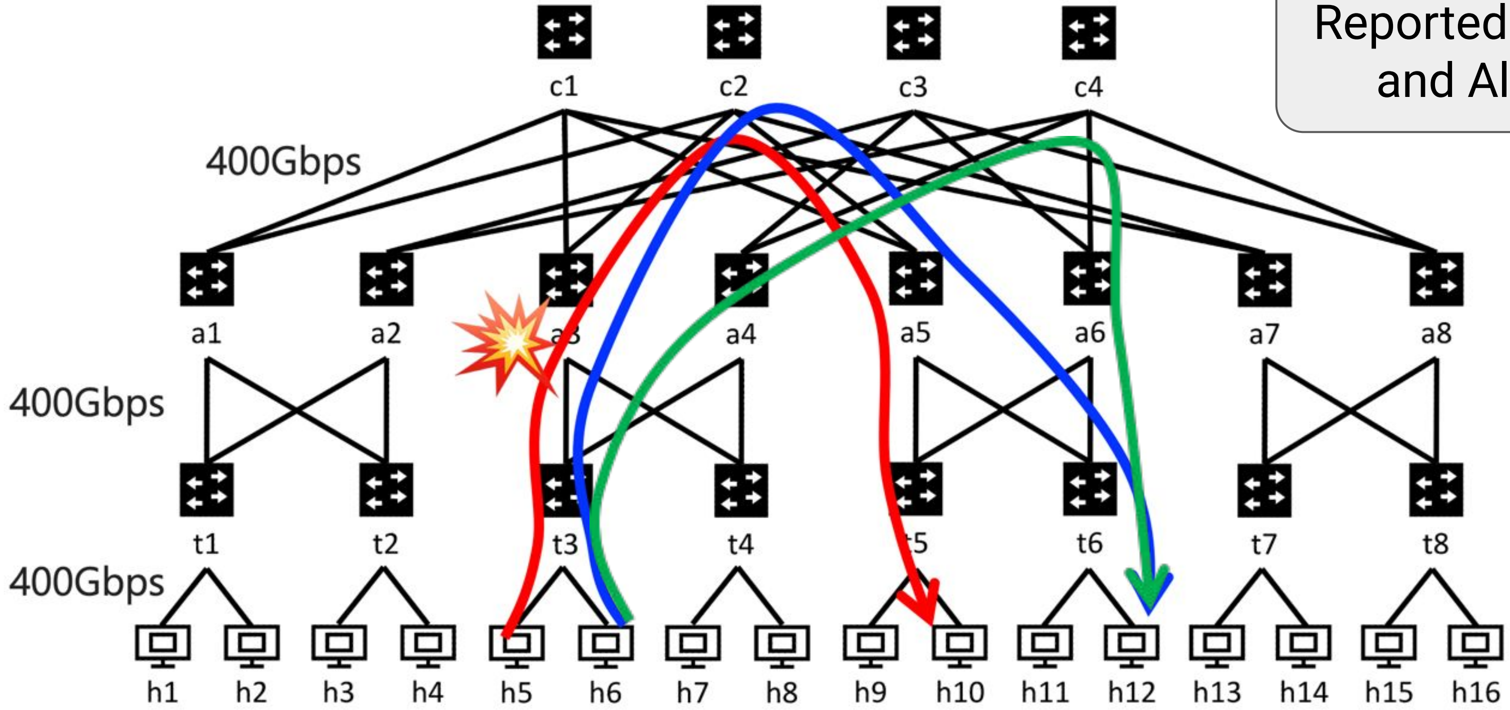
DeepSeek-V3 MoE: all-to-all like

~2015 ~2020 2024 2025 Time

Slowly Evolving Networking

Host transport on RDMA NICs is hard to adapt to better suit ML workloads

- Hardware changes are time-consuming



Reported by Meta and Alibaba

Agenda

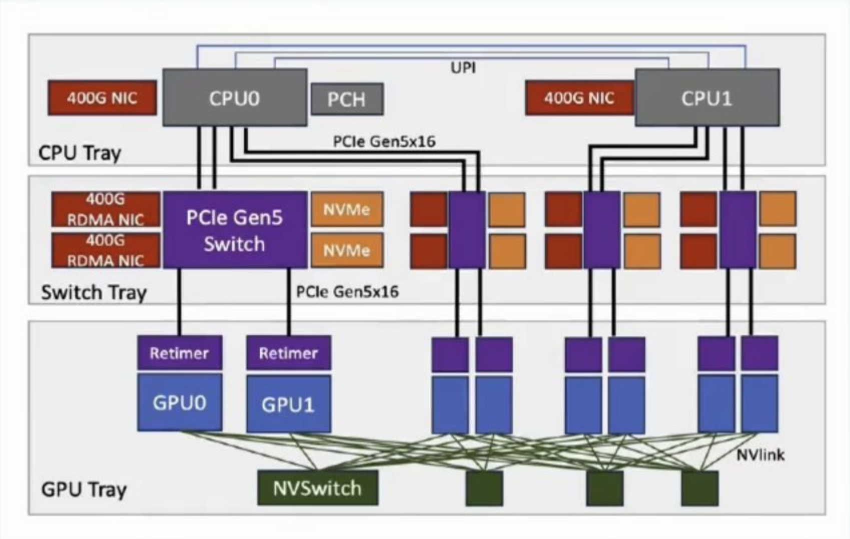
- Overview of Our Solution
- Hardware and Network Topology
- Routing Evolution
- Transport
- Operations and Experiences
- Conclusion and Future Considerations

Meta AI Training Evolution

DISTRIBUTED TRAINING TO GPU FULL SYNC TRAINING	**MODEL COMPLEXITY AND SCALE EXPLOSION**	**VARIED MODEL AND DATA PARALLELISMS**	**HIGH BANDWIDTH & LOW PREDICTABLE LATENCY**

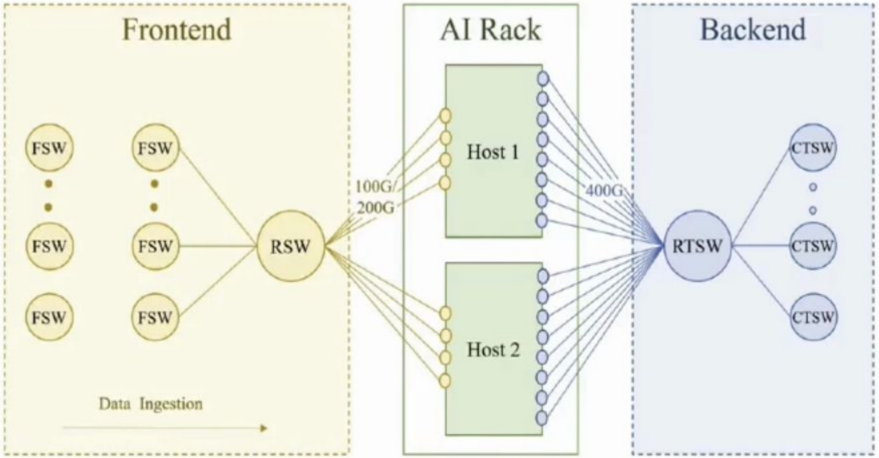
Our Production RDMA network...

PURPOSE BUILT FOR AI WORKLOADS	ROCEV2 TRANSPORT WITH PFC	LARGE SCALE (O(100K)), MULTI-VENDOR DEPLOYMENT	O(10K) SIZE CLUSTERS



Dedicated Backend NIC per GPU

- Grand Teton Open Compute Chassis
- Single Port 400G Backend NICs
- 1 NIC to 1 GPU mapping
- PCIe speed matches between GPU/ NIC



Separate RDMA Backend and Frontend Network

Helps us account for varied growth rates for GPU to GPU synchronization traffic vs Data Ingestion and supporting traffic.

Network Topology

- Dedicated Backend Network isolated from Frontend Network
- 3 Layer Clos
- ToR switches: Shallow Buffers
- Spine Switches: Deep Buffer switches with static carving buffers

Routing Solution

- Originally Static Routing
- Evolved to Enhanced ECMP hashing on Destination QP ID + Collectives Implementing flow multiplexing
- Traffic Engineering on First hop switches
- Future exploration: Flowlet load balancing

Congestion Control

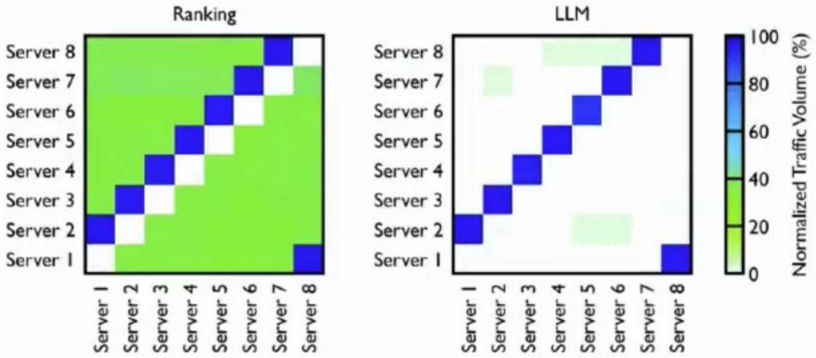
- DCQCN along with PFC 200G Networks
- Pivot away from DCQCN for 400G
- Use Collective library to limit Congestion

Operations and Perf Tuning

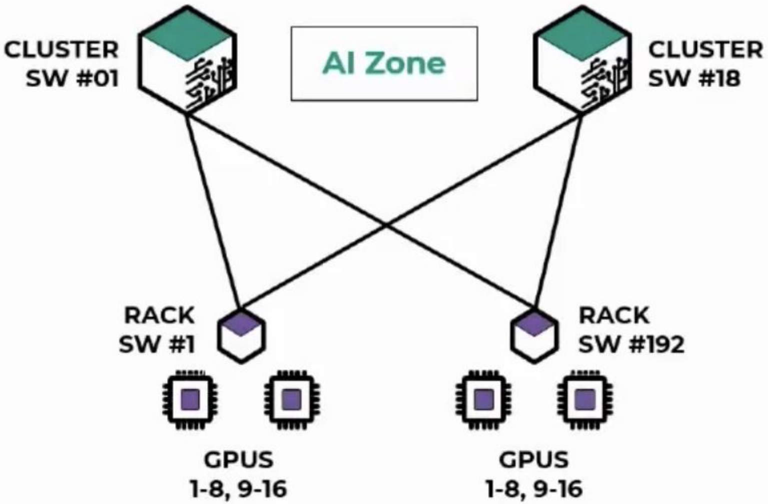
- Network and Communication library (NCCL) tuning
- Performance consistency
- Operational learnings to Scale

Learning #1: Scale and traffic patterns in AI Training

	SCALE OF END-POINTS PER JOB	MODEL / DATA PARALLELISM (SCALE-OUT)	PREDOMINANT TRAFFIC PATTERNS
RANKING FOR RECOMMENDATION	UP TO HUNDREDS	EMBEDDING EXCHANGE & DATA PARALLELISM	FULL MESH + HIERARCHICAL PATTERNS
CONTENT UNDERSTANDING & GENERATIVE AI	UP TO TENS OF THOUSANDS	FULLY SHARDED / CLASSICAL DATA-PARALLELISM PIPELINE PARALLEL	HIERARCHICAL PATTERNS



AI Zone: Built for Ranking Workloads



Rack

- 16 GPUs per rack
- Connected by RTSW Shallow Buffer Switches

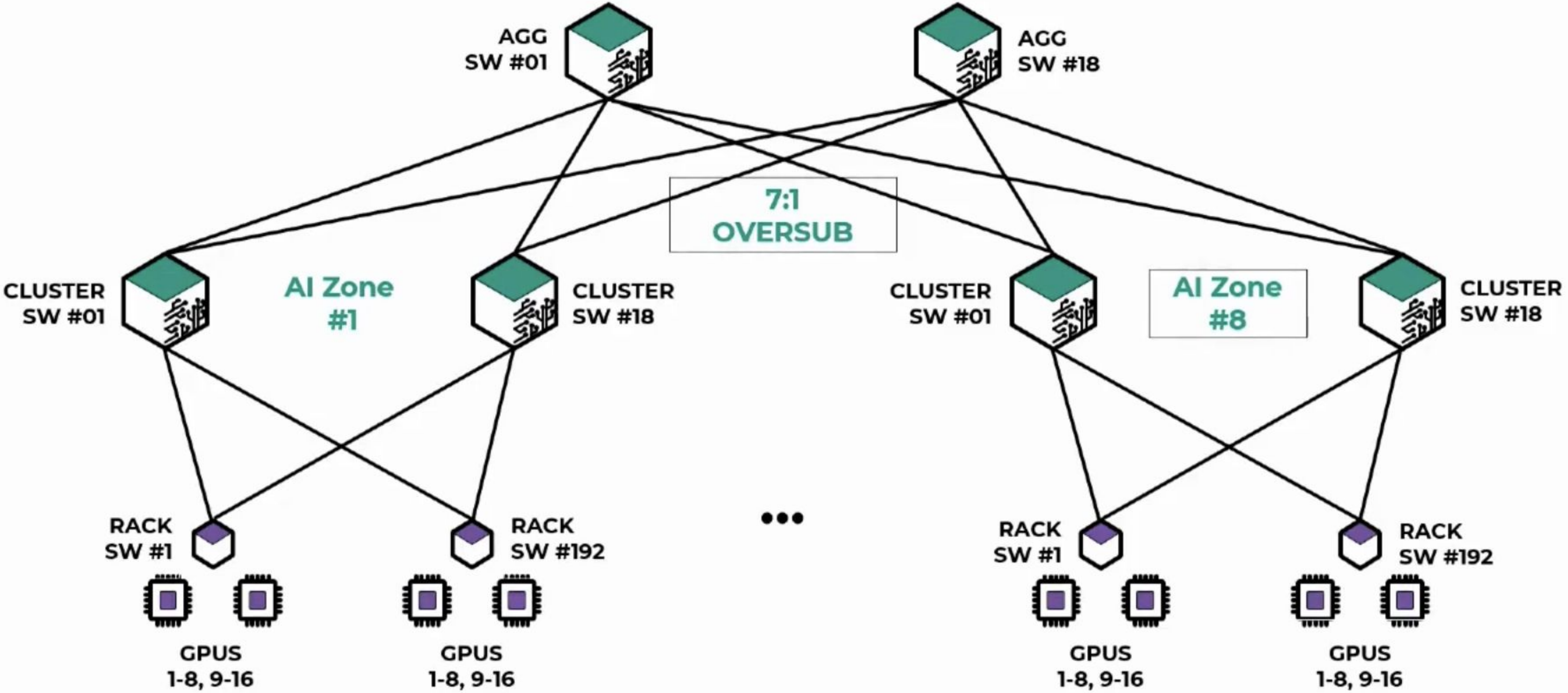
AI Zone

- Upto 18 **CTSWs** connect 256 racks forming ~4000 GPUs
- CTSW: Deep buffer switches with Virtual output queuing architecture

Full Bisection Bandwidth

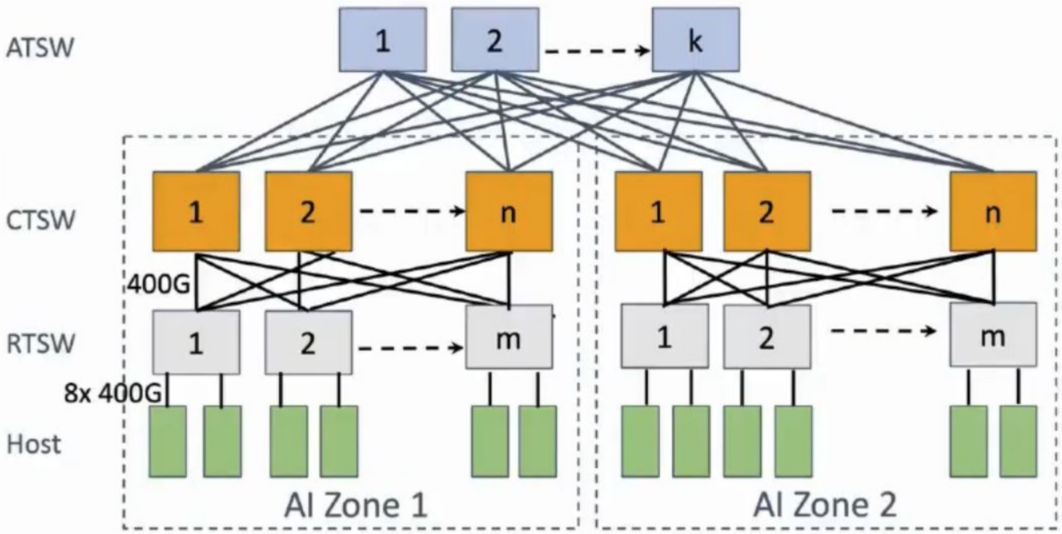
- Needed for Ranking workloads with Full Mesh Network patterns

DC Scale Cluster for GenAI



ToR based 3 stage Clos Topology for Extensibility and Reliability

- **ToR Architecture**
  - RTSW: Shallow buffer Switch running FBOSS
  - 2 Servers per rack limits the switch failure domain
  - Facilitates Usage of DACs for NIC to RTSW connector
- **Spine Switches:**
  - Deep buffer switches with Virtual output queuing architecture
- **AI Zone**
  - Upto 18 **CTSWs** connect 256 racks forming ~4000 GPUs
  - Full bisection bandwidth
- **DC Scale**
  - **ATSW** switches connect up to 8 AI Zones
  - Provides Oversubscribed bandwidth
  - Large jobs with collectives suitable are placed
  - Scheduler spreads job with Network topo awareness to support collective algorithm graph



Learning #2: Low Flow Entropy with Hierarchical Collectives

Collectives	Avg. # of QPs per GPU
`AlltoAll(v)`	15
`AllReduce`	4
`AllGather`	4
`ReduceScatter`	4

Number of Flows / active QPs per NIC
128 GPU Collective

TE Rollout Signal from Production: Ranking Cluster, AI Zone

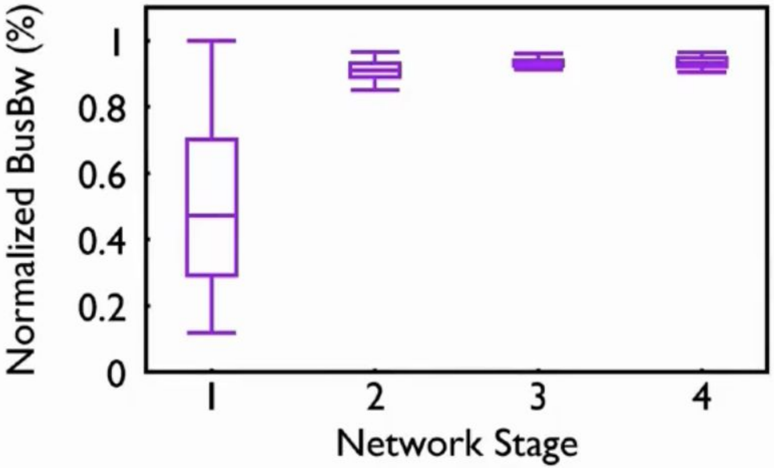
Bandwidth observed by All Reduce Kernel extracted based on Trace Duration observed over the years.

Stage1: Static Routing

Stage2: Under-Subscription 1:2

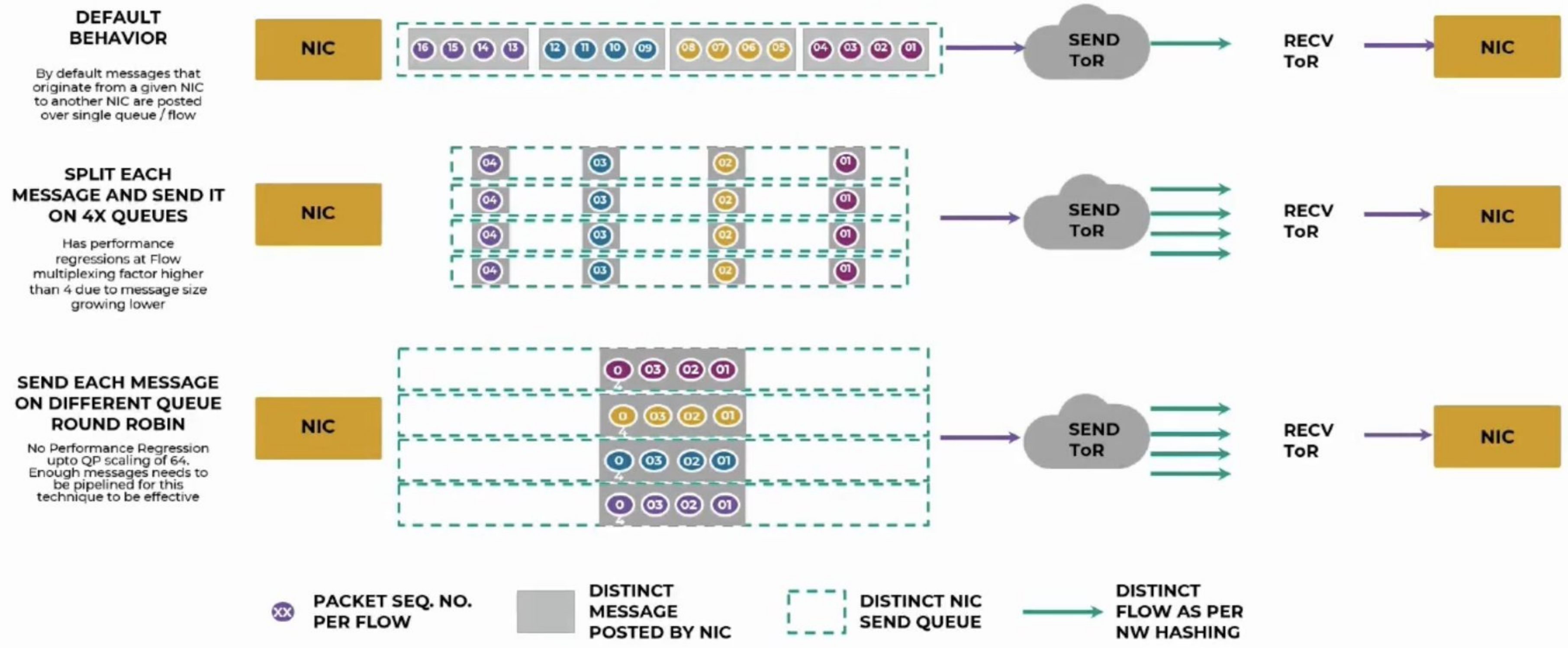
Stage3: TE Rollout

Stage4: Reduce under-subscription to 1 : 1.125

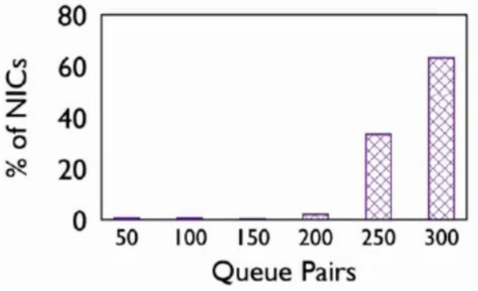


Improving ECMP with Flow Multiplexing

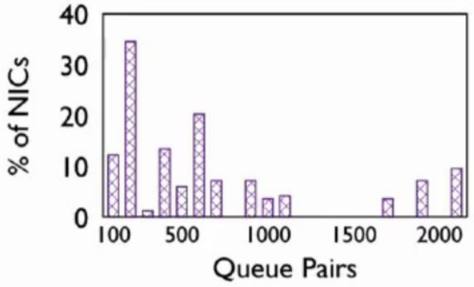
Example of Flow multiplexing of Factor 4 on Elephant flow worth 4 messages 4 pkts each



GENERATIVE AI



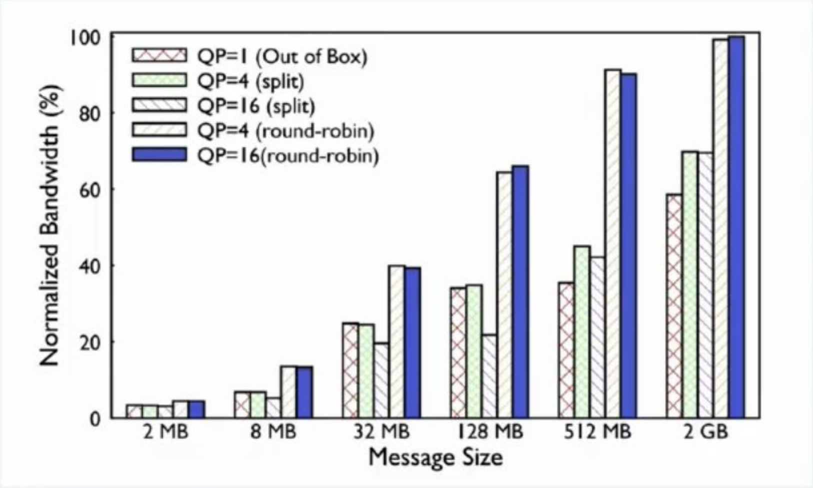
RANKING



QP Scaling Impact

Applying QP Scaling = 4 for Ranking workloads

Applying QP Scaling = 16 for GenAI workloads



PERFORMANCE IMPACT

ECMP hashing with QP ID: Perf

QP=16 helped achieve roofline performance for hierarchal collectives

Learning #3: Variable Congestion Per Collective

Collectives	Buffer occupancy per leaf switch (MB)
AlltoAll(v)	65.6
AllReduce	13
AllGather	22.1
ReduceScatter	19.6

Cumulative Buffer Watermarks per RTSW
128 GPU Collective

Tuning DCQCN did not provide a net benefit

- Momentary congestion.
  - Original Approach: With DCQCN with 200G Networks
  - 2 not net-positive outcomes:
    - Low buffer utilization with perf regression in corner cases
    - Marginal perf benefits (if any) with higher buffer thresholds.

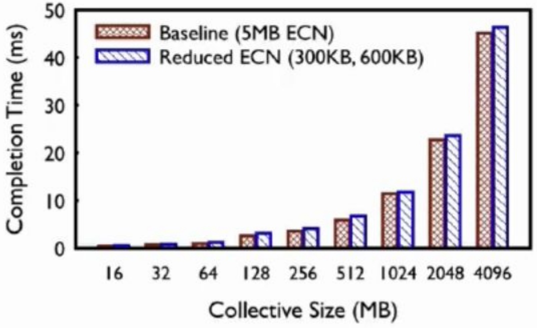


Figure 12: ECN impact on performance: Allreduce completion time(ms) on 32 GPUs comparison with CTSW ECN threshold changes. Baseline uses 5MB as both low and high thresholds. A tighter threshold of 300KB low and 600KB high leads to lower performance.

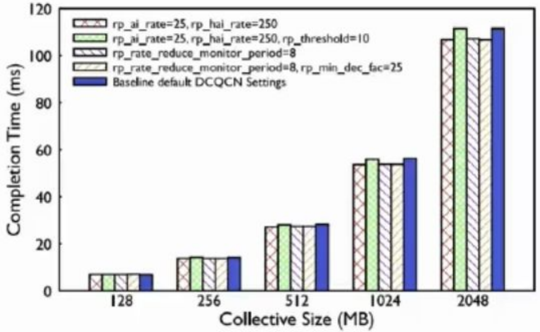
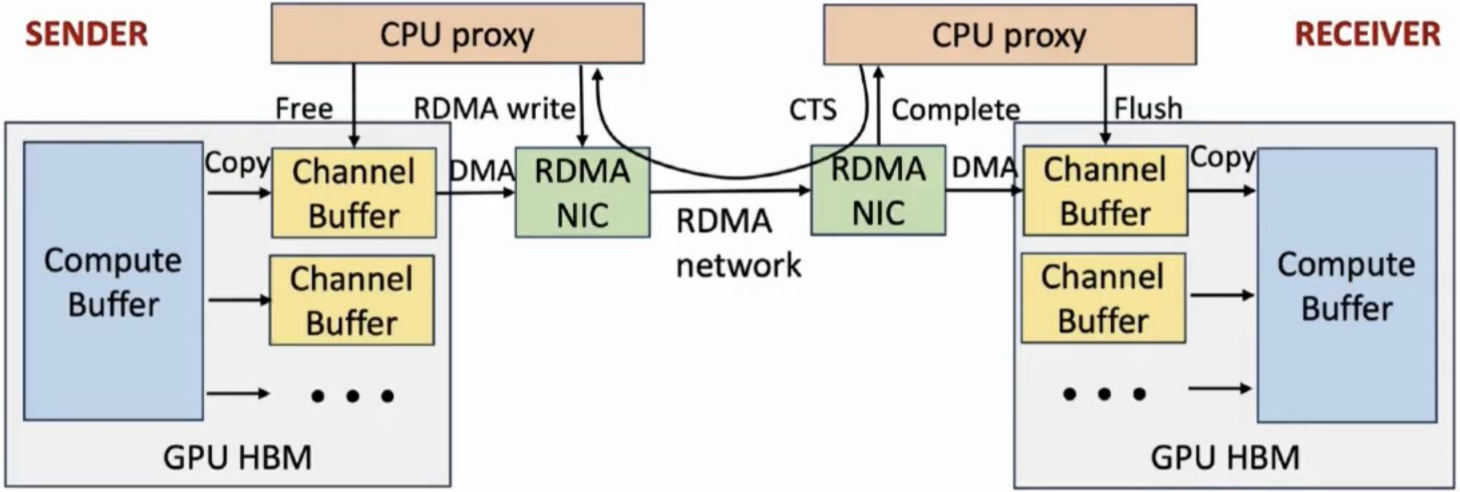


Figure 13: Tuning DCQCN for `AlltoAll` collective

Receiver based Congestion Control with Comms Library



ToR topology with 3 Layer Clos helps support large scale with extensibility and reliability

Enhanced ECMP and TE deal with persistent congestion caused by inefficient load balancing

Collective Library and Static carving of buffers to deal with momentary congestion

Roofline performance can be achieved by Comms Library and Network Co-tuning

Buffers:
Can we scale AI Training with shallow buffers switches without sacrificing on performance and reliability ?

Meta's current deployments are with deep buffer switches scaled deployments operational ease without sacrificing performance.

Routing evolution:
Can we come-up with a generalized load balancing solution that is operationally simple ?

If so will this approach with Network / Switch Centric or End-point custom transport centric or will need both ?

High RTT:
Can large clusters to scale gen AI models involving large network latencies support efficient training ?

Is a lossless approach still feasible at this cable lengths ?

Fungibility:
Can we build common networks to support Ranking, GenAI and Dist. Inference Use-cases ?

These apps and technologies have shared infrastructure for years and the teams behind them frequently work together.

THANK YOU

Meta