

RTGWG  
Internet-Draft  
Intended status: Standards Track  
Expires: 23 April 2026

Z. Han, Ed.  
R. Pang  
Z. Ruan  
X. Yi  
China Unicom  
20 October 2025

Usecase and requirement of deploying PFC and fine-grained flow control  
draft-han-rtgwg-codeployment-pfc-fgfc-01

## Abstract

The demand for lossless network transmission and the application of flow control mechanisms have expanded from DCNs (Data Center Networks) to WANs(Wide Area Networks). To mitigate PFC - related issues in WANs, the fine - grained flow control is proposed. This mechanism aims to achieve precise control at flow / tenant levels, limits flow control to specified paths and slices, and provides intelligent congestion backpressure. As current DCN already adopts PFC mechanisms, the fine-grained flow control in WANs needs to work with PFC in DCNs to achieve end-to-end flow control.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 23 April 2026.

## Copyright Notice

Copyright (c) 2025 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document.

Han, et al.

Expires 23 April 2026

[Page 1]

Internet-Draft Req of PFC and fine-grained flow control October 2025

Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Introduction and Background {#intro and backg}	2
2. Terminology	3
3. Interworking deployment of PFC and fine-grained Flow Control	3
4. Procedure of end-to-end flow control	4
4.1. PFC to fine-grained flow control	4
4.2. Fine-grained flow control to PFC	5
5. Requirement of joint deployment	6
6. Security Considerations	7
7. IANA Considerations	7
8. Informative References	7
Authors' Addresses	7

### 1. Introduction and Background {#intro and backg}

DCNs are typically characterized by a limited network scale, short path and predictable traffic patterns, so flow control mechanisms like PFC (Priority Flow Control) and ECN (Explicit Congestion Notification) operate effectively. With the growth of AI LLM distributed training and inference, lossless transmission of massive data between geographically separated data centers is required [I-D.hs-rtgwg-wan-lossless-uc], and the flow control mechanisms need to be extended from DCNs to WANs. Unlike DCNs, WANs are large-scale with complex topologies, long paths, and diverse traffic type. PFC based on port-level feedback ensures lossless transmission of RDMA protocol, by pausing/resuming specific priority queues to prevent congestion. When using it in the WANs, the backpressure from PFC will cause head-of-line blocking, deadlocks, and congestion spreading, which degrade network throughput [I-D.hs-rtgwg-wan-lossless-uc]. To mitigate these issues, the fine-grained flow control is required for WANs.

Fine-grained flow control improves upon the coarse-grained port-based PFC mechanism. It enables precise control at the flow, tenant, or other granular levels, limits flow control to specified paths and slices, and provides intelligent congestion backpressure with granular parameters (pausing time, and buffer thresholds etc.). These capabilities collectively contribute to achieving efficient and refined flow control in WANs.

This draft focuses on the scenarios where PFC is employed in DCNs and the fine-grained flow control is utilized in WANs. Usecase and requirements for the interworking deployment of PFC and fine-grained flow control mechanisms are described, achieving end-to-end flow control through coordination and policy mapping between DCNs and WANs.

### 2. Terminology

PFC: Priority-based Flow Control

DCN: Data Center Network

WAN: Wide Area Network

RDMA: Remote Direct Memory Access

RoCE: RDMA over Converged Ethernet

### 3. Interworking deployment of PFC and fine-grained Flow Control

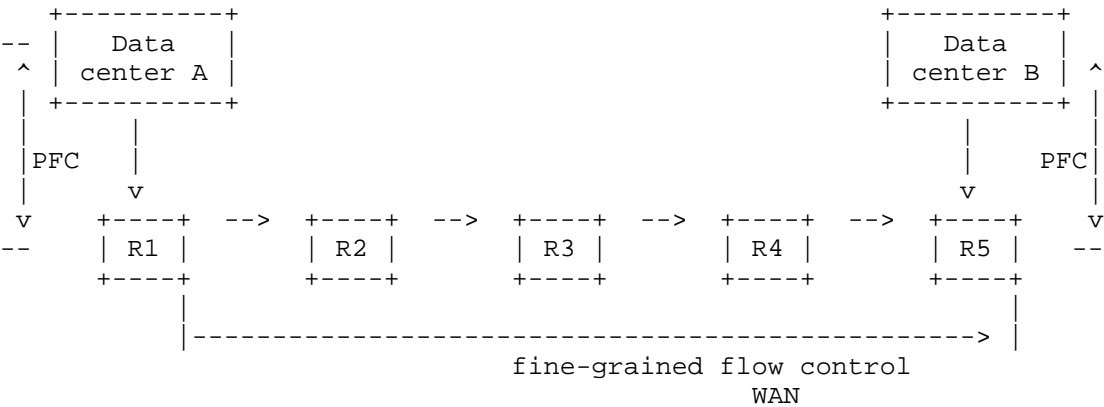


Figure 1: Codeployment of PFC and fine-grained flow control

As shown in Figure 1, there are two data centers, A and B, connected by WAN via nodes R1 -> R2 -> R3 -> R4 -> R5.

The internal nodes of data center A and data center B employ the PFC mechanism. Because most DCN NICs today are optimized for legacy protocols (e.g., Ethernet, DCB) and lack SRv6 processing capabilities. This limitation prevents the direct extension for refined flow control. Hardware/firmware upgrades are needed to enable fine-grained flow control deployment.

WAN nodes R1-R5 deploy fine-grained flow control to avoid PFC backpressure issues, enabling flow/tenant-level congestion handling with granular parameters for precise and intelligent backpressure. WAN nodes support HQOS (Hierarchical Quality of Service) queuing mechanisms and slicing.

Edge nodes R1 and R5 support both PFC and fine-grained flow control, interworking DCN and WAN flow control mechanisms and ensuring seamless end-to-end flow control. The NNI ports of edge nodes R5 and R1 can establish multiple slices, each corresponding to a tenant and supporting 1-8 queues.

### 4. Procedure of end-to-end flow control

#### 4.1. PFC to fine-grained flow control



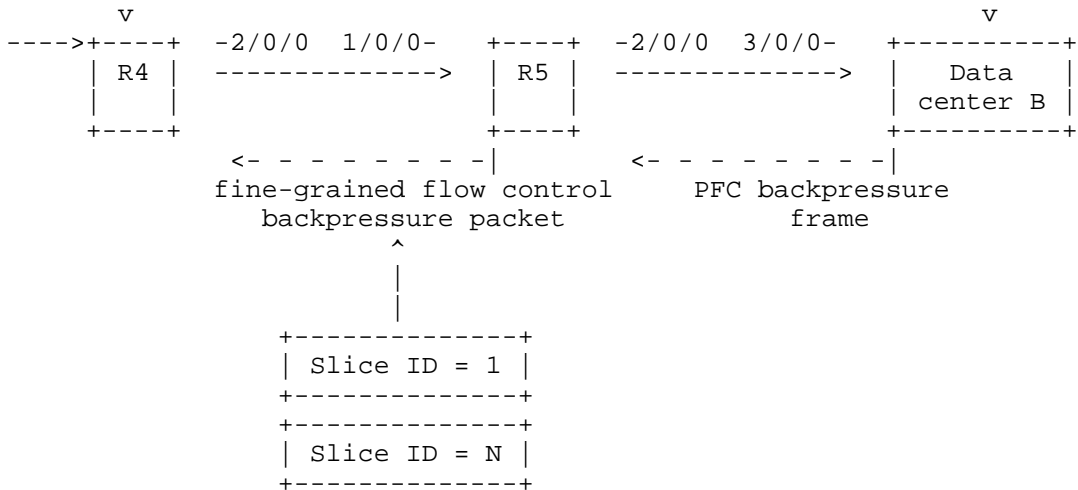
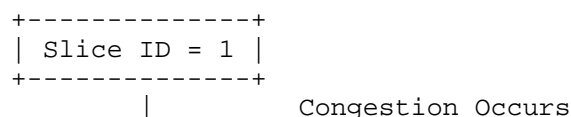


Figure 2: PFC to fine-grained flow control

Edge node R5 responds to the PFC frame sent by the data center and transmits fine - grained flow control packet to the WAN. The process follows these steps:

- 1) When congestion occurs at the incoming port 3/0/0 of data center B.
- 2) The data center B sends a PFC backpressure frame to the 2/0/0 port of edge node R5. The PFC frame carries the queue priority of the traffic to be backpressured, which is af1.
- 3) Edge node R5 needs to support responding to the PFC frame and buffers the traffic with the priority af1 through the 2/0/0 physical port.
- 4) The 1/0/0 port of edge device R5 has multiple slices. When the buffer queue corresponding to the 2/0/0 port of edge device R5 reaches the buffer threshold.
- 5) According to the port, tenant traffic, and slice mapping relationship, the 1/0/0 port of edge device R5 sends a fine - grained flow control backpressure packet to network node R4. The packet carries the tenant traffic information to be backpressured, with the queue priority af1, sliceID, and pause time, etc.
- 6) Based on the congestion handling situation, network node R4 sends fine - grained flow control packets to the upstream WAN nodes as needed.

#### 4.2. Fine-grained flow control to PFC



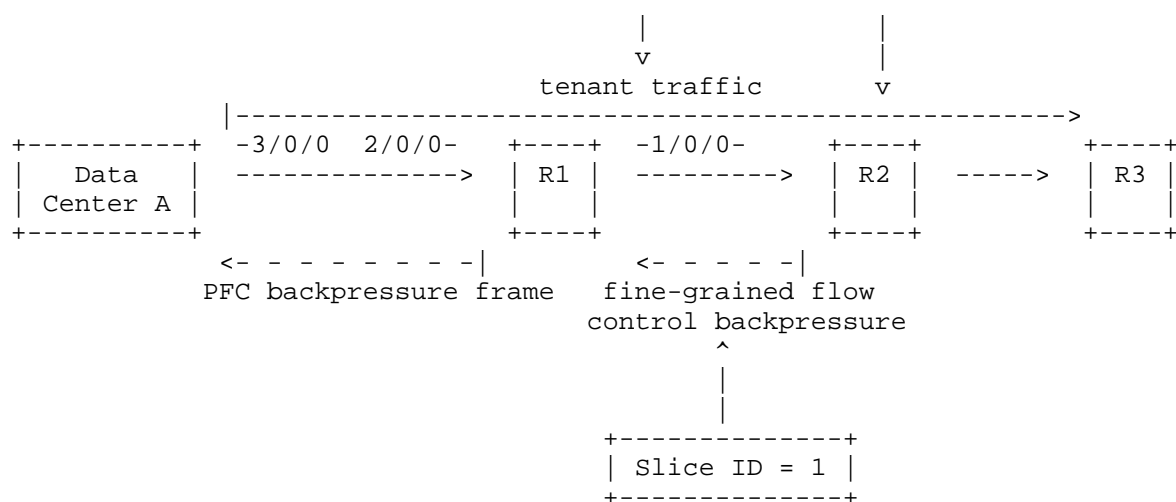


Figure 3: fine-grained flow control to PFC

Edge node R1 responds to fine - grained flow control packet from WAN, then sends PFC frame to the data center. The process follows these steps:

- 1) When congestion occurs in the traffic of queue afl with sliceID = 1 at the egress port of network node R2.
- 2) Network node R2 sends a fine - grained flow control backpressure packet to edge node R1. This packet carries the tenant traffic information to be backpressured, with the queue priority afl, sliceID = 1, and the pause timed, etc.
- 3) Edge node R1 performs traffic control and buffers the tenant traffic with priority afl and sliceID = 1.
- 4) When the buffer queue corresponding to port 1/0/0 of edge device R1 reaches the buffer threshold, port 2/0/0 of edge node R1 sends backpressure to the data center according to the standard PFC packet.
- 5) Data center A performs standard PFC backpressure and stops all traffic with priority afl destined for port 3/0/0.

## 5. Requirement of joint deployment

Edge node needs support the coordination and bidirectional translation between the fine-grained flow control mechanism in the WAN and the PFC mechanism in the DCN, enabling seamless end-to-end flow control across WAN and DCN domains.

Edge node needs to respond to PFC frames from the DCN:

- a) Learn task flow-to-port mappings to identify affected traffic;
- b) Configure appropriate buffer thresholds;
- c) Generate and send fine-grained flow control messages to WAN nodes with granular parameters.

Edge nodes needs to respond to fine-grained flow control messages from the WAN:

- a) Use established flow-to-port mappings to determine target DCN

ports;  
b) Configure appropriate buffer thresholds;

c) Generate and send standard PFC frames to corresponding DCN ports.

## 6. Security Considerations

This document does not introduce any new security considerations.

## 7. IANA Considerations

This document has no IANA actions.

## 8. Informative References

[I-D.hs-rtgwg-wan-lossless-uc]  
Zhengxin, H., He, T., Shi, H., and T. Zhou, "Use Cases and Requirements for Implementing Lossless Techniques in Wide Area Networks", Work in Progress, Internet-Draft, draft-hs-rtgwg-wan-lossless-uc-01, 2 July 2025, <<https://datatracker.ietf.org/doc/html/draft-hs-rtgwg-wan-lossless-uc-01>>.

## Authors' Addresses

Zhengxin Han (editor)  
China Unicom  
Beijing  
China  
Email: hanzx21@chinaunicom.cn

Ran Pang  
China Unicom  
Beijing  
China  
Email: pangran@chinaunicom.cn

Zheng Ruan  
China Unicom  
Beijing  
China  
Email: ruanz6@chinaunicom.cn

Xinxin Yi  
China Unicom  
Beijing  
China  
Email: yixx3@chinaunicom.cn