RTGWG
Internet-Draft
Intended status: Standards Track
Expires: 14 April 2026

Y. Liu
ZTE Corporation

11 October 2025

Proactive Flow Control Point Detection in WAN
draft-liu-rtgwg-wan-flowctrl-detect-00

Abstract

This document proposes a proactive detection mechanism for flow control in WAN, letting the congested node to know precisely which upstream point should the flow control message be sent to.

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 14 April 2026.

Table of Contents

1. Introduction

With the growth of intelligent computing services, scenarios such as disaggregated computing and real-time inference require the lossless transmission of large volumes of bursty traffic over wide-area networks (WANs). To achieve lossless data transmission over WAN, there're quite a few recent works aiming to deploy flow control mechanisms in WAN to avoid packet loss in case of congestion, e.g, [I-D.ruan-spring-priority-flow-control-sid] discusses how to deploy PFC(Priority-based Flow Control, [IEEE802.1Qbb]) in WAN based on SRv6 data plane, and [I-D.liu-rtgwg-srv6-cc] proposes the method of precise/fine-grained flow control to achieve flow control at the network slice [RFC9543] level.

To conclude, to deploy flow control mechanism in WAN, the node facing congestion needs to generate a flow control message and sends it to the upstream point which is able to perform the flow control action (e.g, stop sending the corresponding traffic or reducing the sending rate).

The flow control message sending mechanism may include one of the follows:

* Multicast. Although standard PFC propagates congestion information via Ethernet multicast frames, multicast-based mechanism is not preferred in WAN since it cannot accurately reach upstream nodes, potentially leading to incorrect flow suppression and impacting unrelated services.

* Centralized configuration. The controller, with the awareness of all the node and the path information in the network, can configure the information of the upstream flow control point on each node that may generate the flow control message. But this methods will bring extra burden for the controller in large scale networks.

* Distributed decision. The congested node decides the upstream node itself. The difficulty is that the congested node needs to be aware of the necessary information to make the proper decision.

Based on the above considerations, this document proposes a proactive detection mechanism for flow control in WAN, letting the congested node to know precisely which upstream point should the flow control message be sent to.

The detailed flow control mechanism itself is out of the scope of this document.

2. Terminology

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. Flow Control in WAN

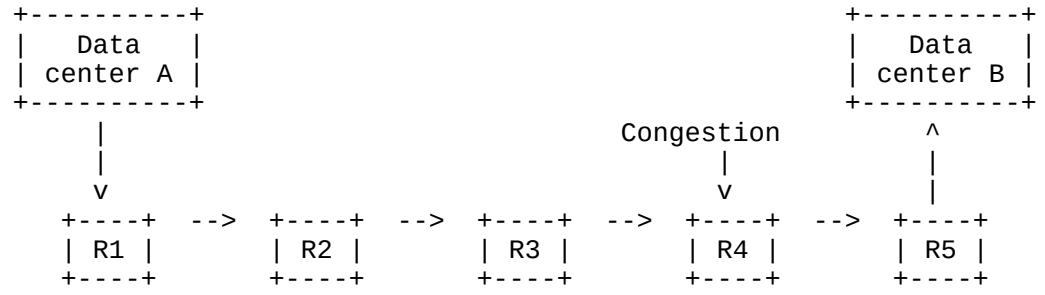


Figure 1

As shown in Figure 1, data center A and data center B are connected via a path(R1-R2-R3-R4-R5) in WAN.

R1,R2,R4 and R5 are able to perform the function of flow control, and R3 is a legacy device which doesn’t support any flow control technology.

When congestion occurs at R4, R4 generates a flow control message (e.g, the PFC pause frame defined in [IEEE802.1Qbb], the congestion notification message defined in [I-D.liu-rtgwg-srv6-cc]) to the nearest upstream stream node that is able to perform flow control, i.e, node R2.

R2 receives the notification and performs the flow control based on the content of the notification message and local capacity. If R2 cannot handle the congestion, a flow control message is forwarded further upstream to R1.

4. Proactive Flow Control Point Detection Method

The basic concept of the proactive flow control point detection method in this document is to send a flow control detection packet along the packet forwarding path.

When receiving the flow control detection packet, the node that is capable of flow control updates the packet with its own information(e.g, the interface address or the corresponding SRv6 ajacency SID of the interface), so the detection packet will always carry the information of the nearest upstream node that’s capable of flow control and the node receiving the detection packet would store this information and use it as the destination of the flow control message when congestion occurs.

4.1. Packet Format

The following information is required in the flow control detection packet:

* Upstream flow control point identifier: indicate the nearest upstream point(interface of the node) that is able to perform flow control for the corresponding traffic flow

* Flow control object identifier: used in the scenario of precise flow control to provide the extra information of flow control object, e.g, if the flow control is at the network slice level, the network slice ID is the flow control object identifier

* Path identifier: used to identify the path of the traffic flow when necessary.

A new Hop-by-Hop option (Section 4.3 of [RFC8200]) type is defined in this document to carry the fields above for flow control point detection. Its format is shown in Figure 2.

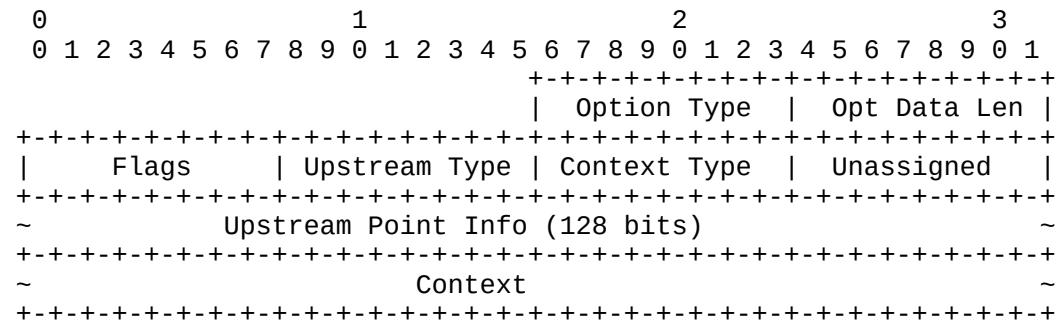
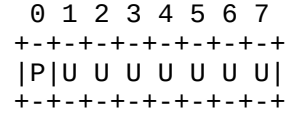


Figure 2

Option Type: 8-bit identifier of the type of option. The type of Flow Control Detection option is TBA.

Opt Data Len: 8-bit unsigned integer indicates the length of the option Data field of this option, in octets.

Flags: 8-bit flags field. The most significant bit is defined in this document.



* P (PFC): The P flag is used to indicate whether this flow control detection packet is used for PFC.

Upstream Type: indicates the type of the upstream point that is able perform flow control for the traffic. When set to 1, the upstream Point Info carries a 128-bit IPv6 interface address, when set to 2, the upstream Point Info carries a 128-bit SRv6 adj-SID.

Upstream Point Info: 128-bit field carrying the corresponding upstream point information based on the value of the Upstream Type.

Context Type: The Context Type field is an 8-bit bitmap that specifies which contexts are included in the Context field of the packet. Each bit in this field corresponds to a specific context. When a bit is set to 1, it indicates the presence of the corresponding parameter, where,

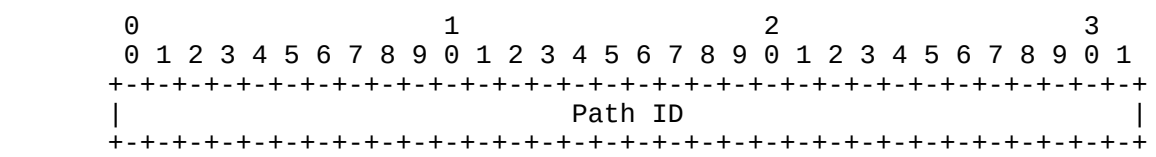
* bit 0: indicates the presence of the path identifier field when set, the format of the path identifier field is shown as in section 4.1.1

* bit 1: indicates the presence of the flow control object identifier when set,the format of the flow control object is shown as in section 4.1.2.

The packet fields defined above can be carried in-band or out-band as long as the packet is forwarded along the same path of the normal traffic flow

4.1.1. bit 0 Context

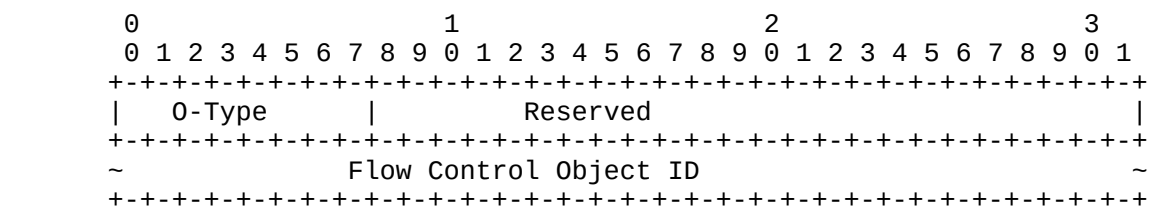
When bit0 of Context Type is 1, the following context is included to indicate the identifier of the path:



Path ID: used to identify a path in the network. The scope of the Path ID is implementation specific.

4.1.2. bit 1 Context

When bit1 of Context Type is 1, the following context is included to carry the identifier of the flow control object when precise flow control mechanism is used:



O-Type: 8-bit field, indicates the type of Flow Control Object ID. When the value of the O-Type is 1, it indicates that Flow Control Object ID carries a 32-bit network slice ID.

4.2. Processing Procedures

As in Figure 1, the traffic of network slice 1 is forwarded along an SRv6 path, the corresponding SID list is <SID-R12,SID-R23,SID-R45>, whereas SID-R12 is the adjacency SID of R1 for the adjacency between R1 and R2, SID-R23 and SID-R45 are also the corresponding adj-SID on R2 and R4.

And precise flow control is enabled in the network to control the congestion at the network slice level.

R1,R2,R4 and R5 are able to perform flow control at the network slice level, and R3 is a legacy device which doesn’t support any flow control technology.

To detect the flow control point along the path of slice 1, R1 sends a flow control detection packet in band with the traffic of slice 1, since R1 is capable of flow control, R1 puts SID-R12 into the Upstream Point Info and the Flow Control Object ID field is set to slice-ID 1.

When R2 receives the packet, it stores the mapping between slice 1 and SID-R12, and updates the Flow Control Object ID with its own information, i.e, SID-R23.

Since R3 doesn’t recognize the flow control detection packet, it just forwards the packet based on the SID-list and slice-ID of the packet.

When R4 receives the packet, it stores the mapping between slice 1 and SID-R23, and updates the Flow Control Object ID with its own information, i.e, SID-R45.

When R5 receives the packet, it stores the mapping between slice 1 and SID-R45, and since R5 is the endpoint, it stops processing the packet further.

When congestion occurs at R4 in slice 1, R4 would generate a flow control message for slice 1, and based on the local information, R4 finds the information of upstream flow control point, i.e, SID-R23, and uses it as the destination of the flow control message.

When R2 receives the flow control message with DA set as local adj-SID SID-R23, R2 perform the flow control for slice 1 on the port related with SID-R23 based on the context of the flow control message.

If R2 is not able to control the congestion and generates a flow control message further, R2 would send the message with DA set to SID-R12 based on the local information.

5. Security Considerations

The security considerations with IPv6 Hop-by-Hop Options header are described in [RFC8200],, [RFC7045][RFC9098], [RFC9099], [RFC9673]. This document introduces a new IPv6 Hop-by-Hop option which is either processed in the fast path or ignored by network nodes, thus it does not introduce additional security issues.

6. IANA Considerations

This document requests IANA to assign a new option type from "Destination Options and Hop-by-Hop Options" registry [IANA-HBH].

Hex Value	Binary Value act	chg	rest	Description	Reference
TBA	00	0	tba	Flow Control Detection Option	[this document]

7. References

7.1. Normative References

[IANA-HBH] IANA, "Internet Protocol Version 6 (IPv6) Parameters", <https://www.iana.org/assignments/ipv6-parameters/ipv6-parameters.xhtml>.

[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>.

[RFC7045] Carpenter, B. and S. Jiang, "Transmission and Processing of IPv6 Extension Headers", RFC 7045, DOI 10.17487/RFC7045, December 2013, <https://www.rfc-editor.org/info/rfc7045>.

[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>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

[RFC9098] Gont, F., Hilliard, N., Doering, G., Kumari, W., Huston, G., and W. Liu, "Operational Implications of IPv6 Packets with Extension Headers", RFC 9098, DOI 10.17487/RFC9098, September 2021, <https://www.rfc-editor.org/info/rfc9098>.

[RFC9099] Vyncke, ., Chittimaneni, K., Kaeo, M., and E. Rey, "Operational Security Considerations for IPv6 Networks", RFC 9099, DOI 10.17487/RFC9099, August 2021, <https://www.rfc-editor.org/info/rfc9099>.

[RFC9673] Hinden, R. and G. Fairhurst, "IPv6 Hop-by-Hop Options Processing Procedures", RFC 9673, DOI 10.17487/RFC9673, October 2024, <https://www.rfc-editor.org/info/rfc9673>.

7.2. Informative References

[I-D.liu-rtgwg-srv6-cc] Liu, Y., Peng, S., Lin, C., and X. Min, "Congestion Control Based on SRv6 Path", Work in Progress, Internet-Draft, draft-liu-rtgwg-srv6-cc-00, 9 October 2025, <https://datatracker.ietf.org/doc/html/draft-liu-rtgwg-srv6-cc-00>.

[I-D.ruan-spring-priority-flow-control-sid] Ruan, Z., Han, M., Zhengxin, H., and Ying, "Priority-based Flow Control SID in SRv6", Work in Progress, Internet-Draft, draft-ruan-spring-priority-flow-control-sid-01, 27 June 2025, <https://datatracker.ietf.org/doc/html/draft-ruan-spring-priority-flow-control-sid-01>.

[IEEE802.1Qbb] IEEE, "IEEE Standard for Local and metropolitan area networks--Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks--Amendment 17: Priority-based Flow Control", DOI 10.1109/IEEESTD.2011.6032693, IEEE Std 802.1Qbb-2011, September 2011, <https://standards.ieee.org/ieee/802.1Qbb/4361.html>.

[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>.

Author’s Address

Yao Liu
ZTE Corporation
China
Email: liu.yao71@zte.com.cn