SPRING
Internet-Draft
Intended status: Standards Track
Expires: 15 December 2025

Z. Ruan, Ed.
M. Han, Ed.
Z. Han
Y. Liu
China Unicom
17 Oct 2025

Priority-based Flow Control SID in SRv6
draft-ruan-spring-priority-flow-control-sid-02

Abstract

To address the issue of lossless transmission for cross-data center services, the PFC (Priority-based Flow Control) mechanism can be extended to wide-area networks (WANs) to solve packet loss caused by network congestion and the problem of backpressure signal transmission between WAN and RoCEv2 network in data centers.

By leveraging SRv6 and slicing 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.

PFC is a hop-by-hop mechanism. Given that most current WAN devices do not support PFC function, this document proposes defining a new type of SRv6 SID End.X.PFC to indicate the PFC-capable interface in the WAN network.

Table of Contents

1. Introduction

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.

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:

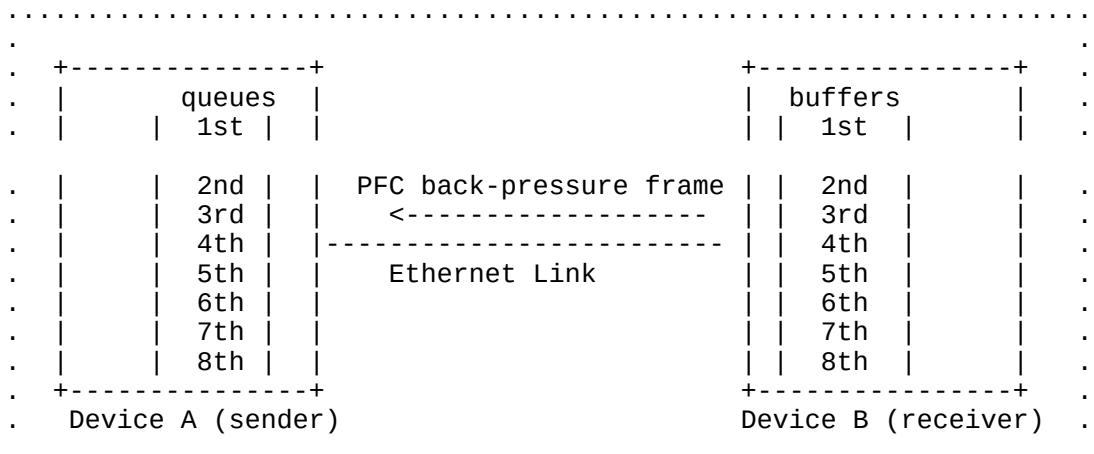

Figure 1: Diagram of PFC Working Mechanism

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 constrains.

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 network slicing.

b. Legacy Device Constraints Currently, most WAN routers do not support the processing of back-pressure signals except some new versions of routers. Take the Figure 2 diagram as an example.The direction of traffic is R1 -> R2-> R5. Among them,the two devices 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 closest to itself among the upstream devices of the current traffic that support the processing of back-pressure signals.

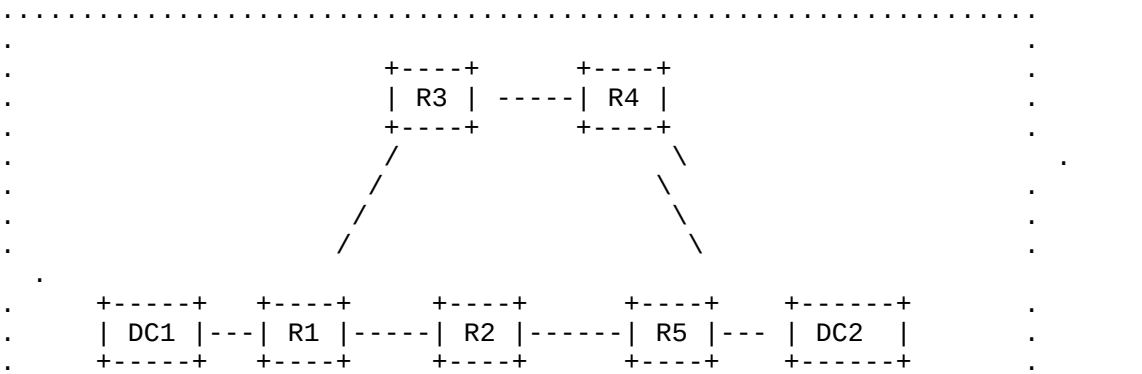

Figure 2: Topo for Cross-DC WAN network

1.2. Conventions and Definitions

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. Abbreviations and definitions used in this document: WAN Wide Area Network SRv6 Segment Routing over IPv6 [RFC8986] SID Segment Identifier ROCEv2 RDMA over Converged Ethernet version 2 [IBTA-ROCEv2] PFC Priority-based Flow Control RDMA Remote Direct Memory Access DC Data Center

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 SID for Flow Control in WAN network

This section defines a new SRv6 SID named End.X.PFC, and provides specific descriptions regarding the functions and behaviors of this SID.

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 SID, 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.

Any SID instance of this behavior is associated with a set, J, of one or more L3 adjacencies.

When N receives a packet destined to S and S is a local End.X.PFC SID, 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. }
```

Note that End.X.PFC SIDs should be announced using IGP or BGP-LS in a similar way to the announcement of End.X SIDs, while they need to be distinguished from the End.X SID by both the network nodes and the

network controller. The detailed protocol extension will be described in a separate document.

3. Using Flow Control SID

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 generate the SID of End.X.PFC and advertise it by IGP and BGP-LS protocols.

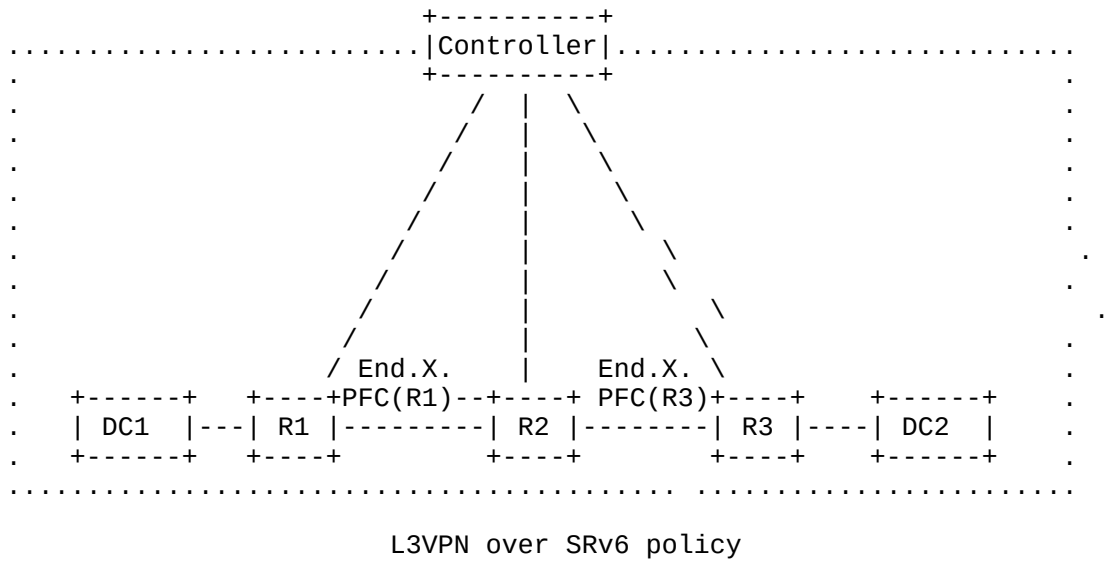

Figure 3:End.X.PFC Working Flow In WAN Network

Assume that a tenant needs to read data via RDMA protocol from DC2 to DC1. 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 calculates and distributes an SRv6 Policy for the inter-DC service. The segment list is {R1.End.X.PFC, R2.End, R3.End.X.PFC},Then the controller identifies the upstream-downstream relationship between R1 and R3 in the path. It 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 action. Similarly, the controller pre-deploys a reverse tunnel from R1 to R3 for carrying PFC frames.

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.X}). Upon detecting an End.X.PFC type SID in the upstream path, R3 dynamically creates a reverse tunnel with a segment list of {R3.End.X, R2.End, R1.End.X.PFC} to carry PFC frames.

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. Security Considerations

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

5. OAM Considerations

TBD

The End.X.PFC SID should support response to some OAM packets, such as ping,tracert.etc

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]

7. Normative References

[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>.

[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>.

[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>.

Contributors

Authors' Addresses

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