

Use cases and Requirement for Flow Control Collaboration Across DCNs and WAN

[draft-han-rtgwg-codeployment-pfc-fgfc-02](#)

[draft-han-rtgwg-fine-grained-backpressure-01](#)

Zhengxin Han, Ran Pang, Yi Yue, Jie Dong, Zheng Ruan, Quan Xiong

Zhengxin Han (Presenter, China Unicom)

RTGWWG@IETF 125, March 2026

Background: Lossless transmission and flow control expanding from DCN to WAN

- PFC and ECN are adopted in Data Centers (DCs) for lossless transmission.
- With the growth of distributed AI/ML training and inference across geographically separated DCs, the demand for congestion-free data transfer has expanded from DCNs to WANs.
- Unlike DCNs, WAN poses significant challenges to achieving congestion-free data transmission.

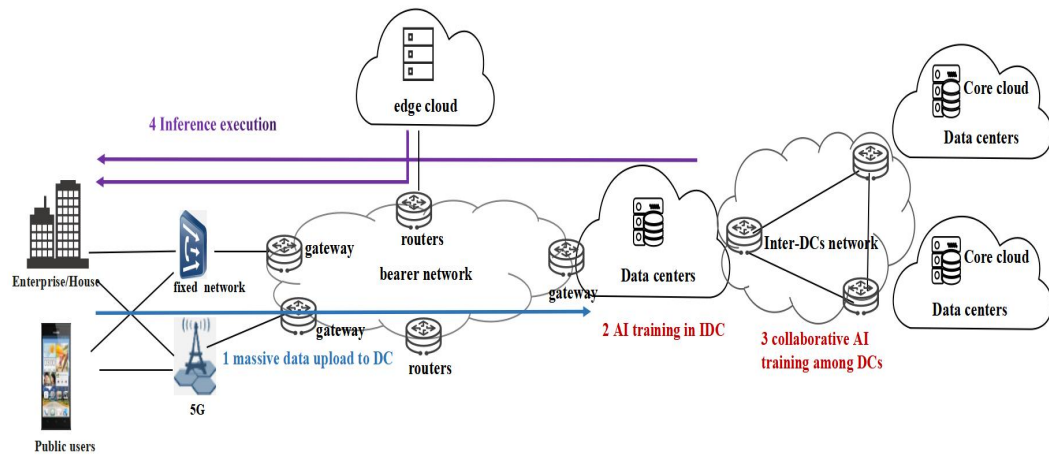


Figure 1: Distributed AI training and inference across multiple AIDCs

Characteristic	DCN	WAN
Network Scale	Limited	Large-scale
Topology	Simple, Predictable	Complex, Dynamic
Path Length	Short	Long (Long RTT)
Traffic Pattern	Predictable	Diverse, Frequent Micro-bursts

Table 1: DCN vs WAN

Limitations of directly applying PFC to WAN

- Mechanism of PFC: port-level feedback for hop-by-hop backpressure over Ethernet, independently controlling up to eight priority queues by pausing or resuming them to prevent congestion.
- Limitations of PFC: including head-of-line blocking, deadlocks, and congestion spreading when applied to WAN, degrade network throughput and utilization.

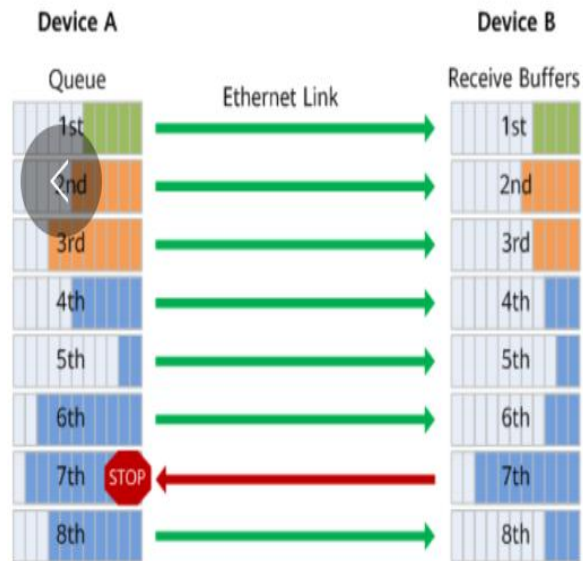


Figure 1 : Mechanism of PFC

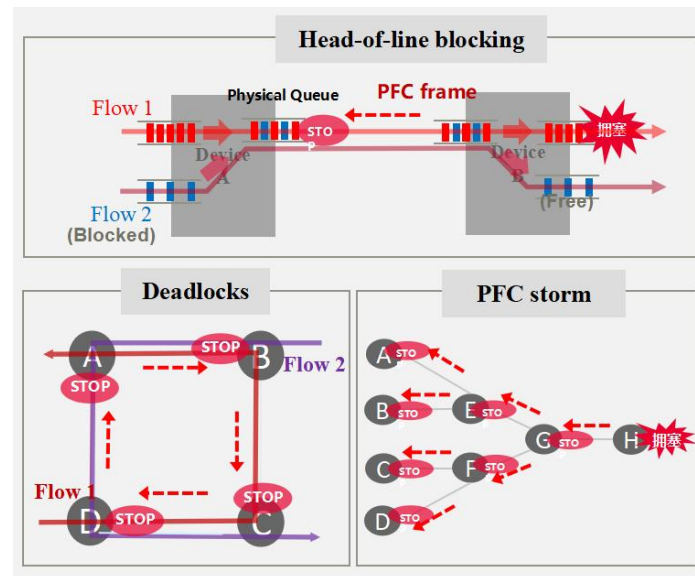


Figure 2: Limitations of PFC

- **Head-of-Line Blocking:**
Port-based backpressure affects all traffic sharing the same queue.
- **Congestion Spreading:**
backpressure propagates to unrelated nodes, with uncontrollable scope.
- **Deadlocks:**
Multipathing may create dependency cycles, leading to deadlock.

➤ **PFC as a Layer 2 mechanism is not suitable for direct deployment in IP WAN.** ◀

Fine-grained flow control for WAN: an enhancement of PFC

[draft-han-rtgwg-fine-grained-backpressure](#)

- ❑ **Fast and precise L3 congestion control** mechanism for IP WAN.
- ❑ **Key Features:**
 - operates in the data plane for **millisecond-level congestion response**.
 - **enables precise flow control at tenants**, flows or other granular levels.
 - **limits flow control to specific paths and slices**.
 - provide intelligent congestion backpressure with granular parameters based on traffic prediction.
- ❑ Extends PFC via network protocols (e.g., ICMPv6,UDP) to carry backpressure message.

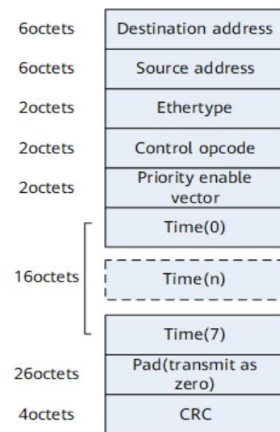


Figure 1:PFC backpressure frame

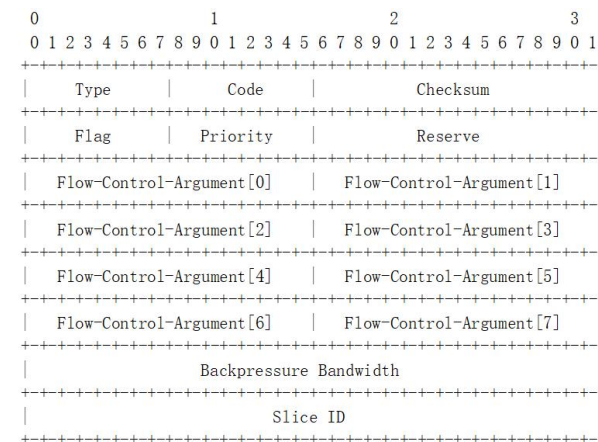
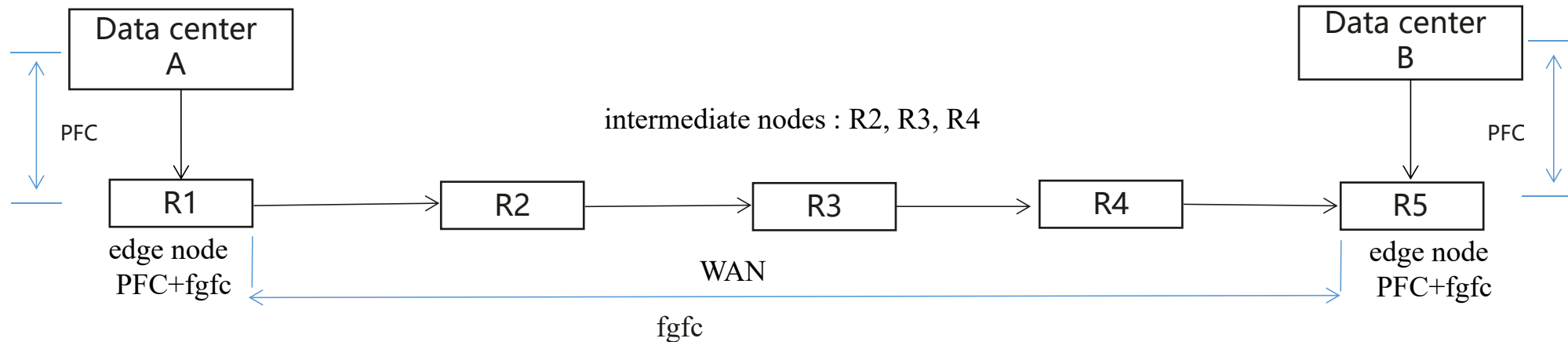


Figure 2: ICMPv6-based Backpressure Message Format (as example)

Codeployment of PFC and fine-grained flow control

□ Motivation (Why?)

- For DCs interconnection over WAN scenario, it requires to **achieve congestion-free data transfer across data centers**.
- **PFC within DCN + fgfc in WAN**. R1 and R5, as the edge nodes between DCN and WAN, need to coordinate PFC and fgfc for **end-to-end flow control**.



Two types of collaborative deployment scenarios

Scenario 1: Single-Hop direct interconnection (no intermediate node participation)

- Tunnel established between edge node R1 and R5 to propagate fgfc message.

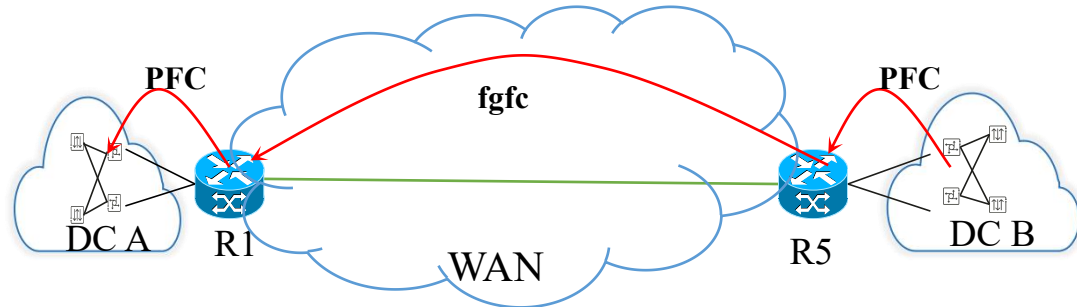


Figure1: no intermediate nodes in WAN

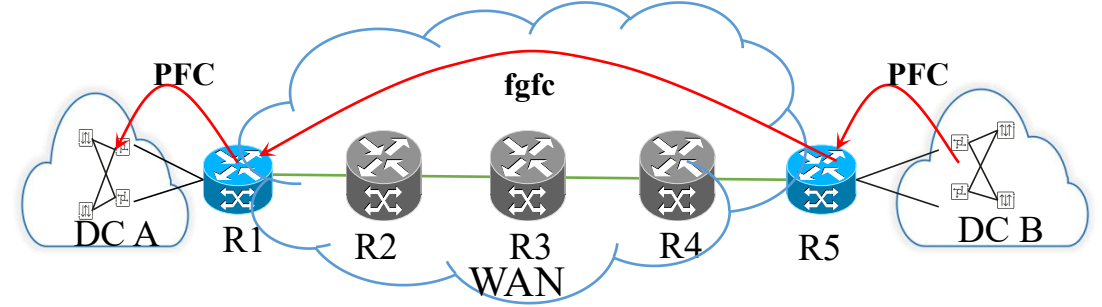


Figure2: legacy intermediate nodes (no support for fgfc) in WAN

Scenario 2: Multi-Hop interconnection (with intermediate node participation)

- Depends on the fgfc capability of intermediate nodes, supports both hop-by-hop and cross-hop backpressure.

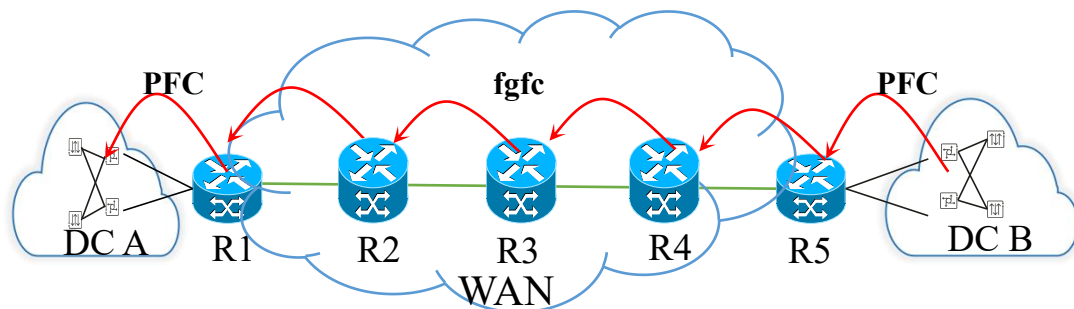


Figure3 :all intermediate nodes in WAN support fgfc

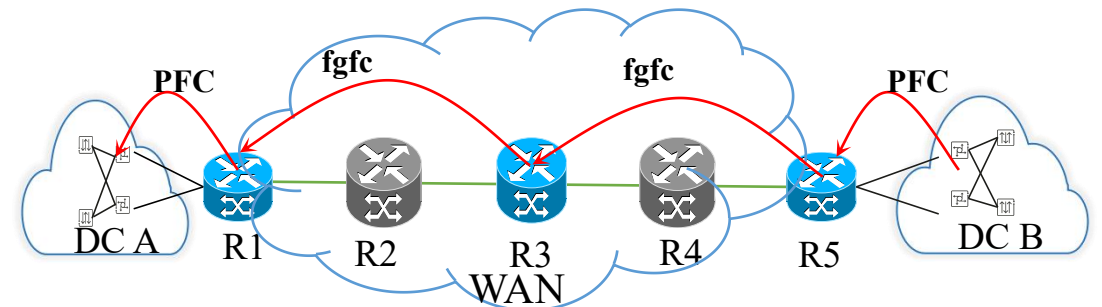
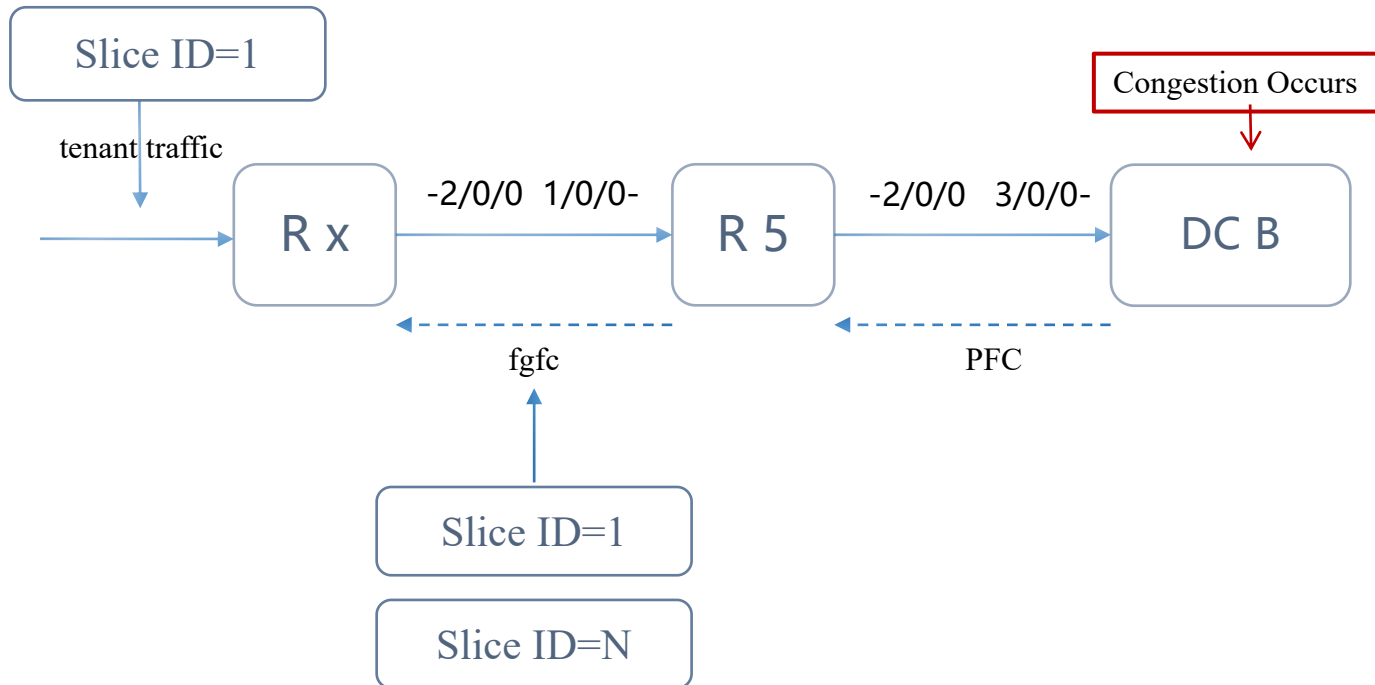


Figure 4: few intermediate nodes (e.g.,R2) in WAN support fgfc

Interworking from PFC to fine-grained flow control

Situation	DCN congestion → WAN backpressure
Edge Node	R5 translates standard PFC → fgfc

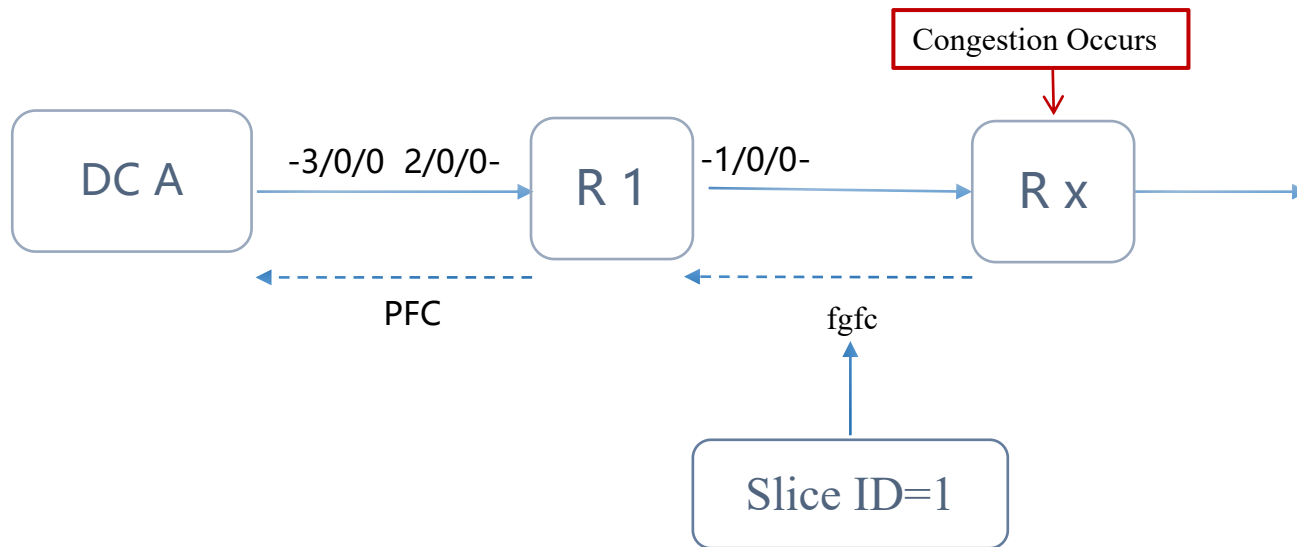
Steps



1. Congestion occurs at DC B incoming port 3/0/0.
2. DC B sends PFC frame to R5 port 2/0/0 (priority af1).
3. R5 respond PFC frame, buffers af1 traffic on port 2/0/0.
4. R5 1/0/0 supports network slicing. When reaches the buffer threshold, R5 1/0/0 sends fgfc packet (af1, sliceID, and pause time, etc.) to upstream node Rx (supports fgfc) .
5. Based on congestion status and scenario, Rx may send fgfc to upstream node, or converts to PFC and forwards to DC.

Interworking from fine-grained flow control to PFC

Situation	WAN congestion → DCN backpressure
Edge Node	R1 translates fgfc → standard PFC



Steps

1. Congestion occurs at Rx (af1, sliceID=1).
2. Rx sends fgfc packet to R1 (af1, slice1, pause time, etc.).
3. R1 responds fgfc packet and buffers tenant traffic with af1 and sliceID = 1.
4. When R1 1/0/0 reaches the buffer threshold, R1 port 2/0/0 sends PFC frame to DC A.
5. DC A performs PFC and stops all traffic with af1 to port 3/0/0.

Requirements of edge node for collaborative deployment

□ Requirement 1: coordination & bidirectional translation between WAN and DCN flow control

- **Protocol Conversion:** PFC ↔ fgfc (fine-grained flow control)
- **Semantic Mapping:** port-level ↔ tenant/flow-level
- **Policy Coordination:** DC policies ↔ WAN Policies
- **Network Capability:** support network slicing and HQoS (Hierarchical Quality of Service)

□ Requirement 2: respond to DCN PFC frames

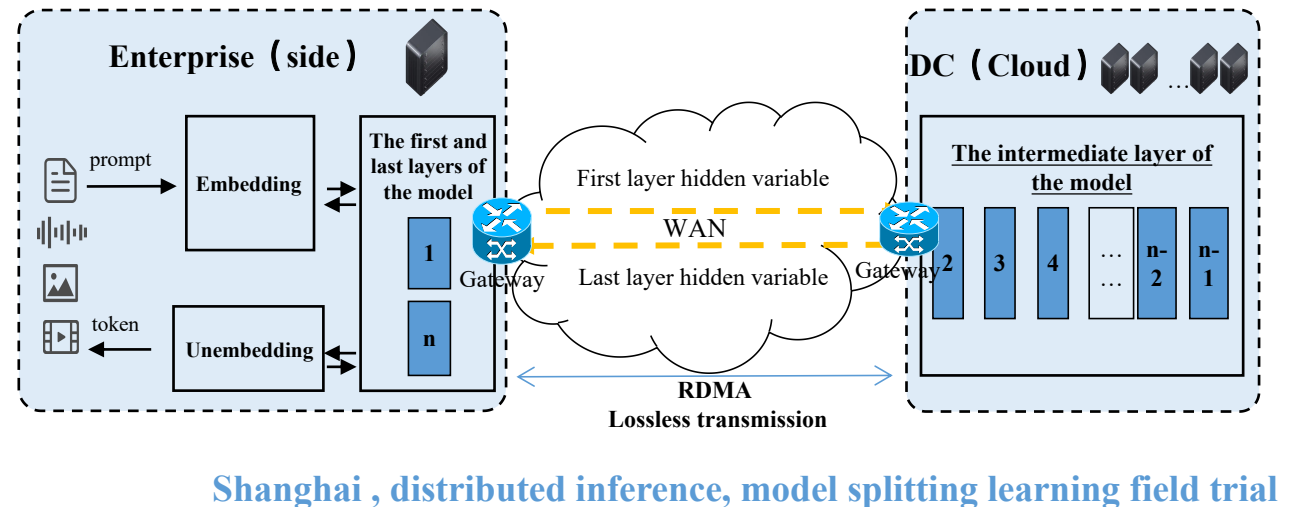
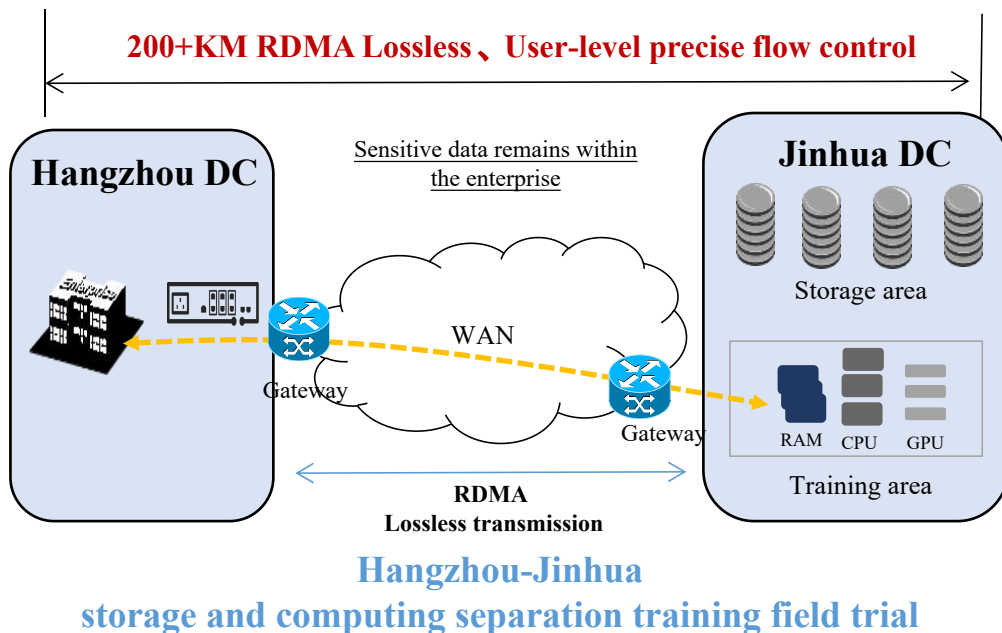
- **Learn flow-to-port mappings:** identify affected tenant traffic and establish multiple slices
- **Manage buffer resources:** configure buffers and thresholds
- **Generate fgfc packets:** send to upstream WAN nodes

□ Requirement 3: respond to WAN fgfc messages

- **Use flow-to-port mappings:** determine target DCN device ports
- **Manage buffer resources:** configure buffers and thresholds
- **Generate standard PFC frames:** send to corresponding DCN device ports

Field trials : Intelligent computing over WAN

- Distributed AI training and inference trials have been conducted over the WAN, **and achieved over 95% computational efficiency.**
- **Fine-grained flow control in WAN and PFC in AIDCs were leveraged** to guarantee RDMA performance over long distance.
- The **gateways met the requirements for collaborative deployment**, successfully achieving end-to-end flow control.



Next steps

- Ask for more reviews and comments.
- Welcome discussion and contributions.
- Maintain alignment with FANTEL and PFC-based related congestion control drafts.
- Promote network testing and validation.

THANKS !