Insights from Production: Scheduled Ethernet Fabric in Large AI Training Clusters

OCP GLOBAL SUMMIT | OCT 15-17, 2024
SAN JOSE, CA

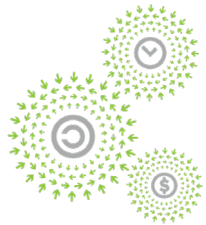
Networking

NETWORKING

Insights from Production: Scheduled Ethernet Fabric in Large AI Training Clusters

Pengfei Huo (ByteDance, Sr. Network Architect)

Henry-Xiguang Wu (Broadcom, Switch Product Marketing)



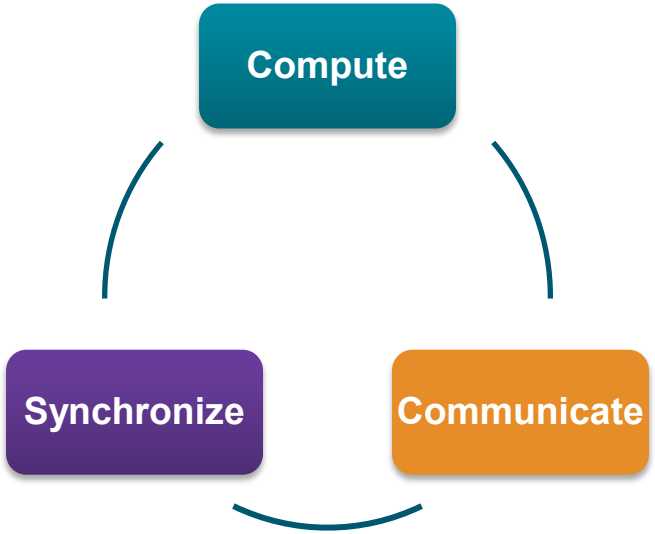
OPEN PLATINUM™

Agenda

- Challenges for AI training network
- Introduction of Scheduled Ethernet Fabric
- ByteDance’s Production Insights
- Standardization and Ecosystem
- Call to Actions

What Makes AI Networks Unique?

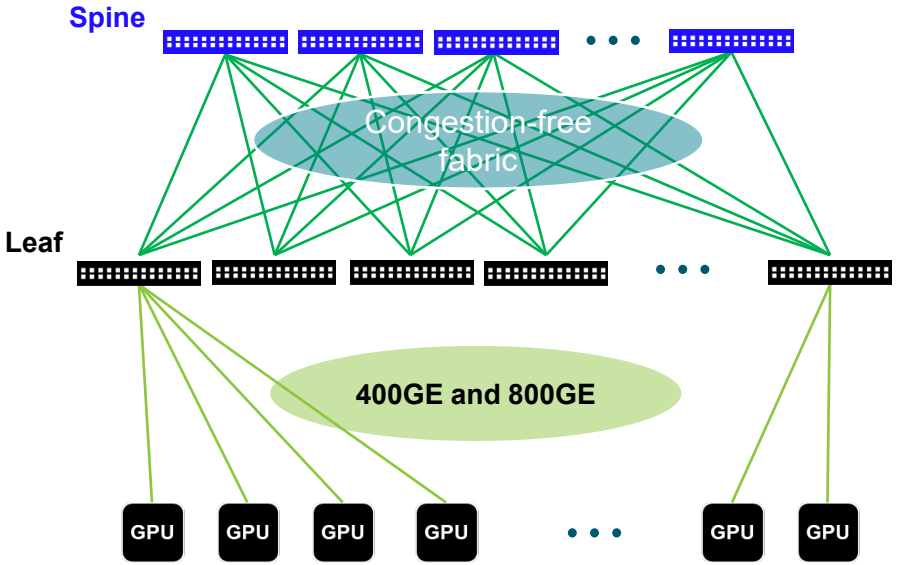
- Fewer flows (low entropy)
- High bandwidth flows
- Synchronized and bursty traffic
- Links are saturated in micro-seconds (<< RTT)
- Training jobs run for long periods of time (hours, days)
- Tail latency impacts job completion time significantly



GPU to GPU Communication Drives High Bandwidth Utilization

Scheduled Ethernet Fabric – Receiver based Scheduling

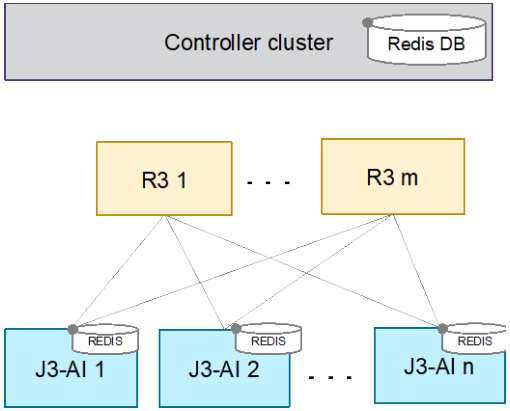
- Switch scheduled fabric
  - Standard Ethernet I/O
  - Receiver based scheduling
  - Leaf: switching, forwarding, queuing, scheduling
  - Spine: forwarding at low power
  - OCP compliant system
- Leaf deployment options
  - ToR/MoR - in the XPU racks
  - In the network rack, with spines



* 800G Distributed Disaggregated Chassis Routing System Evolution, https://www.opencompute.org/documents/ddc-v3-ocp-base-specification-revision-5-0-pdf

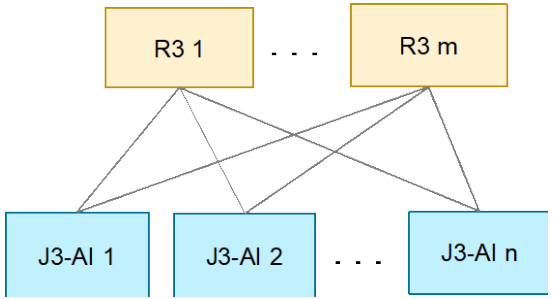
NOS Options

Centralized State Synchronization



- Every J3-AI sends local ARP and port information to the Redis DB on the Controller which syncs it to all other J3-AI. Full mesh BGP distributes routes.

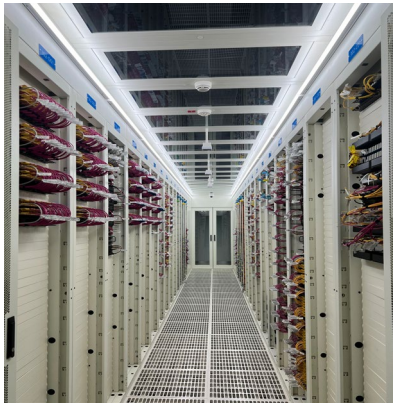
Distributed State Synchronization



- Full mesh connectivity between J3-AIs for state synchronization to distribute ARP and port information. Full mesh BGP distributes routes.

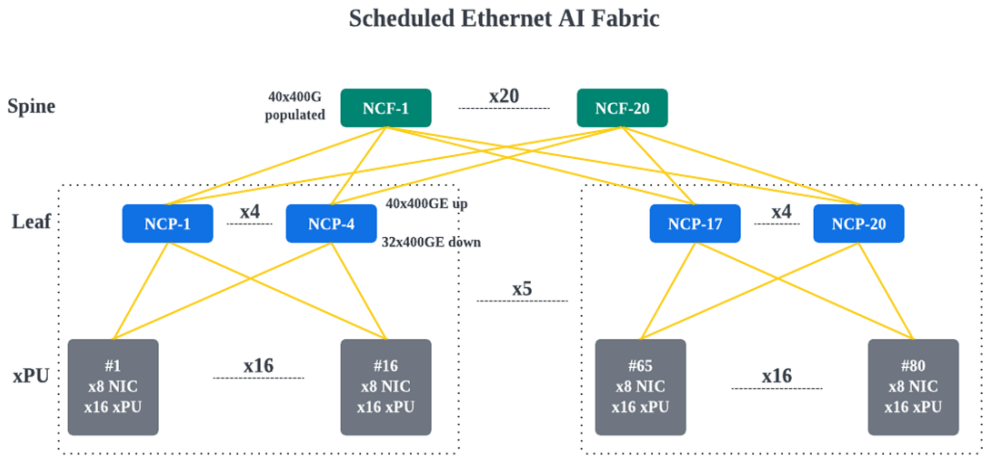
The Journey to Scheduled Fabric

- 2022 Testing with Jericho2&Ramon Chassis
  - HBM is not mandatory for AI, On-chip buffer works well
- 2023 POC Jericho2C+/Ramon Scheduled Fabric with 128 x 200G XPU*
  - Excellent performance at AI traffic
  - Comprehensive tests covering internal CC/credit mechanism/impact of cable lengths/compatibility/ dynamic PFC/gRPC/etc …
  - Identified areas to optimize for AI application from NOS perspective
- 2024 June - First Scheduled Fabric with J2C+/Ramon in Production
  - ByteDance drove the implementation of distributed control plane
- 2024 Aug - Jericho3/Ramon3 POC completed
- More excitement to go



* More info at https://www.youtube.com/watch?v=zC8OQVHtlIQ

Architecture of The 1st Scheduled Fabric in Production



- 1280 xPUs cards at a speed of 400GE
- Distributed management system
- Two-layer Architecture
- 4-Rail connection
  - Single-rail performs on par with multi-rail in SF, single low digit % performance premium at multi-rail
  - Users are free to choose multi or single rail
- N+2 redundancy at Spine
  - Performance degradation is linear at failure
- No need for DC-QCN
  - No struggling on parameter tuning
  - PFC in between NCP and NIC/endpoint
- QP number/entropy does not impact load balancing

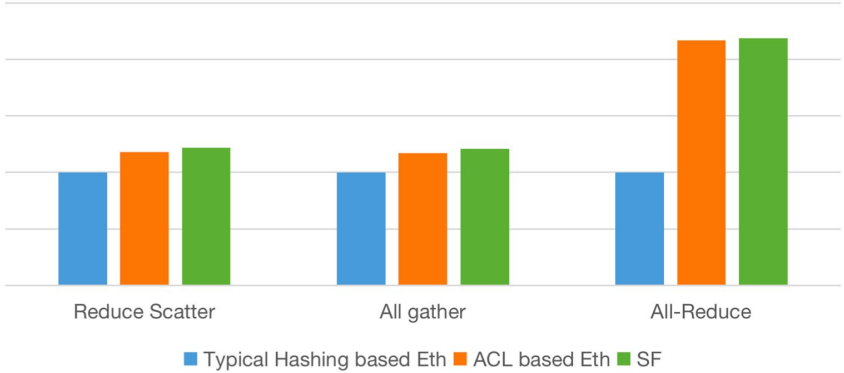
Performance in the Production Cluster

- Rather than static PFC, dynamic PFC enables more controls to user and better performance

Static Threshold ✗

Dynamic Adjusted Threshold ✓

- Collective communication performance based on 128 XPU cards



Reduce Scatter - 22% improved

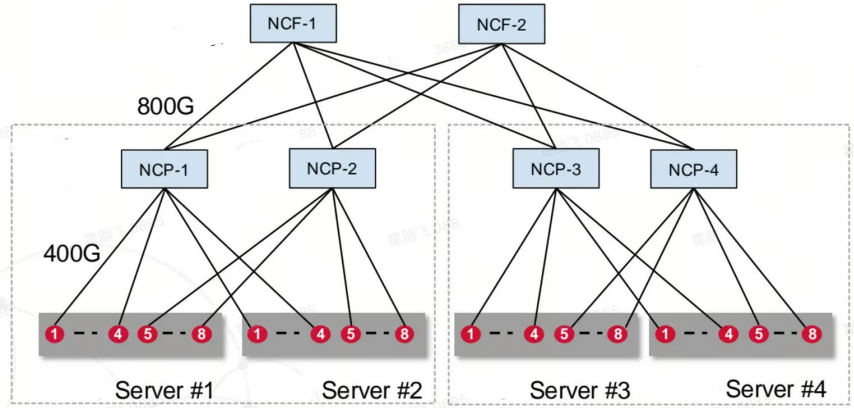
All Gather - 21% improved

All Reduce - 119% improved

* ACL based hashing is hard to manage and scale in real application

Performance in Jericho3 POC

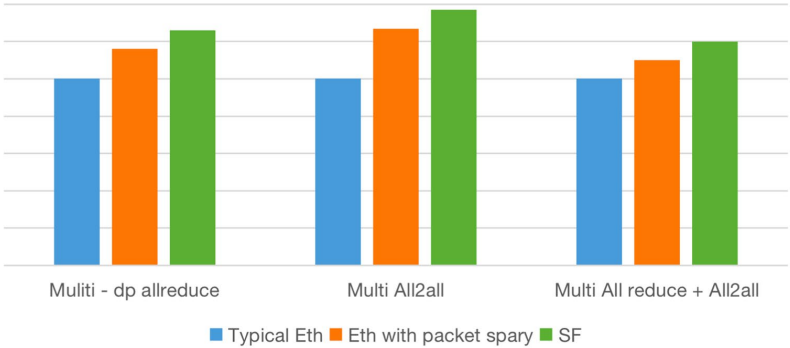
- XPU Servers
  - 8 XPU cards per Server
  - 400G NIC per XPU card
- Network
  - Jericho3 NCP: 3.2T downlink
  - Ramon3 NCF
  - 2 Rails, very four NICs from each server belongs to one rail



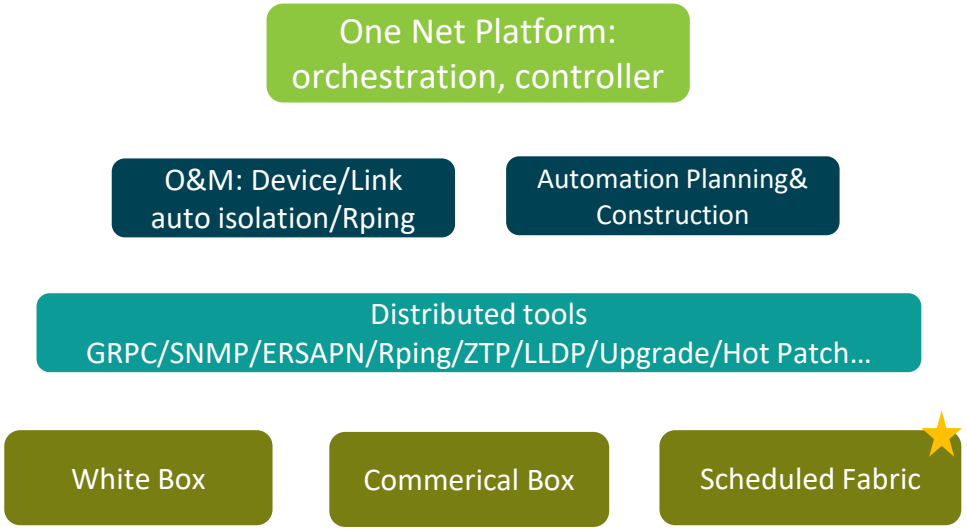
Multi- DataParallel allReduce - 26% improved

Multi- All2all - 37% improved

Multi- All2all&allreduce - 20% improved



Seamless Operation & Maintenance



System Layer - One Platform
- Unified orchestration, O&M&C

Solution Layer
- Unified O&M and Construction Solution/Process

Tools Layer - Unified standard for each function
- Independently upgrade GRPC/Sflow/ERSPAN/ZTP/link-verfiy

Device Layer - Different Switches

Unified Construct & Operate across Scheduled and non-Scheduled Fabric systems

Scheduled Fabric is managed as individual boxes

Standardization Initiative

Proprietary control planes from vendors create challenges in construction, operation, and supply

ByteDance response – Disaggregation and Standardization

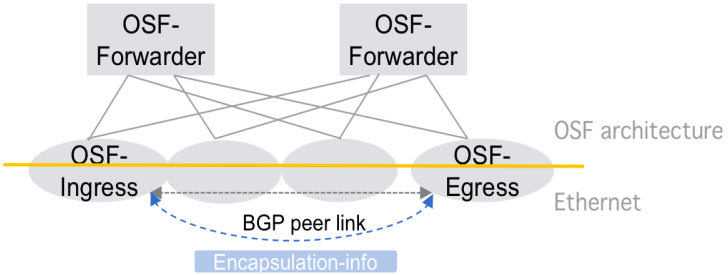
DONE

Step-1: Control Plane from Centralized to Distributed mode

WIP

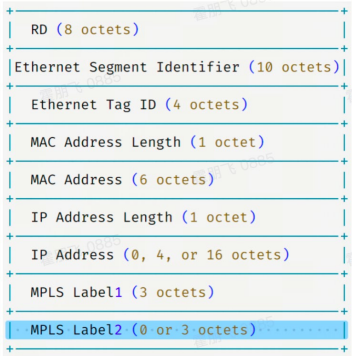
Step-2: Control plane Interoperability

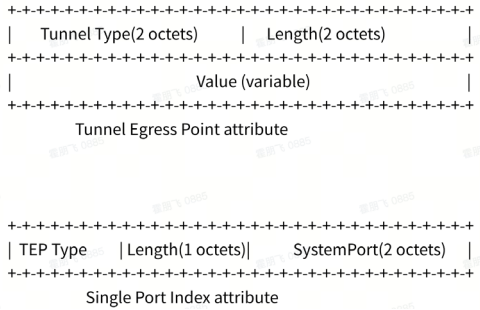
- OSF (Open Scheduled Fabric) Framework

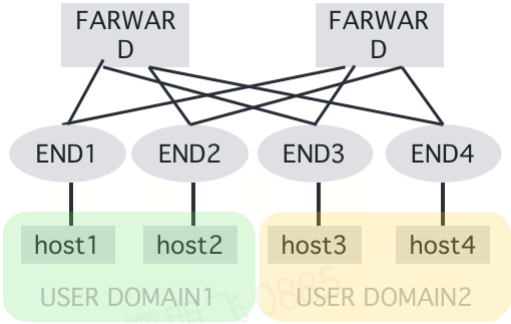


Open Scheduled Fabric Framework

- Easy to develop - solely leverage EVPN with extended Community value
- One protocol covers all scenarios - L2, L3, Multi-tenant







Call to Action

ByteDance H3C DRIVENETS Ruijie Networks Accton BROADCOM

- Welcome to OSF ecosystem
  - A receiver based option besides UEC, Standard Ethernet with DCQCN or DLB/GLB and other Ethernet solutions
  - Software is planned to be contributed to SAI/SONiC
- MSA is ongoing
  - aiming to standardize the hardware resource allocation, e.g. MOD port range,
- Please contribute to
  - 800G Distributed Disaggregated Chassis Routing System Evolution (V3), https://www.opencompute.org/documents/ddc-v3-ocp-base-specification-revision-5-0-pdf
  - A OSF Framework for Artificial Intelligence (AI) Network, https://datatracker.ietf.org/doc/html/draft-hcl-rtgwg-osf-framework-00
  - BGP Extension for Tunnel Egress Point, https://datatracker.ietf.org/doc/html/draft-hcl-idr-extend-tunnel-egress-point-00
  - Distributed Forwarding in a Virtual Output Queue (VOQ) Architecture, https://github.com/sonic-net/SONiC/blob/master/doc/voq/architecture.md

Thank you!