SPRING
Internet-Draft
Intended status: Standards Track
Expires: 31 August 2026

Z. Ruan, Ed.
China Unicom
Y. Liu, Ed.
ZTE
M. Han, Ed.
Z. Han
Y. Liu
China Unicom

27 February 2026

SRv6 behavior extention for Flow Control in WAN
draft-ruan-spring-priority-flow-control-sid-03

Abstract

In the wake of the continuous emergence of novel AI technologies, including collaborative training, distributed inference and the like, the scenario in which different data centers communicate via the RDMA (Remote Direct Memory Access) protocol has emerged as a new requirement. similar to data center networks, wide area networks (WANs) should also be equipped with certain flow control capabilities, nodes in the wide area network (WAN) need to quickly and accurately notify upstream traffic nodes of their congestion status. Upon receiving the congestion notification, upstream nodes can take responsive actions to mitigate network congestion. By extending SRv6[RFC 8986] technology, specific services can be carried over SRv6 policies, which enables the precise delivery and response of congestion signals.

For traffic carried over SRv6 policies, the service traffic path is explicitly orchestrated via the Segment List in the SRH header. By parsing the sequence of the Segment List, downstream nodes or controllers can accurately identify the upstream nodes and interfaces corresponding to the current traffic, thus enabling the rapid notification of congestion information to the upstream. Since each SID in the Segment List is a 128-bit IPv6 address that represents a specific behaviour (e.g., End, End.X), extending existing behaviours to mark their capability to receive and process congestion signals allows flexible control over the delivery mode of congestion signals in the WAN.

The specific packet formats and contents of congestion signals vary in various ways and are continuously being updated. This document takes the most commonly used PFC backpressure notification mechanism as an example to elaborate on the overall workflow of WAN congestion information notification and the extensions to the SRv6 protocol. PFC technology is widely deployed in RoCEv2 networks within data

centers to report congestion signals. By leveraging SRv6 and NRP technologies, it is possible to effectively control the propagation range and path of PFC backpressure signals in wide area networks, thereby avoiding issues such as deadlocks and the excessive propagation of congestion scope.

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.

Copyright Notice

Table of Contents

1. Introduction

With the continuous emergence of new AI technologies such as collaborative training and distributed inference, the scenario in which different data centers communicate via the RDMA (Remote Direct Memory Access) protocol has emerged as a new requirement.

Given the high sensitivity of RDMA technology to packet loss, PFC technology is widely deployed in RoCEv2 networks within data centers. The working flow of PFC is shown in Figure 1, which is mainly composed of the following steps:

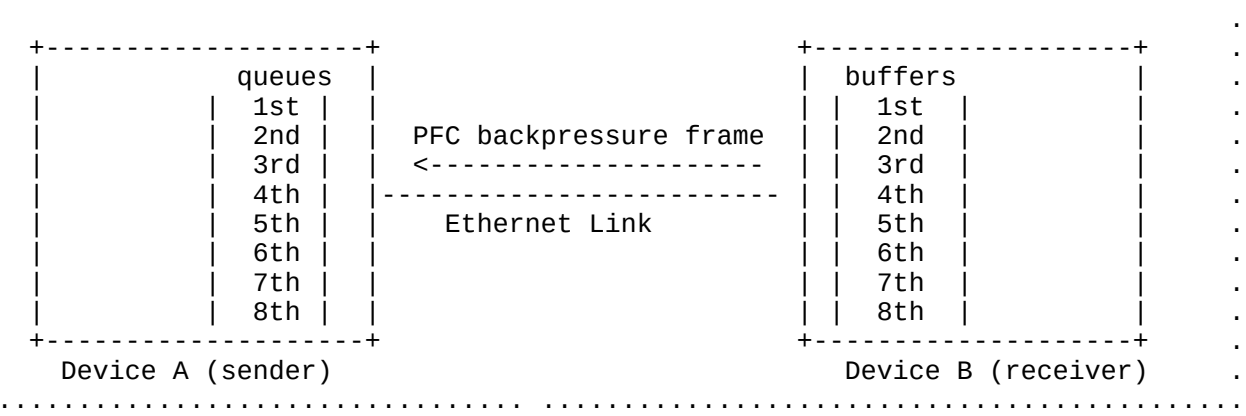

Figure 1: Diagram of PFC Working Mechanism

a. Devices supporting PFC have multiple priority queues on the transmit interface, and the receive interface has an equal number of receive buffers.

b.When a receive buffer on a downstream device (such as Device B) becomes congested, that is, the queue buffer is consumed quickly and exceeds a certain threshold (such as 1/2 or 3/4 of the port queue buffer), the corresponding mechanism will be triggered.

c.Device B detects congestion sends a back-pressure signal "STOP" to the upstream device (Device A) in the data-entry direction.

d.After receiving the back-pressure signal, the upstream device (Device A) stops sending the packets of the corresponding priority queue according to the signal indication and stores the data in the local interface buffer. If the consumption of the local interface buffer of Device A also exceeds the threshold, it will continue to apply back-pressure to the upstream.

e.When the congestion situation of the receive buffer is alleviated, that is, the used buffer of the queue is reduced below the PFC threshold, the receiving device (device B) will send a PFC back-pressure stop message to the upstream to notify the upstream device to send packets again and resume the traffic transmission of the corresponding priority queue.

1.1. Requirements

In the scenario of cross-data center communication, back-pressure frames may need to be propagated across wide area networks. The transmission conditions of wide area networks are much more complicated than those of data center networks, and thus will face some constraints.

a. Tenant-Granular Back-pressure In the wide area network scenario, a physical link may carry the services of multiple tenants simultaneously.In order to avoid the mutual influence of traffic among different tenants,back-pressure signaling should support tenant-level granularity,this can be achieved by leveraging the technology of SRv6[RFC 8986]and Enhanced VPN[draft-ietf-spring-sr-for-enhanced-vpn-10] .

b. Legacy Device Constraints There is a wide variety of devices in the wide area network, and many of them do not support congestion notification,upgrading all the equipment is uneconomical and difficult to implement.As a result, in many scenarios, backpressure packets cannot be transmitted hop by hop as in the data center network.Therefore, a more flexible method for conveying congestion signals is needed.

Take the Figure 2 diagram as an example.The direction of traffic is R1 -> R2 -> R3-> R4-> R5. Among them, R1 and R5 support the generation and processing of back-pressure signals, while the device R2,R3,R4 does not support it. When congestion occurs on the interface between R5 and DC2, if the back-pressure signal can be transmitted to the corresponding interface of R1 in a timely manner, then it can be ensured that there will be no packet loss in the traffic. Therefore, a mechanism is needed that enables R5 to perceive the device and interface among the upstream devices of the current traffic that support the processing of back-pressure signals.

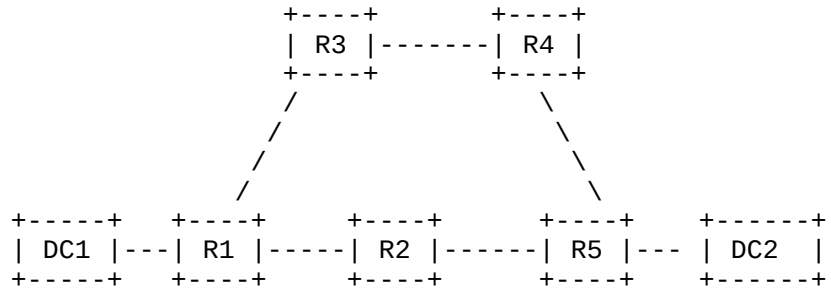

Figure 2: Topo for Cross-DC WAN network

1.2. Conventions and Definitions

1.3. 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. End.X.PFC behavior

The "Endpoint with L3 cross-connect and Priority-based Flow Control" behavior (abbreviated as "End.X.PFC") is a variant of the End.X behavior defined in [RFC8986]. Based on the original End.X behavior, it incorporates additional meanings to facilitate the identification of interfaces in the network that possess the capability to handle PFC packets. Its main use is to identify a PFC-capable interface within the wide area network. By advertising this information across the network, it enables other devices and controllers in the network to implement traffic control strategies more easily.

The specific behavior of a device upon receiving an End.X.PFC packet can be divided into two scenarios:

case1:When the device acts as an intermediate node in the SRv6 path and the End.X.PFC is the non-final hop SID in the Segment list, the End.X.PFC behaviour exhibits the same forwarding characteristics as the End.X behaviour, that is, forwarding the traffic to the next hop through the specified L3 adjacency interface.

case2:When the device acts as the final destination node of the SRv6 tunnel for PFC backpressure signals and the End.X.PFC is the final hop SID in the Segment list, the End.X.PFC behaviour parses the inner payload of the encapsulated packet, extracts the PFC backpressure frame, and executes the corresponding PFC flow control action according to the content of the frame.

When N receives a packet destined to S and S is a local End.X.PFC behavior, N does the following:

```
 S01.  When an SRH is processed {
  S02.  If (Segments Left == 0) {
  S03.  Stop processing the SRH, and proceed to process the next
           header in the packet, whose type is identified by
           the Next Header field in the routing header.
  S04.  }
  S05.  If (IPv6 Hop Limit <= 1) {
  S06.     Send an ICMP Time Exceeded message to the Source Address
            with Code 0 (Hop limit exceeded in transit),
            interrupt packet processing, and discard the packet.
  S07.  }
  S08.  max_LE =(Hdr Ext Len / 2) - 1
  S09.  If ((Last Entry > max_LE) or (Segments Left > Last Entry+1)) {
  S10.      Send an ICMP Parameter Problem to the Source Address
            with Code 0 (Erroneous header field encountered) and
            pointer set to the Segments Left field, interrupt
            packet processing,and discard the packet.
  S11.  }
  S12.  Decrement IPv6 Hop Limit by 1
  S13.  Decrement Segments Left by 1
  S14.  Update IPv6 DA with Segment List[Segments Left]
  S15.  Submit the packet to the IPv6 module for transmission to the new
        destination via interface J
  S16. }
```

When processing the Upper-Layer header of a packet matching a FIB entry locally instantiated as an End.X.PFC SID, N does the following:

```
S01.  If (Upper-Layer header type == 143(Ethernet) ) {
S02.    Remove the outer IPv6 header with all its extension headers
S03.     If(Destination MAC==01-80-C2-00-00-01)
S04.       Interface J will perform flow control actions based on the
           content in the Priority - Flow Control (PFC) frames.
S05. } Else {
S06.  Process as per Section 4.1.1 defined in [RFC8986]
S07. }
```

3. Using End.X.PFC Behavior For Congestion Notification

In the topology shown in Figure 3, it is assumed that the edge devices R1 and R3 of the wide area network support the processing of PFC (Priority-based Flow Control) frames, while R2 does not. R1 and R3 can configure the behavior of End.X.PFC locally and advertise it by IGP and BGP-LS protocols.

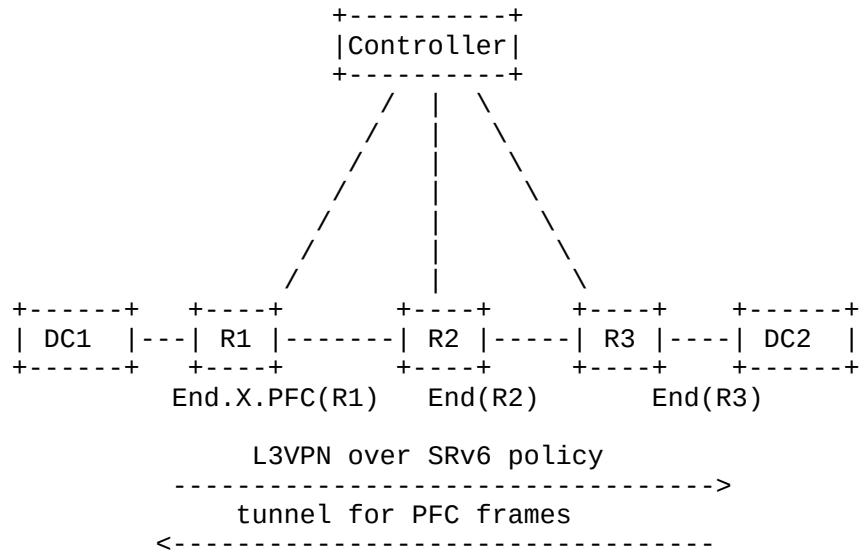

Figure 3: Using the End.X.PFC Behaviour for flow control

Assume that a tenant needs to transmit data over RDMA protocol from DC1 to DC2. The operator can deploy an L3VPN over SRv6 Policy service in the wide area network to carry this traffic.

When congestion occurs at interface between R3 and DC2, or when R3 receives a PFC back-pressure signal from DC2 and its buffer exceeds the set threshold, R3 needs to propagate a back-pressure signal upstream. There are two scenarios for how R3 can accurately send the back-pressure signal to the corresponding interface on R1:

Scenario 1: Controller-Pre-deployed Tunnels for PFC Back-pressure Frames The controller first calculates an SRv6 policy {R1.End.X.PFC,R2.End,R3.End} between R1 and R3 ,After the SRv6 Policy is provisioned and programmed, the controller verifies the nodes involved in the Policy, and retrieves that R1 and R3 are devices with PFC capability.The controller can identify the upstream-downstream relationship between R1 and R3 in the segment list, then pre-deploys an SRv6 tunnel from R3 to R1 with a segment list of {R3.End , R2.End,

R1.End.X.PFC } . When R3 generates a PFC back-pressure frame, the frame is encapsulated into this tunnel. Upon reaching the final hop R1, R1 processes the End.X.PFC behavior. Similarly, the controller pre-deploys a reverse tunnel from R1 to R3 for carrying PFC frames.

To improve the orchestration efficiency of the controller, similar to End.X.PFC, a new End.PFC behavior(eg.R3.End.PFC) can be defined to identify nodes with PFC capability.By using this behavior, the nodes along the path that require the creation of reverse tunnels can be more easily identified by the controller.

Scenario 2: Device-Auto-triggered Tunnel Creation for PFC Back-pressure Frames When R3(PFC-capable node) receives the first data packet of an SRv6 policy, it analyzes the segment list in the SRH header (e.g., {R1.End.X.PFC , R2.End, R3.End }). Upon detecting an End.X.PFC behavior in the upstream path, R3 dynamically creates a reverse tunnel with a segment list of {R3.End ,R2.End, R1.End.X.PFC } to carry PFC frames.

A reverse tunnel can be based on SRv6 Policy or an SRv6 BE. In the SRv6 BE scenario, it is only required to encapsulate R1.End.X.PFC into the destination IP of the packet.

It should be noted that, in order to avoid packet loss on devices that do not support the PFC (Priority-based Flow Control) functionality, network slicing technology [RFC9543] can be utilized. Tenant-level slices can be deployed on the interfaces traversed by SRv6 to provide independent queues and bandwidth resources.If slicing technology is used, the information of the reverse tunnel should also include the corresponding slicing information, such as the slice ID, etc

4. OAM Considerations

It is necessary to perform network quality probing between the two devices that need to exchange backpressure signals. As illustrated in Figure 3, the downstream node is required to conduct real-time probing of the reachability and network status of the relevant behaviours on the upstream node. This capability enables the downstream node to dynamically adjust its own watermarks and the size of reserved buffers.

5. Security Considerations

The security considerations of SRv6 in RFC8754 [RFC8986] apply to this document.

6. IANA Considerations

This document defines a new SRv6 Endpoint behavior called END.X.PFC.

IANA is requested to allocate four new code points from the "SRv6 Endpoint Behaviors" sub-registry in the "Segment-routing with IPv6 data plane (SRv6) Parameters" registry:

Value	Hex	Endpoint Behavior	Reference
TBA	TBA	End.X.PFC	[This ID]
TBA	TBA	End.X.PFC with PSP	[This ID]
TBA	TBA	End.X.PFC with USP	[This ID]
TBA	TBA	End.X.PFC with PSP & USP	[This ID]
TBA	TBA	End.X.PFC with USD	[This ID]

7. References

7.1. 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>.

[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>.

[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>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

7.2. Informative References

[I-D.ietf-spring-sr-for-enhanced-vpn] Dong, J., Miyasaka, T., Zhu, Y., Qin, F., and Z. Li, "Segment Routing based Network Resource Partition (NRP) for Enhanced VPN", Work in Progress, Internet-Draft,

draft-ietf-spring-sr-for-enhanced-vpn-10, 15 December 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-spring-sr-for-enhanced-vpn-10>.

[RFC9543] Farrel, A., Ed., Drake, J., Ed., Rokui, R., Homma, S., Makhijani, K., Contreras, L., and J. Tantsura, "A Framework for Network Slices in Networks Built from IETF Technologies", RFC 9543, DOI 10.17487/RFC9543, March 2024, <https://www.rfc-editor.org/info/rfc9543>.

Authors' Addresses

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

Yao Liu (editor)
ZTE
Nanjing
China
Email: liu.yao71@zte.com.cn

MengYao Han (editor)
China Unicom
Beijing
China
Email: hanmy12@chinaunicom.cn

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

Ying Liu
China Unicom
Beijing
China
Email: liuy619@chinaunicom.cn