RTGWG Working Group
Internet-Draft
Intended status: Standards Track
Expires: 18 August 2026

X. He
L. Deng
China Telecom

14 February 2026

PFC PAUSE Frame Forwarded Transparently in Wide Area Networks
draft-he-rtgwg-wan-pfc-00

Abstract

This document describes a solution for transparent forwarding of PFC PAUSE frames in wide area networks, which does not require the nodes in wide area networks to support PFC flow control capabilities.

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 18 August 2026.

Copyright Notice

Table of Contents

1. Introduction 2
2. Conventions 3
2.1. Requirements Language 3
2.2. Terminology 3
3. Transparent Forwarding of PFC PAUSE Frames in WANs 3
3.1. Flow Control Mechanism For PFC Frame 4
3.2. PFC PAUSE Frame Processing 5
4. IANA Considerations 7
5. Security Considerations 7
6. References 7
6.1. Normative References 7
6.2. Informative References 7
Authors' Addresses 7

1. Introduction

Remote Direct Memory Access (RDMA) is a method of accessing memory on a remote system without interrupting the processing of the Central Processing Unit (CPU) on that system. RDMA enables lower latency and higher throughput on the network and lower CPU utilization for the servers and storage systems. Currently, RoCEv2 (RDMA over Converged Ethernet Version 2) is widely deployed in lossless networks in intelligent computing centers, providing packet loss free data transmission services for high-performance computing (HPC) and AI model training and inference scenarios.

With the rapid growth in demand for computing and storage resources in AI big models and distributed storage, intelligent computing centers are interconnected through wide area networks (WANs) to provide multi-DCs collaboration to compensate for the limitations of insufficient computing and storage resources in a single DC. The interconnection of artificial intelligence Data Centers (AIDCs) through WANs are becoming a new network structure gradually accepted by the industry, providing wide area lossless transmission for emerging application scenarios. Priority-based Flow Control(PFC)[IEEE8021Q-2022] technology is widely deployed in RoCEv2 networks to aviod packet loss caused by congestion. However, the deployment of PFC in WANs may lead to head-of-line blocking, deadlocks, and even congestion diffusion over a wider range, which will degrade network performance. On the other hand, WANs need to provide differentiated services for various applications, and there exist differences in buffering capacity from different nodes as well as link delay metrics between two nodes, leading to inconsistent parameters configuration of node, which makes network operation and maintenance more complicated. Therefore, PFC mechanism is not suitable for large-scale deployment in WANs.

This document describes a solution for transparent forwarding of PFC PAUSE frames in wide area networks, which does not require the nodes in WANs to support PFC flow control capabilities. As a result, end-to-end flow control between AIDCs interconnected through MANs can be realized with minimal impact on network performance.

2. Conventions

2.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2.2. Terminology

Abbreviations used in this document:

AI: Artificial Intelligence

AIDC: Artificial Intelligence Data Center

DC: Data Center

MAC: Media Access Control

P: Provider

PE: Provider Edge

PFC: Priority-based Flow Control

RDMA: Remote Direct Memory Access

RoCEv2: RDMA over Converged Ethernet version 2

SR-MPLS: Segment Routing Based on Multiprotocol Label Switching

SRv6: Segment Routing over IPv6

VXLAN: Virtual Extensible Local Area Network

WAN: Wide Area Network

3. Transparent Forwarding of PFC PAUSE Frames in WANs

3.1. Flow Control Mechanism For PFC Frame

The PFC is referred to as classical stepwise back pressure with dedicated Ethernet pause frame, which is widely deployed in RoCEv2 networks to aviod packet loss caused by congestion. The PFC PAUSE frame format is shown in Figure 1.

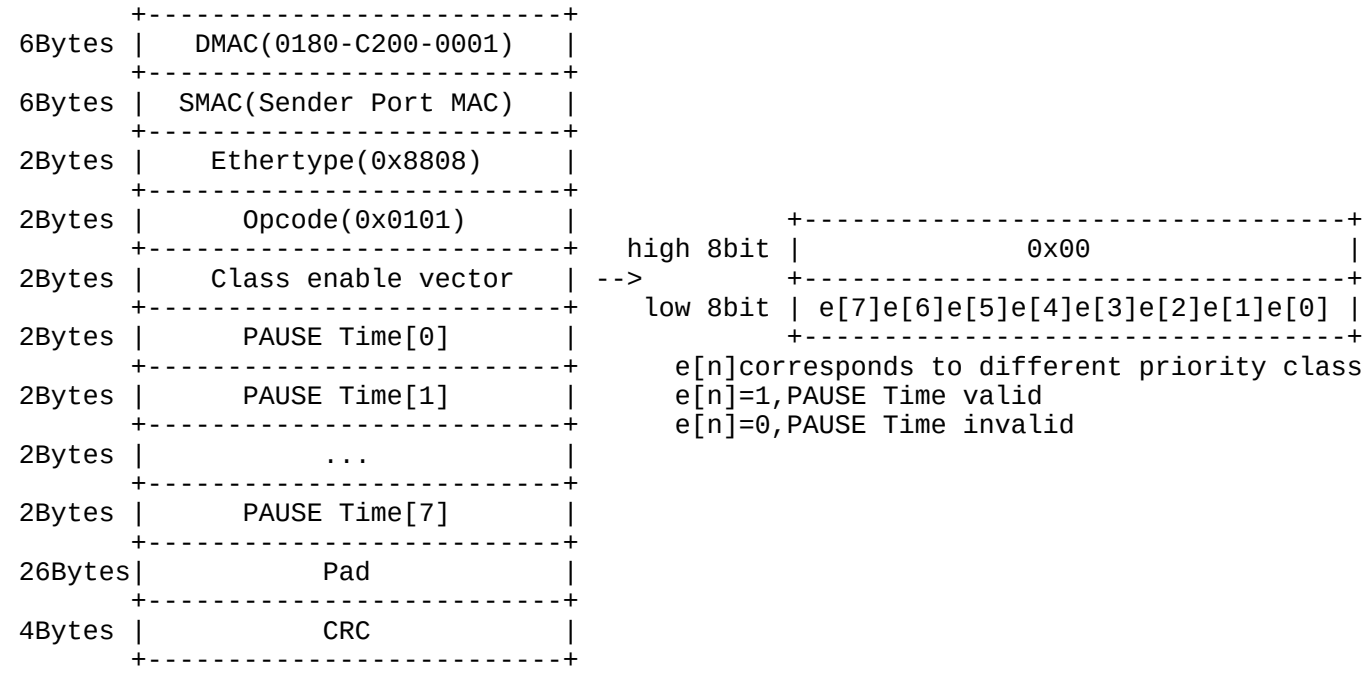


Figure 1: PFC PAUSE Frame Format

With this flow control mechanism, the congested node asks the directly connected upstream network node to pause the data traffic by a dedicated Ethernet pause frame called PFC frame, and then the upstream network node may stepwise ask its directly connected upstream network node to pause the data traffic by a PFC frame, until the most upstream network node may ask the directly connected traffic sender to pause the data traffic by a PFC frame. [IEEE8021Q-2022] details how this kind of flow control mechanism works.

Typically, when two AIDCs are interconnected through WANs, VPN tunnels (e.g., SR-MPLS, SRv6, VXLAN) are established between the ingress PE and egress PE to carry massive RDMA traffic between DCs, as shown in Figure 2.

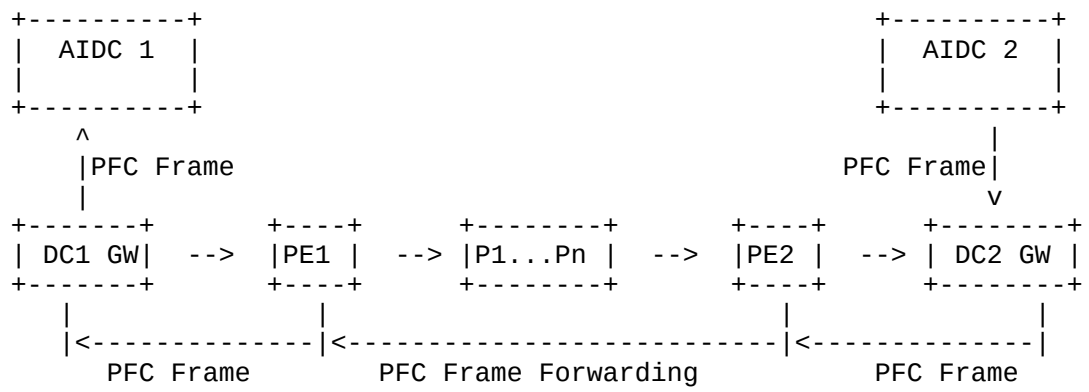


Figure 2: AIDCs Interconnected Through WANs

3.2. PFC PAUSE Frame Processing

When congestion occurs in the destination AIDC, the PFC frames are stepwise sent to the destination DC gateway. Similarly, the destination DC may stepwise ask its directly connected upstream egress PE node to pause the data traffic by sending a PFC frame. In Figure 2, AIDC 2 sends the PFC frames to DC2 gateway, and in turn, DC2 gateway sends the PFC frames to PE2 When congestion occurs at the recieved port.

When the egress PE node of WAN receives a PFC frame, it needs to parse a PFC frame and determine that it is a legal PFC frame, that is, besides its correct frame format, its destination MAC address must be the multicast address: 0180-C200-0001 and the source MAC address must be its directly connected downstream DC gateway port MAC address (some vendors also use device system MAC address). Otherwise, the egress PE node must discard this illegal PFC frame.

The egress PE node encapsulates the PFC frame based on tunnel encapsulation protocol, then forwards it to the immediate transit node, which in turn forwads it transparently to the upstream node until it reaches the ingress PE node.

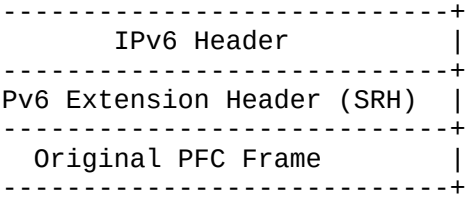
The ingress PE node decapsulates the PFC frame and replaces the source MAC address in the original PFC frame with the MAC address of its port directly connected to the source DC gateway, then forwards it to the source DC gateway.

In order to ensure that the PFC frames can be forwarded to the ingress PE quickly, it is preferable to configure the highest priority for the encapsulated PFC frames such that the PFC frames are not discarded in case of network congestion.

Similarly, the source DC gateway needs to parse the forwarded PFC frame and determine that it is a legal PFC frame, that is, besides its correct frame format, its destination MAC address must be the multicast address: 0180-C200-0001 and the source MAC address must be its directly connected ingress PE port MAC address(some vendors also use device system MAC address). Otherwise, the source DC gateway must discard this illegal PFC frame.

the source DC gateway sends the PFC frames to the source AIDC (AIDC1 in Figure 2) When congestion occurs at the recieved port. Consequently, end-to-end flow control between AIDCs can be realized across WANs.

An example is that two AIDCs are interconnected through SRv6 tunnel in WANs. The encapsulated PFC frame format is depicted as follows:



Due to the much longer transmission distance of WANs compared to Internal DCs , the PFC frames forwarded from the egress PE to the ingress PE require a significant transmission delay. The destination DC gateway still needs to receive the data traffic continuously sent from the source DC gateway until the source DC gateway receives the PFC frames and pauses sending the corresponding priority data traffic. The amount of data received by the destination DC gateway is positively correlated with the transmission delay of PFC frame. To avoid packet loss caused by overflow in the receiving port queue, the destination DC gateway needs to reserve more buffer for the corresponding priority queue of the receiving port based on WAN transmission delay of PFC frame.

The reserved buffer setting for the priority queue of the receiving port at the destination DC gateway is required to meet the following condition.

The buffer size of the priority queue reserved for the receiving port > (the average receiving rate of the corresponding priority flow at the receiving port - the average sending rate of the corresponding priority flow at the sending port) * the forwarding delay of the PFC frame from the destination DC gateway to the source DC gateway.

4. IANA Considerations

This document has no IANA actions.

5. Security Considerations

This document does not introduce any new security considerations.

6. References

6.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

6.2. Informative References

[IEEE.802.1Q.2022] IEEE, "IEEE Standard for Local and Metropolitan Area Networks--Bridges and Bridged Networks", IEEE 802-1q-2022, DOI 10.1109/IEEESTD.2022.10004498, 30 December 2022, <https://ieeexplore.ieee.org/document/10004498>.

Authors' Addresses

Xiaoming He
China Telecom
Email: hexm4@chinatelecom.cn

Lijie Deng
China Telecom
Email: denglj4@chinatelecom.cn