FANTEL Scenarios in IP MAN

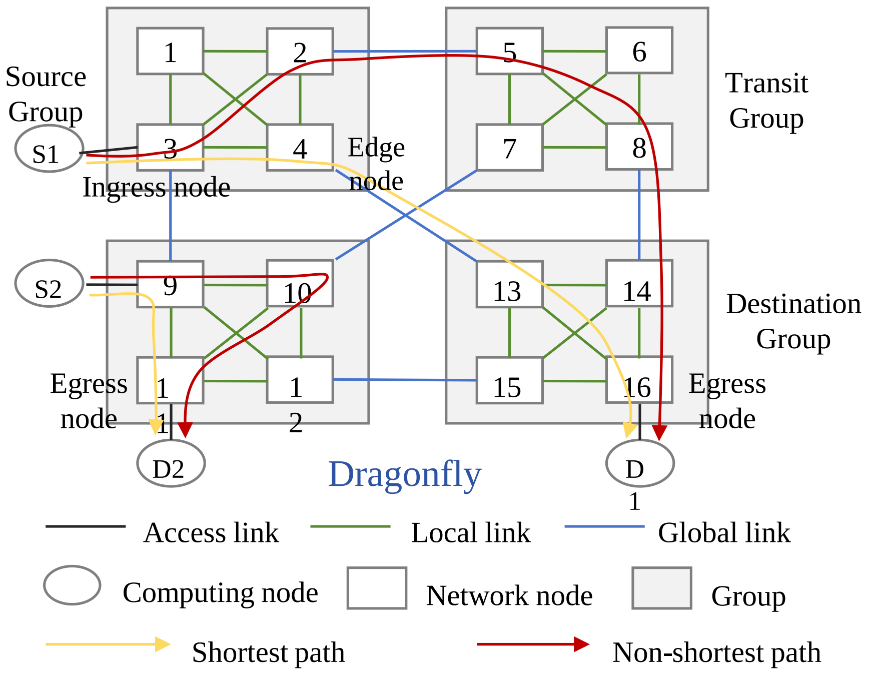
Yongqing Zhu

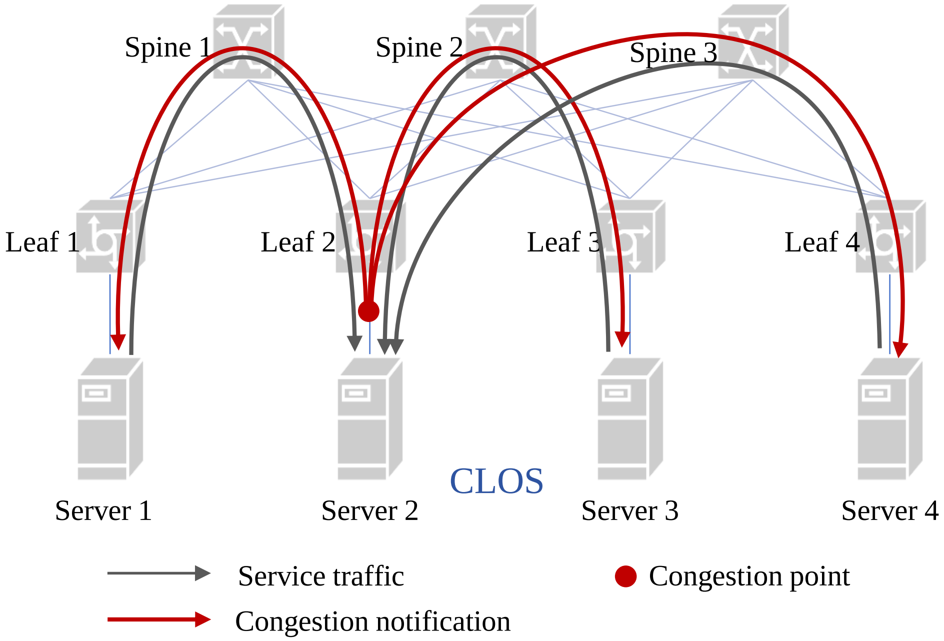
zhuyq8@chinatelecom.cn

IETF122 Sidemeeting

Fast and precise notificaton in AI Data Center (AIDC)

- **With the booming of AI, China Telecom is actively building related infrastructures, such as AIDC, computing power scheduling platform et al., to meet service requirement**





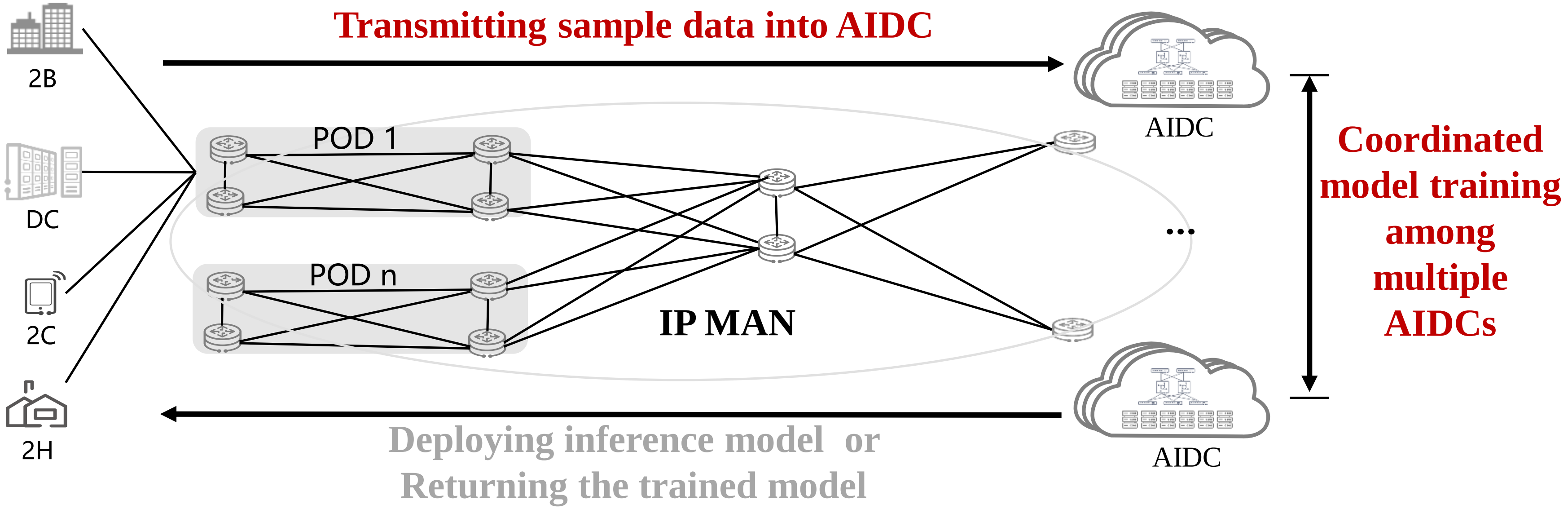
- Dynamically adjusting traffic forwarding path based on ***fast and precise notifications*** for network status changes

- Dynamically adjusting packet forwarding rate based on ***fast and precise notifications*** for congestion condition

AI service requirements for IP MAN

Main scenarios related to AI services in IP Metroplitan Area Network

- Transmiting sample data into AIDC: Uploading the sample data from the customer to the AIDC for model training
- Coordinated model training among multiple AIDCs (in case one AIDC can’t meet the requirement of model training)
- Deploying the inference model in IP MAN or returning the trained model to the customer
- Customer access the deployed inference model (if it’s deployed in IP MAN in the form of cloud service)
  - The inference model is usually deployed in the edge DC near the customer, ensuring access latency can be guaranteed
  - There are some upstream bandwidth requirements (if access the multimodal inference model)



Transmitting sample data into AIDC

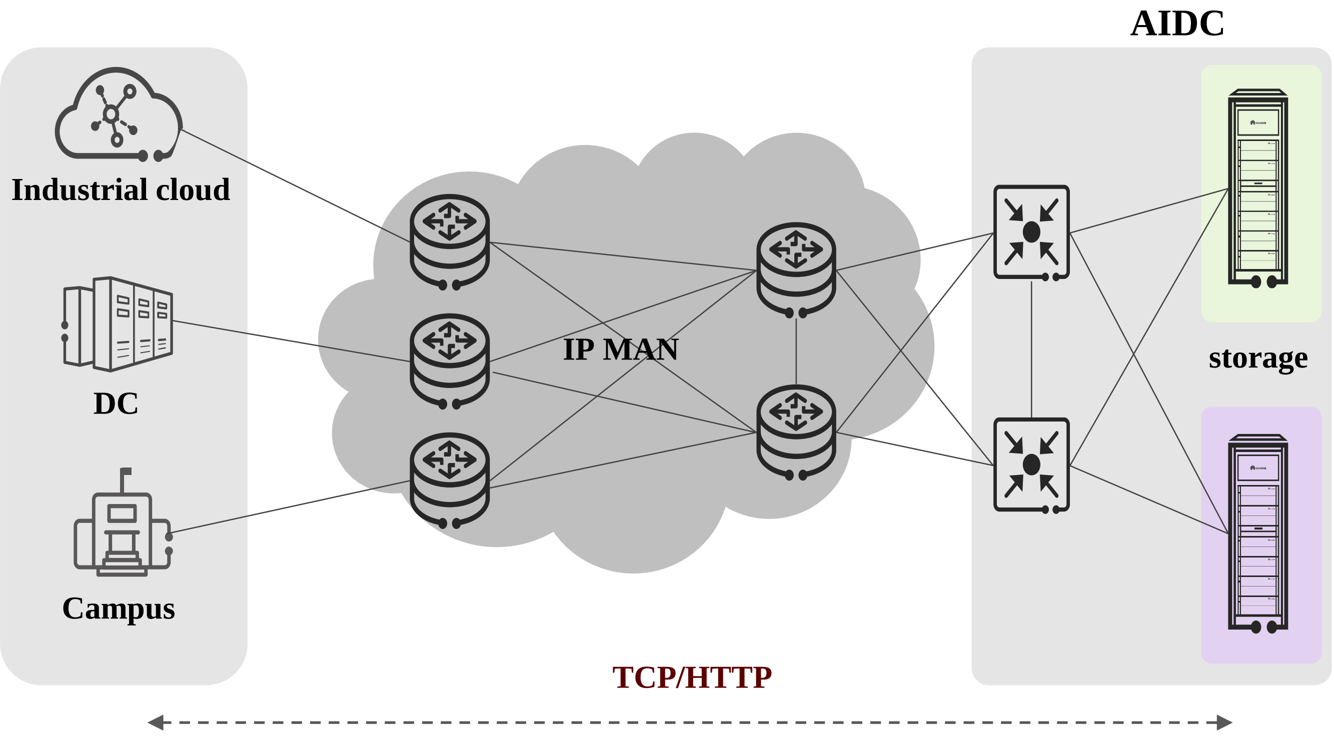
Coordinated model training among multiple AIDCs

Deploying inference model or Returning the trained model

Except that some hyperscalers train foundation models in their private AIDCs, most enterprises prefer leasing computing resources to meet their requirements

Scenario 1: Transmitting sample data to AI cluster

- **Challenge: How to upload sample data into AIDC (usually in storage) in a cost-effective way?**
  - The amount of sample data is very large (usually in the unit of petabytes), but the leaseing time of the computing resources is limited
    - Requiring the transmission of sample data in a relatively short time (e.g. an hour), but the B/W of the customer’s leased line is usually small and fixed



Flexibly providing services

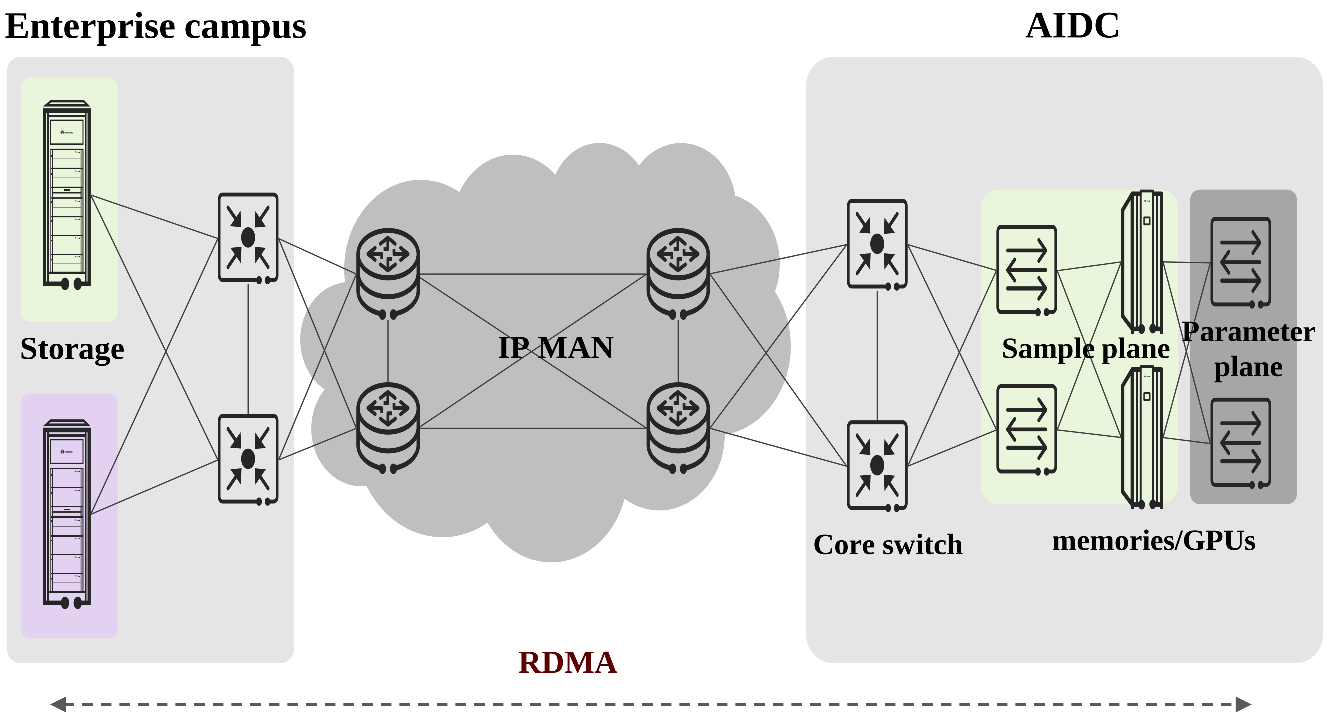
- ✓ Dynamically adjusting the service B/W in the unit of time

Ensuring high throughput

- ✓ Fully utilizing available link bandwidth via SRv6 multipath
- ✓ Achieving efficient traffic engineering and load balancing via ***fast network status notification***

Scenario 2: Transmitting sample data to the memories/GPUs of servers

- **On the base of scenario 1, how to transmit the data into memories/GPUs in real time?**
  - Needs to transmit the data cost-effectively via RDMA with security: Similar to scenario 1, but the customer cares more about the privacy and security of the data
    - Eliminating packet loss: RDMA protocol is extremely sensitive to packet loss (0.1% packet loss will lead to 50% reduction of throughput)
    - High security



Eliminating packet loss

✓ Enabling flow-based precise congestion control via ***immidiate congestion notification***

✓ Deploying NRP technologies to assure the appropriate resources for data trasmission

Improving data security

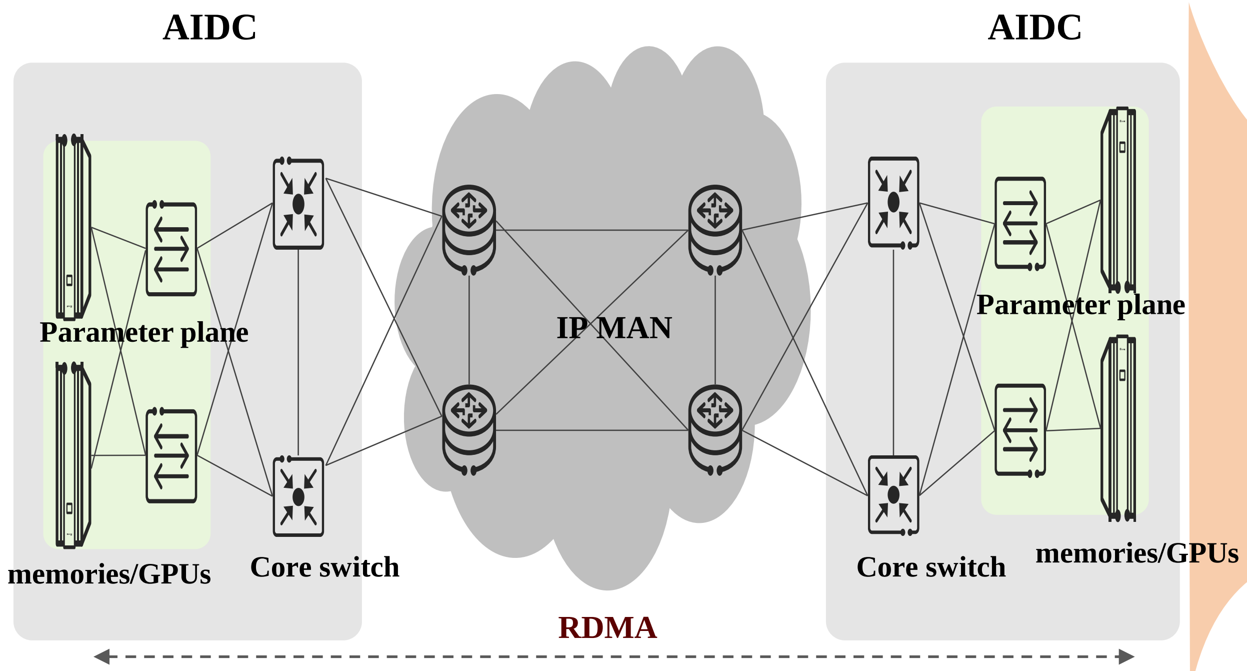
✓ Preventing data leakage via data encryption technologies

Scenario 3: Coordinated model training across AIDCs

- **How to transmit the data (such as parameter data) cost-effectively across AIDCs?**
  - **Time sensitive and packet-loss sensitive**: similar to the scenario 2
  - **Cost-effective transmission**
    - Although the amount of data across AIDCs is very large, it can be lessened by mode optimization (PP/DP/EP et al.)
    - The impact of traffic bursts can be alleviated by increasing the buffer of router

	Number of Parameters & GPUs	
	100B &1K	**1000B &10K**
Data parallelism	3.2Tbps	25.6Tbps
Pipeline Parallelism	4.8Tbps	51.2Tbps

bandwidth requirements for parameter synchronization



Ensuring high throughput

- ✓ Ensuring transmission speed by deploying 400G/800G links
- ✓ Fully utilizing available link bandwidth via SRv6 multipath
- ✓ Achieving efficient traffic engineering and load balancing via ***fast network conditioni notification***

Eliminating packet loss

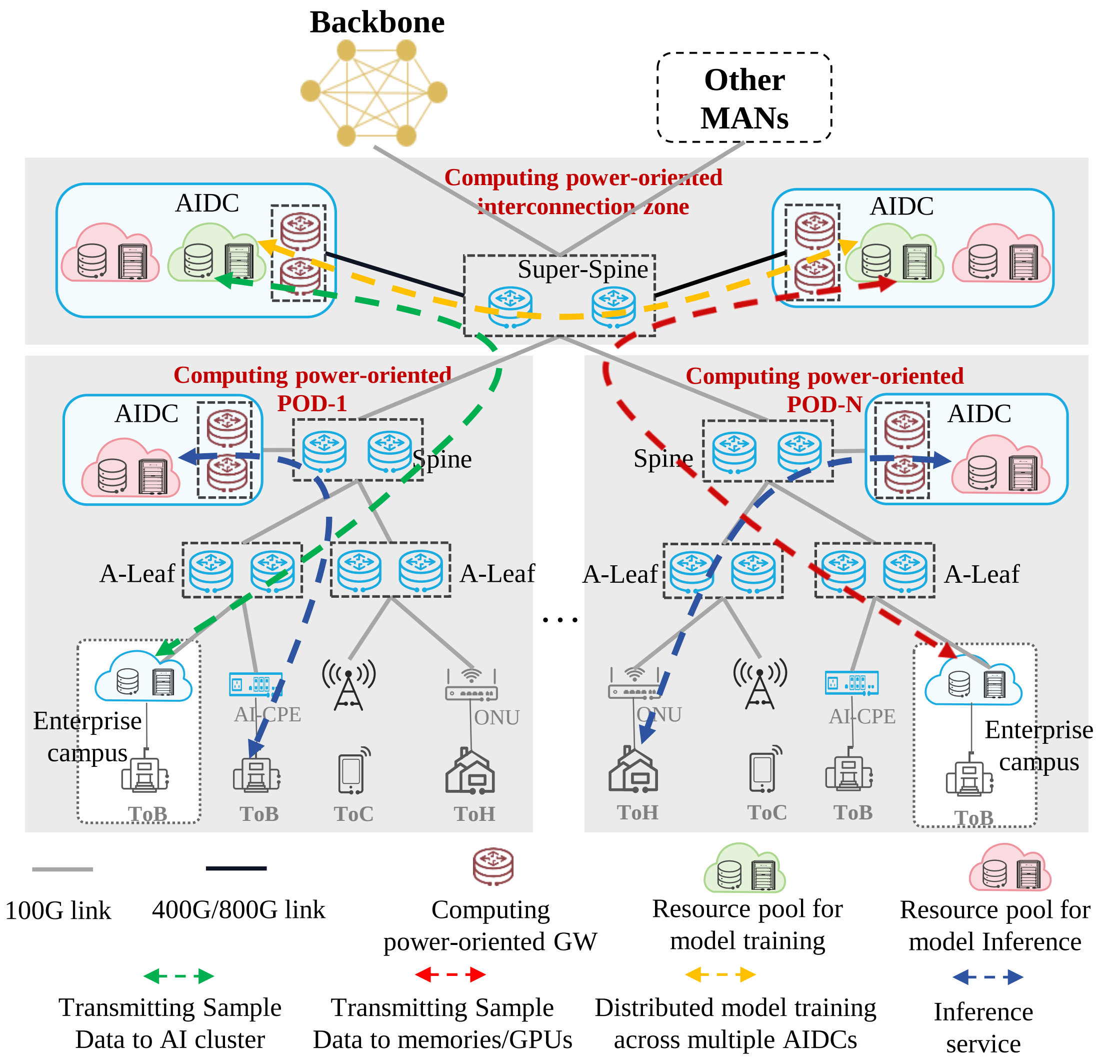
- ✓ Enabling flow-based precise congestion control via ***precise fault control &congestion notification***

Improving bandwidth convergence

- ✓ Mitigating traffic bursts by increasing port buffer

China Telecom is pushing the computing service-oriented MAN forward

Based on the Cloudified IP MAN, the computing service-oriented IP MAN supports various *computing services & real-time services(e.g video)* by optimizing network architecture and enhancing network capabilities


Backbone
Other MANs
Computing power-oriented interconnection zone
AIDC
Super-Spine
AIDC
Computing power-oriented POD-1
AIDC
Spine
Computing power-oriented POD-N
Spine
AIDC
A-Leaf
A-Leaf
A-Leaf
A-Leaf
…
Enterprise campus
AI-CPE
ONU
ToB
ToB
ToC
ToH
ONU
AI-CPE
Enterprise campus
ToH
ToC
ToB
ToB
100G link
400G/800G link
Computing power-oriented GW
Resource pool for model training
Resource pool for model Inference
Transmitting Sample Data to AI cluster
Transmitting Sample Data to memories/GPUs
Distributed model training across multiple AIDCs
Inference service

Network Architecture

- **"Building Blocks" Architecture: easy to expand & flexible to deploy**
  - **Computing service-oriented POD:** based on Spine-Leaf architecture, enabling elastic network scaling and rapid traffic steering
  - **Computing service-oriented interconnection zone:** through computing service-oriented GW to enable fast and standardized integration
- **Network and Service Decoupling: services not affected by network**
  - **Overlay service:** SRv6/EVPN-based (unified protocol stack)
  - **Underlay network:** enhanced network capabilities

Key Capabilities

- **High throughput:** SRv6 multipath, smart traffic engineering based on fast notification
- **"0" packet loss:** precise flow control based on fast notification
- **High security:** tenant-level slicing, data encryption

Takeaway

- **IP MAN covers all the scenarios of AI and real-time services**
- **Besides high-speed links, IP MAN needs more precise technologies, such as precise fault control and fast congestion notification, to meet the new requirements of AI services**
  - **We need the technology to effectively achieve high throughput and lossless transmission**
- **China Telecom is pushing the evolution of IP MAN to meet the requirement of all kinds of real-time services (including computing-related and video service)**
- **Next Steps**
  - Develop and trial the related technologies to meet the customer’s service requirement
  - Push forward the standardization of the related solution
  - deploy the related solution in the next few years

Let's push it forward together!

Questions and Feedback are Welcome

Thanks!