Google

# At-scale AI DCN Design Considerations

JK Lee (Google)
leejk@google.com

*IETF 121, AIDC*
*November 7th, 2024*

# This talk is more about **physical design and deployment**

How can we:

1. deliver the **performance (both bandwidth and latency) at scale**?
2. **deploy** this network design in the **physical world**, with realistic constraints?
3. **speed up** the deployment at scale?

at the same time.

*Sharing design goals/requirements, rather than specific solutions.*

(Focus is on AI datacenter networks, but wider-area interconnects have similar issues)

# #1: Performance at scale

**Bandwidth at scale** → multiplicative increase

- Per-accelerator BW demand
- Per-rack accelerator density increases (thanks to liquid cooling)
- Scaling law of LLM persists (100ks of accelerators)

**Latency at scale**

- Collective performance is sensitive to latency and flow collisions
- Need as many accelerators as possible under small hop distances and fiber reaches

# #2: Implication on physical design & deployments

**Multiplicative BW**-per-machine-rack demand vs. linear SERDES speed increase

- # of network racks to deploy increases faster than # machine racks
- # of fibers to deploy and (manually) connect increases fast

**Latency at scale** → max fanout with bounded latency

- Dense logical topology
- Dense and optimal physical placement of racks and fibers
- At both a datacenter level and a campus level
- While honoring physical constraints like secure conduit, fiber path diversity, power and cooling infrastructure

# #3: How to speed up deployment at scale?

- Common rack design
- Fiber design to minimize the manual porting work
- Optimize the space-time workflow of technicians and equipment
- Incremental network deployment aligned with DC infrastructure build steps and machine delivery
- Retain design and deployment flexibility around demand/reqt changes vs supply chain issues
- SW testing: *Digital Twins* of the physical-world dimensions and constraints

More details in “*Physical Deployability Matters,*” J. Mogul and J. Wilkes, HotNets 2023

# Resiliency

- **Main goal: avoid 1) forced checkpoint rollback or 2) job slowdown** wrt any kind of network failures
- Proactive link monitoring & link qual at scale
- Need clear signal for link flapping, link down, link up
- Fail-static fabric + host-driven reaction helps
    - Deridex @ SigComm'23
- Align failure domains; align maintenance windows
- Prioritize repair SLO. Fail fast and repair fast
- E2e coordination between network monitoring/mgmt system, application scheduler and job scheduler

# Telemetry

**There is no one-size-fits-all solution for**

- Telemetry for congestion control and reliable transport (e.g., CSIG @ IETF)
- Telemetry for performance monitoring and optimization
- Telemetry for triaging and troubleshooting

First two calls for host- or application-driven top-down telemetry solutions

**Bottom-up fabric telemetry is critical for speedy triaging** in at-scale AI DC

- What, where, when and why. (Who is less critical)
- Lightweight solution that works on different platforms

# Summary

## At-scale deployment and operation is critical

- Fabric design for speedy deployment and easy operation
- Operation policy to handle application/host/fabric SW failures & upgrades
- SW to automate deployment and operation

## Collaboration opportunities

- Forums to share pain points and ideas (like AIDC)
- Telemetry: standard metric definition, streaming interface (e.g., gNPSI)