Handling inter-DC/Edge AI-related network traffic

(IDEA traffic handling)

IETF 121 side meeting, November 7th, 2024

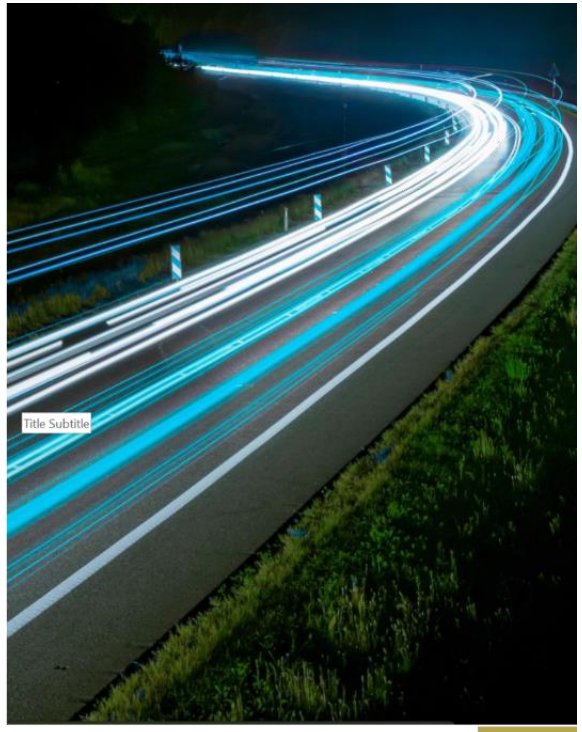
AI-related network traffic growth

Increase in bandwidth demand

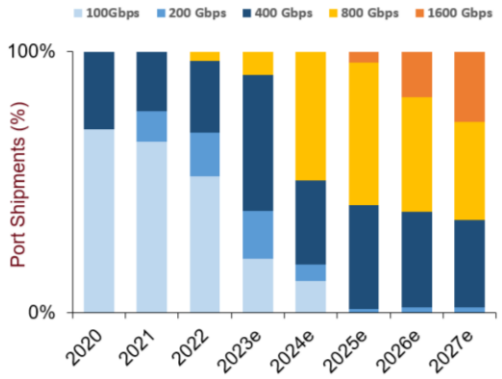
- Demand associated with front end and back end traffic
  - Back end ⇔ model training
  - Front end ⇔ model transfer, inference

AI Traffic Characteristics - Continued

- **Average cluster size is growing:**
  - AI models growing 1000X every 3 years
  - Cluster size quadrupling every 2 years
- **Amount of network bandwidth per accelerator:**
  - Growing from 200/400/800 Gbps today to more than 1Tbps in the near future
- **AI traffic growth rate:**
  - Up to 10X every 2 years @ some Cloud SP networks



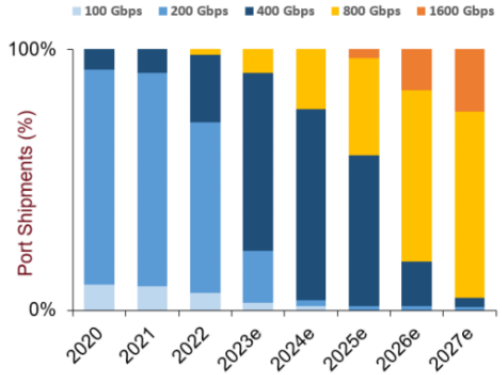
Migration to High-Speed in AI Networks (Front-End)



Preliminary Forecast (2023-2027)

- All ports are Ethernet
- Nearly 2/3 of the ports will be at 800 Gbps speeds and above by 2027

Migration to High-Speed in AI Networks (Back-End)



* Includes both Ethernet and InfiniBand
* Source: Dell'Oro Group AI Networks Report

Preliminary Forecast (2023-2027)

- InfiniBand and Ethernet will coexist
- Nearly all ports will be at 800 Gbps speeds and above by 2027
- Triple-Digit CAGR for network bandwidth

AI-related network traffic growth

Need for distribution pattern (adapted from Dell' Oro)

AI XPU size	Server I/O 10-100s of XPUs	Rack scale 1000s of XPUs	DC Scale 10K+ of XPUs
Type of applications	Small AI apps	Moderate AI apps	Large AI apps
AI Network options	CXL – NVLink PCIe	AI leaf Ethernet or IB	AI Spine Ethernet or IB

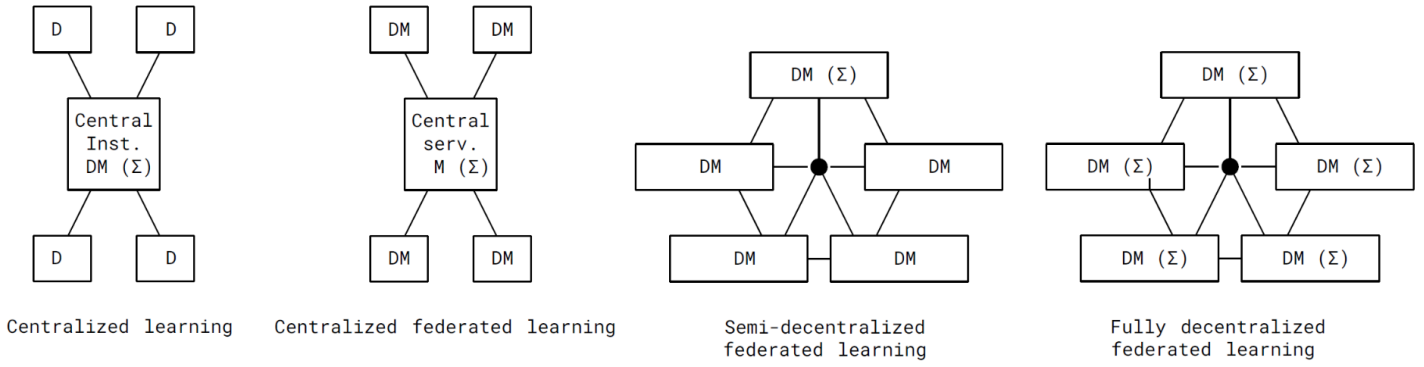
Legacy AI workloads

Modern / LLM / generative AI workloads ➔
Distribution at datacenter scale,
Network specialization to cope with specific AI requirements

AI-related network traffic growth

Need for distribution pattern (adapted from Dell' Oro)

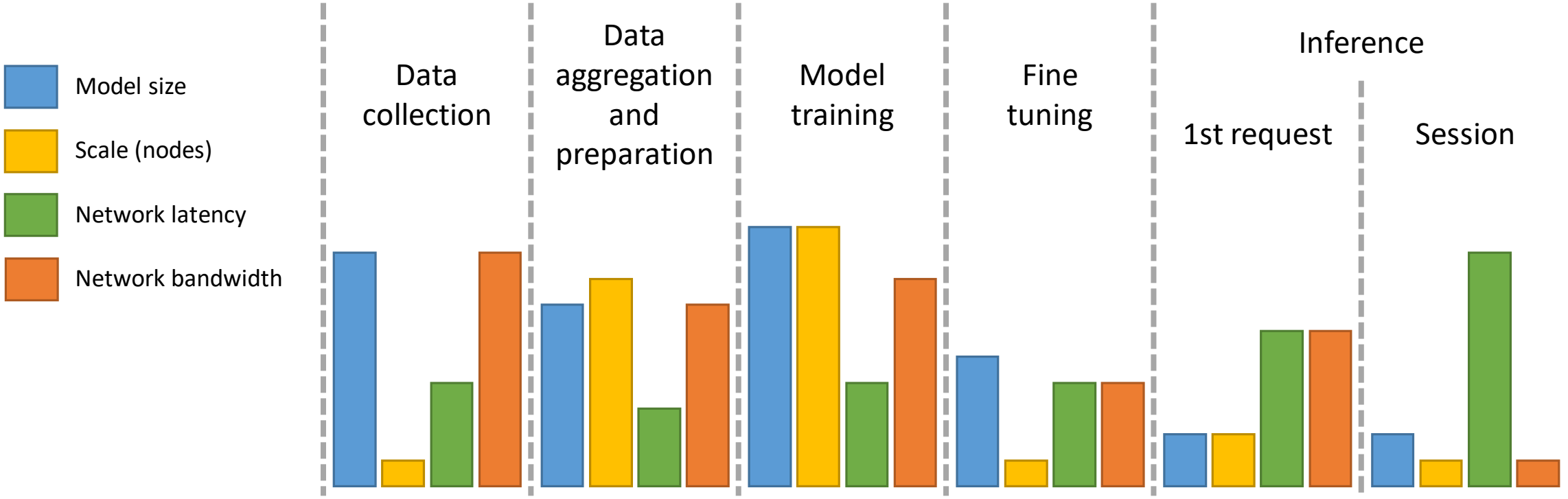
AI XPU size	Server I/O 10-100s of XPUs	Rack scale 1000s of XPUs	DC Scale 10K+ of XPUs	Inter-DC/Edge scale 100K+ of XPUs
Type of applications	Small AI apps	Moderate AI apps	Large AI apps	Larger scale LLM / GenAI apps
AI Network options	CXL – NVLink PCIe	AI leaf Ethernet or IB	AI Spine Ethernet or IB	*To be defined*



From centralized AI distributed accross datacenters to decentralized AI

Networking for AI initiatives and AI lifecycle

Traffic characteristics (Source: Meta, Juniper)



Training is centralized in large clusters

- Training is distributed among GPUs
- High network bandwidth
- JCT latency is important, but network latency is not critically sensitive

Inference is small clusters or edge/distributed

- Inference is mostly on 1 CPU/GPU, except in LLMs where it may be a few
- Inference overall result latency is important, but inference is in a single GPU or single server (8 GPUs), not crossing a large network fabric

Work on AI-related network challenges in the IETF

- At last IETF meeting, we organized a side meeting on *Inter-DC AI: Requirements and Challenges*
  - Participants: ~60 people
  - Follow-up discussions with participants showing interest in this topic from equipment vendors, operators and other involved parties
- General interest for AI-related network challenges in general in the community
  - Several side meetings during IETF 121 week:
    - Use of AI to manage networks: AI4Net, Large Language Models for Networking, AI control
    - Networks supporting AI workloads: 6gip AI/MLNet, Net4AI, our initiative
  - HP-WAN BoF on Monday with AI-related use case
- Draft introducing Network challenges related to AI traffic outside the datacenter
  - https://datatracker.ietf.org/doc/draft-aft-ai-traffic/ entitled *Handling inter-DC/Edge AI-related network traffic: Problem statement*

Agenda for today

13:30 – 13:35 (5 min)	Meeting setup and introduction of the topic *(Huawei)*
13:35 – 13:50 (15 min)	*Accommodating LLM Service over Heterogeneous Computational Resources* by Binhang Yuan *(Together AI – HKUST)*
13:50 – 14:00 (10 min)	*RDMA proxy Inter-DC Over WAN Using Gateway* by Rubing Liu (*H3C*)
14:00 – 14:15 (15 min)	*Considerations on Inter-DC Network Requirements* by Yisong Liu *(China Mobile)*
14:15 – 14:30 (15 min)	*Research Progress of Intelligent Computing Networks in China* by Liang Guo *(Chief Engineer of Cloud Computing and Big Data Research Institute, CAICT)*
14:30 – 14:40 (10 min)	*Enabling Inter-DC AI-Networking with Service provider optical slicing and programmable pluggables* by Oscar Gonzalez de Dios *(Telefónica)*
14:40 – 15:00 (20 min)	Discussion on challenges to address in IETF and conclusion