RTGWG Working Group
Internet-Draft
Intended status: Standards Track
Expires: 31 August 2026

X. Min
ZTE Corp.
K. Zhang
China Mobile

27 February 2026

Congestion Notification for Pause
draft-xiao-rtgwg-congestion-notification-for-pause-01

Abstract

This document describes the necessity and feasibility to introduce a mechanism of congestion notification for pause. After receiving the L2 pause frames from the destination data center gateway, the egress provider edge node sends the congestion notifications to the upstream provider nodes and the ingress provider edge node in a format defined in this document. The upstream provider nodes and the ingress provider edge node must pause the forwarding of IP flows identified by the congestion notifications. And then the ingress provider edge node may send the L2 pause frames to the source data center gateway.

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 31 August 2026.

Table of Contents

1. Introduction

IP based VPN [RFC2764] is often used to interconnect Data Center Networks (DCN), in which case the IP based VPN is also referred to as IP WAN. In the DCN, Priority-based Flow Control (PFC) [IEEE8021Q-2022] is a widely deployed mechanism for congestion control. However, the PFC as an L2 pause mechanism is not suitable to be deployed in IP WAN, so an L3 pause mechanism is needed for use in IP WAN.

This document describes the necessity and feasibility to introduce a mechanism of congestion notification for pause. Specifically, the problem statement is described in Sections 1 and 3, and the format of the congestion notification message sent from the Provider Edge (PE) node to the Provider (P) and/or PE node is defined in Section 4, and the solution on how the PE node knows the addresses of the destined P and/or PE node is defined in Section 4 and 5.

2. Conventions Used in This Document

2.1. Abbreviations

CE: Customer Edge

DC: Data Center

DCN: Data Center Networks

DoS: Denial-of-Service

IPC: IP Pause Capability

LSA: Link State Advertisement

P: Provider

PE: Provider Edge

PFC: Priority-based Flow Control

RI: Router Information

SRH: Segment Routing Header

SRv6: Segment Routing over IPv6

2.2. 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.

3. Congestion Notification Mechanisms

As a congestion notification for pause mechanism used in DCN, the PFC is referred to as classical stepwise back pressure with dedicated Ethernet pause frame, as shown in Figure 1.

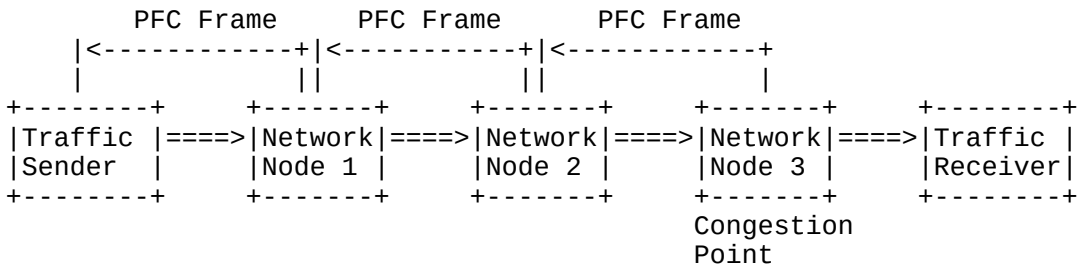

Figure 1: Classical Stepwise Back Pressure with Dedicated Ethernet Pause Frame within DC

With this congestion notification mechanism, the congested network node (Netwok Node 3 in Figure 1) asks the directly connected upstream network node (Network Node 2 in Figure 1) 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 (Network Node 1 in Figure 1) may ask the directly connected traffic sender to pause the data traffic by a PFC frame. [IEEE8021Q-2022] details how this kind of congestion notification mechanism works.

In the IP WAN for DC interconnect, the congestion notification mechanism triggered by the PFC frames from the destination DC gateway is referred to as back pressure with dedicated IP pause packet, as shown in Figure 2.

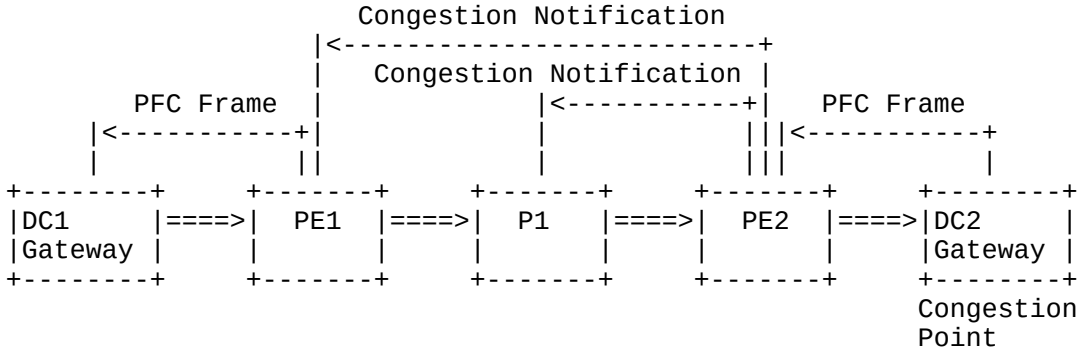

Figure 2: Back Pressure with Dedicated IP Pause Packet within WAN

With this congestion notification mechanism, while detecting congestion the congested egress Customer Edge (CE) node (DC2 gateway in Figure 2) asks the directly connected upstream egress PE node (PE2

in Figure 2) to pause the data traffic by sending PFC frames, and in response to receiving the PFC frames from DC2 which is in congestion, the egress PE generates IP flow pause packets corresponding to the IP flows which cause the congestion in DC2, and then the egress PE asks the upstream P node (P1 in Figure 2) and/or the upstream ingress PE node (PE1 in Figure 2) to pause (buffer) the data traffic of IP flows by sending the IP flow pause packets, until the ingress PE node may ask the directly connected upstream ingress CE node (DC1 gateway in Figure 2) to pause the data traffic by sending PFC pause frames. Note that the upstream P node and/or the upstream ingress PE node receiving the IP flow pause packets must be on the forwarding path of the IP flows and must have the buffering capability for the IP flows causing congestion. This document details how this kind of congestion notification mechanism works.

4. Congestion Notification for Pause Packet Format

Once receiving the L2 pause frames from the destination DC gateway, the egress PE node needs to determine which IP flows cause the congestion. How the egress PE node figure out the IP flows causing congestion is implementation specific and outside the scope of this document. For each IP flow causing congestion, the egress PE node needs to identify the ingress PE node and the P nodes traversed by the IP flow and send congestion notification for pause message to each identified P/PE node. With respect to different WAN technologies, there are different ways for the egress PE node to identify the on-path PE and P nodes. When Segment Routing over IPv6 (SRv6) [RFC8754] is deployed in the WAN, the egress PE node can use Segment Routing Header (SRH) to identify the on-path PE and P nodes; When native IPv6 is deployed in the WAN, the egress PE node can only use the source IP address to identify the ingress PE node.

The congestion notification for pause message sent from the egress PE node to the identified on-path PE and P nodes can be a UDP message or an ICMP message, if a UDP message it’s formatted as follows:

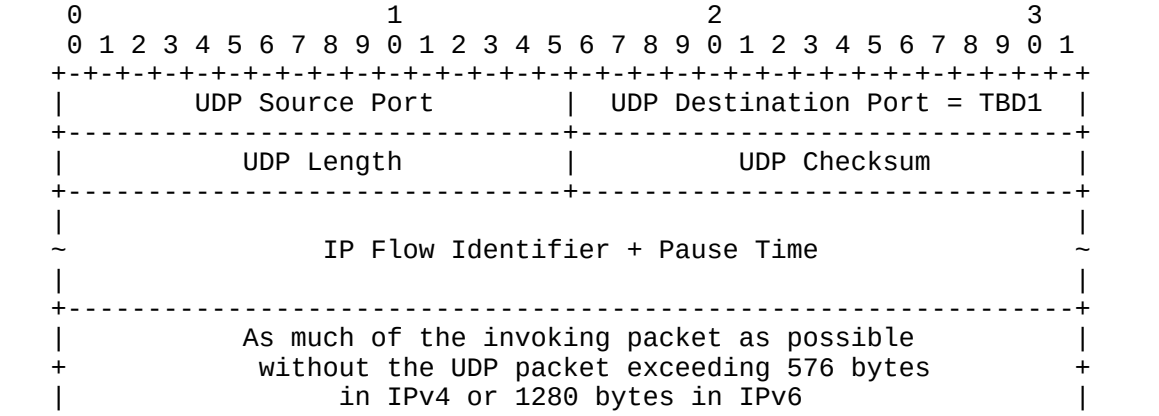

Figure 3: Congestion Notification for Pause Message Format

UDP Header: The UDP header as specified in [RFC768] includes the UDP source port, UDP destination port, UDP length, and UDP checksum. A well-known UDP destination port (TBD1) needs to be allocated for this Congestion Notification Message.

IP Flow Identifier: When SRv6 is deployed in the WAN, the IP Flow Identifier includes the source IP address and the SRH; When native IPv6 is deployed in the WAN, the IP Flow Identifier includes the source IP address, destination IP address, and protocol number.

Pause Time: This field can be either copied from the PFC Pause frames receiving from the DC gateway, or calculated based on the buffer size of the destined node advertised by IGP.

5. Advertising IP Pause Capability Using IGP

Considering that not all WAN routers support buffering IP flows, before the egress PE node can send the congestion notification for pause message to the on-path PE and P nodes, the egress PE node has to know which on-path P/PE nodes support buffering IP flows. The on-path P/PE nodes can notify the egress PE node of its support of buffering IP flows by advertising its IP Pause Capability (IPC) in advance.

5.1. Advertising IP Pause Capability Using IS-IS

The PE and P nodes advertise their support of buffering IP flows by inserting a new IPC sub-TLV into the IS-IS Router Capability [RFC7981]. This sub-TLV SHOULD only be advertised once in the Router Capability TLV. This sub-TLV SHOULD be advertised WAN domain wide. The IP Pause Capability sub-TLV is structured as shown in Figure 4.

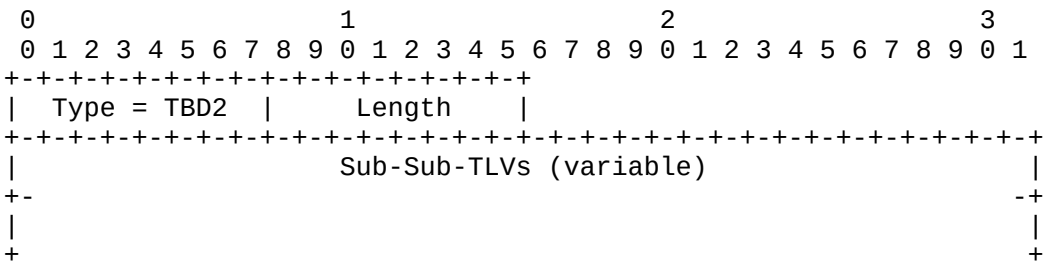

Figure 4: IP Pause Capability Sub-TLV

where:

Type: TBD2.

Length: Variable, in octets, depending on the sub-sub-TLVs.

The only supported sub-sub-TLV is the Buffer Size Sub-Sub-TLV. The Buffer Size advertised in the Buffer Size Sub-Sub-TLV represents the supported maximum IP flows' buffering space. Only a single Buffer Size Sub-Sub-TLV MAY be advertised in the IP Pause Capability Sub-TLV. If more than one Buffer Size Sub-Sub-TLV is present, all the Buffer Size Sub-Sub-TLVs MUST be ignored. The Buffer Size Sub-Sub-TLV is structured as shown in Figure 5.

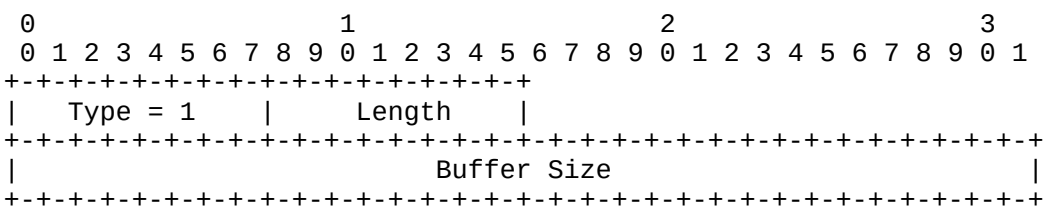

Figure 5: Buffer Size Sub-Sub-TLV

where:

Type: 1.

Length: This field MUST be set to 4.

Buffer Size: This field indicates the maximum IP flows' buffering space supported by the advertising node. The unit for this field is KB (Kilo Bytes).

5.2. Advertising IP Pause Capability Using OSPF

The PE and P nodes advertise their support of buffering IP flows by advertising a new IPC TLV of the OSPF Router Information (RI) Opaque Link State Advertisement (LSA) [RFC7770]. This TLV is applicable to both OSPFv2 and OSPFv3. This TLV SHOULD only be advertised once in the RI Opaque LSA. This TLV SHOULD be advertised WAN domain wide. The IP Pause Capability TLV is structured as shown in Figure 6.

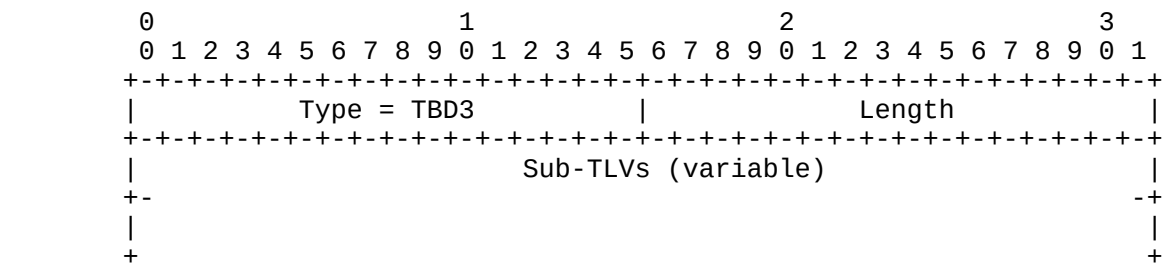

Figure 6: IP Pause Capability TLV

where:

Type: TBD3.

Length: Variable, in octets, depending on the sub-TLVs.

The only supported sub-TLV is the Buffer Size Sub-TLV. The Buffer Size advertised in the Buffer Size Sub-TLV represents the supported maximum IP flows' buffering space. Only a single Buffer Size Sub-TLV MAY be advertised in the IP Pause Capability TLV. If more than one Buffer Size Sub-TLV is present, all the Buffer Size Sub-TLVs MUST be ignored. The Buffer Size Sub-TLV is structured as shown in Figure 7.

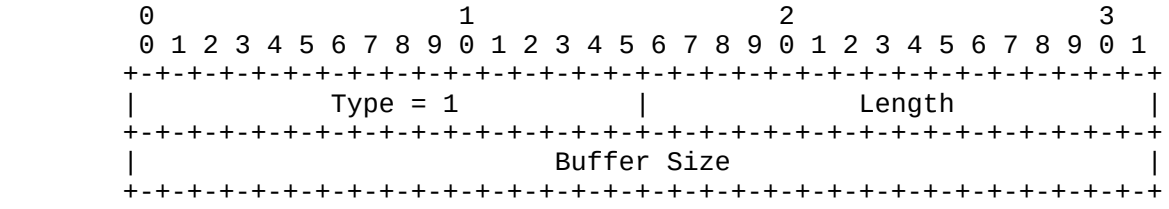

Figure 7: Buffer Size Sub-TLV

where:

Type: 1.

Length: This field MUST be set to 4.

Buffer Size: This field indicates the maximum IP flows' buffering space supported by the advertising node. The unit for this field is KB (Kilo Bytes).

6. Security Considerations

The congestion notification for pause from PE node receiving PFC frames to P/PE nodes MUST be applied in a specific controlled domain. A limited administrative domain provides the network administrator with the means to select, monitor, and control the access to the network, making it a trusted domain.

To avoid potential Denial-of-Service (DoS) attacks, it is RECOMMENDED that implementations apply rate-limiting policies when generating and receiving congestion notification for pause messages.

A deployment MUST ensure that border-filtering drops inbound congestion notification for pause message from outside of the domain and that drops outbound congestion notification for pause message leaving the domain.

A deployment MUST support the configuration option to enable or disable the congestion notification for pause feature defined in this document. By default, the congestion notification for pause feature MUST be disabled.

7. IANA Considerations

7.1. A well-known UDP Port

This document requests the following allocations from IANA:

A well-known UDP port number TBD1 from the System Ports range of the "Service Name and Transport Protocol Port Number" registry [RFC6335] is requested to be assigned to the Congestion Notification for Pause Message. Specifically, IANA is requested to assign a UDP port as shown below for which the Assignee and Contact is the IESG and the IETF Chair, respectively.

Service Name	Port Number	Transport Protocol	Description	Reference
Congestion Notification for Pause	TBD1	udp	Receiver Port for Congestion Notification for Pause	Section 4 of THIS_DOCUMENT

Table 1: Service Name and Transport Protocol Port Number Registry

7.2. IS-IS IP Pause Capability Sub-TLV

This document requests IANA to make the following registration in the "IS-IS Sub-TLVs for IS-IS Router CAPABILITY TLV" registry:

Value	Description	Reference
TBD2	IP Pause Capability	This document

Table 2: New Sub-TLV in IS-IS Sub-TLVs for IS-IS Router CAPABILITY TLV Registry

7.3. IS-IS Sub-Sub-TLVs for the IP Pause Capability Sub-TLV Registry

IANA is requested to create the "IS-IS Sub-Sub-TLVs for IP Pause Capability Sub-TLV" registry under the "IS-IS TLV Codepoints" grouping for the assignment of sub-TLV types for the IP Pause Capability sub-TLV specified in this document. This registry defines sub-sub-TLVs for the IP Pause Capability sub-TLV (TBD2) advertised in the IS-IS Router CAPABILITY TLV (242).

The registration procedure is "Expert Review", as defined in [RFC8126]. Guidance for the designated experts is provided in [RFC7370]. The Buffer Size sub-sub-TLV is defined by this document, and the initial contents of the registry are as follows:

Value	Description	Reference
0	Reserved	This document
1	Buffer Size	This document
2-255	Unassigned	

Table 3: IS-IS Sub-Sub-TLVs for IP Pause Capability Sub-TLV Registry

7.4. OSPF IP Pause Capability TLV

This document requests IANA to make the following registration in the "OSPF Router Information (RI) TLVs" registry:

Value	Description	Reference
TBD3	IP Pause Capability	This document

Table 4: New TLV in OSPF Router Information (RI) TLVs Registry

7.5. OSPF IP Pause Parameter Sub-TLVs Registry

IANA is requested to create the "OSPF IP Pause Parameter Sub-TLVs" registry under the "Open Shortest Path First (OSPF) Parameters" grouping. This registry defines sub-TLVs for the IP Pause Capability TLV (TBD3).

The registration procedures are that the values in the range 1-34999 are to be allocated using the "Standards Action" registration procedure defined in [RFC8126], and the values in the range 35000-65499 are to be allocated using the "First Come First Served" registration procedure. The Buffer Size sub-TLV is defined by this document, and the initial contents of the registry are as follows:

Value	Description	Reference
0	Reserved	This document
1	Buffer Size	This document
2-65499	Unassigned	
65500-65534	Experimental	This document
65535	Reserved	This document

Table 5: OSPF IP Pause Parameter Sub-TLVs Registry

8. Acknowledgements

The authors would like to acknowledge Xiangyang Zhu and Yao Liu for the very helpful discussion.

9. References

9.1. 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>.

[RFC6335] Cotton, M., Eggert, L., Touch, J., Westerlund, M., and S. Cheshire, "Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry", BCP 165, RFC 6335, DOI 10.17487/RFC6335, August 2011, <https://www.rfc-editor.org/info/rfc6335>.

[RFC7370] Ginsberg, L., "Updates to the IS-IS TLV Codepoints Registry", RFC 7370, DOI 10.17487/RFC7370, September 2014, <https://www.rfc-editor.org/info/rfc7370>.

[RFC768] Postel, J., "User Datagram Protocol", STD 6, RFC 768, DOI 10.17487/RFC0768, August 1980, <https://www.rfc-editor.org/info/rfc768>.

[RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <https://www.rfc-editor.org/info/rfc7770>.

[RFC7981] Ginsberg, L., Previdi, S., and M. Chen, "IS-IS Extensions for Advertising Router Information", RFC 7981, DOI 10.17487/RFC7981, October 2016, <https://www.rfc-editor.org/info/rfc7981>.

[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>.

9.2. Informative References

[IEEE8021Q-2022]
IEEE, "IEEE Standard for Local and Metropolitan Area Networks--Bridges and Bridged Networks", DOI 10.1109/IEEESTD.2022.10004498, IEEE Std 802.1Q-2022, December 2022, <https://ieeexplore.ieee.org/document/10004498>.

[RFC2764] Gleeson, B., Lin, A., Heinanen, J., Armitage, G., and A. Malis, "A Framework for IP Based Virtual Private Networks", RFC 2764, DOI 10.17487/RFC2764, February 2000, <https://www.rfc-editor.org/info/rfc2764>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

Authors' Addresses

Xiao Min
ZTE Corp.
Nanjing
China
Phone: +86 18061680168
Email: xiao.min2@zte.com.cn

Kan Zhang
China Mobile
Beijing
China
Email: zhangkan@chinamobile.com