SPRING
Internet-Draft
Intended status: Standards Track
Expires: 8 January 2026

Y. Liu
China Mobile
S. Peng
Huawei
C. Lin
New H3C Technologies

7 July 2025

Congestion Control Based on SRv6 Path
draft-liu-spring-srv6-cc-01

Abstract

This document describes a congestion control solution based on SRv6. It defines mechanisms for congestion notification and flow control within an SRv6-based network, optimizing congestion handling through hierarchical congestion control messages along SRv6 paths.

Table of Contents

1. Introduction

The SRv6 network needs a reliable and efficient mechanism for handling congestion across different segments. Current congestion control techniques lack the ability to handle congestion in a fine-grained, per-path manner. This draft proposes a solution that uses SRv6 path segments and slicing to notify upstream nodes and take actions to reduce congestion. The key idea is to notify upstream nodes about congestion and enable flow control based on SRv6 segments (SID lists). This process is integrated with the SRv6 network's slicing capabilities to provide fine-grained control over network traffic, ensuring lossless transmission of data across SRv6 network.

1.1. 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. Background and Motivation

Priority Flow Control (PFC) provides hop-by-hop, priority-based traffic control. Compared to the traditional Ethernet Pause mechanism, PFC enables more precise flow management by creating multiple virtual channels on a link, each of which can be paused or resumed independently, ensuring that traffic of different priorities does not interfere with one another.

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. In interconnected wide-area networks (WANs), when network congestion occurs, the congestion status must be quickly propagated upstream to both head-end devices and edge devices, enabling hop-by-hop reduction of sending rates. These intelligent computing WANs typically use SRv6 Policies for transport. However, once traffic enters a policy, traditional PFC mechanisms face the following three major challenges:

1. Imprecise Congestion Notification: PFC propagates congestion information via Ethernet multicast frames. In WANs with complex topologies, multicast-based congestion signaling cannot accurately reach upstream SRv6 nodes, potentially leading to incorrect flow suppression and impacting unrelated services.

2. Long Path Latency: WAN paths are long and have significant latency. If congestion signals must be sent all the way back to the data center or edge devices, it results in prolonged traffic degradation. Therefore, upstream control on the SRv6 path is needed to respond promptly.

3. Control Overhead at the Head Node: A single head node in the WAN may manage numerous SRv6 paths. If all congestion messages are sent back to the head node, it could become a processing bottleneck. Performing distributed traffic control at intermediate nodes along the SRv6 path can alleviate the burden on the head node.

3. SRv6 congestion notification Mechanism

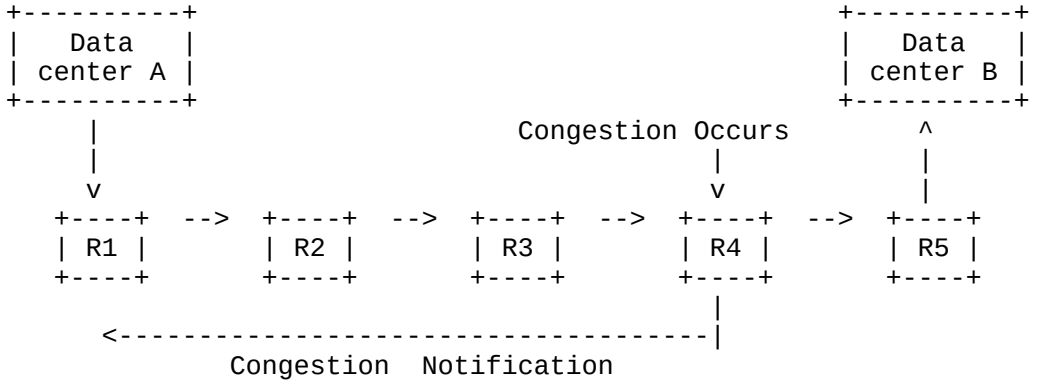

Figure 1: Congestion Notification in SRv6 Network

Consider two data centers, A and B, connected via an SRv6 path defined as R1 -> R2 -> R3 -> R4 -> R5, as shown in Figure 1. The process follows these steps:

1. The head node R1 encapsulates the SID list (SRv6 path) containing R2 -> R3 -> R4 -> R5. It may optionally carry an SRv6 path segment (PSID) and starts forwarding the data. The source address is R1, and the destination address is the SID of R2.

2. Intermediate nodes (R2, R3, R4) forward data according to the SID list, with each node checking its local SID table for forwarding and slice-related information.

3. When a node, such as R4, faces congestion(such as queue overload), it sends a congestion notification message to the previous node in the SID list (R3), including congestion-related information such as SID list, PSID, and slice ID.

4. R3 receives the notification and adjusts the forwarding rate based on local capacity. If R3 cannot handle the congestion, the notification is forwarded further upstream to R2 and so on.

5. If no node can manage the congestion, the head node R1 adjusts the path load balancing or selects an alternate path to mitigate the congestion.

4. Congestion Notification Message Format

The congestion notification message can be encapsulated in either ICMPv6 [RFC4443] or UDP [RFC768] messages. Regardless of the encapsulation format, they contains following fields:

1. *Checksum*: Used for error-checking the packet.

2. *Flags*: Contains special flags. The first bit is S bit, indicating the presence of a Slice ID. Other bit is not defined.

3. *SRv6 Source Address*: The source address of the SRv6 message.

4. *Segment Routing Header (SRH)*: This is the routing header for SRv6, which defines the path the packet should take. The specific details of SRv6 path segments are described in [RFC8754].

5. *Slice ID*: The identifier for the slice experiencing congestion.

6. *Flow Queue Size*: The size of the congestion queue at the node.

7. *Flow Utilized Ratio*: The utilization ratio for the flow, showing how much of the available queue is being used.

8. *Behavior TLV*: Variable-length fields that define the actions of user queues to take in response to congestion, such as reducing or pausing traffic. See Section 4.3 for the specific format.

4.1. ICMPv6 message format

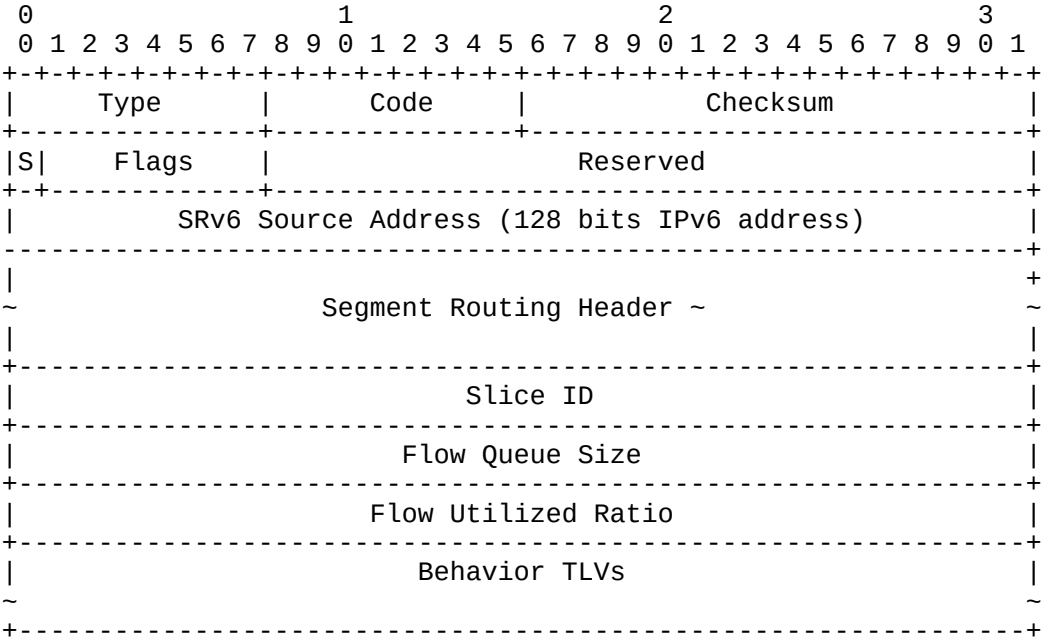

Figure 2: Congestion Notification in ICMPv6

Where:

Type and *Code*: These fields indicate the specific congestion notification type and its sub-type, providing details about the kind of congestion event being reported.

4.2. UDP packet

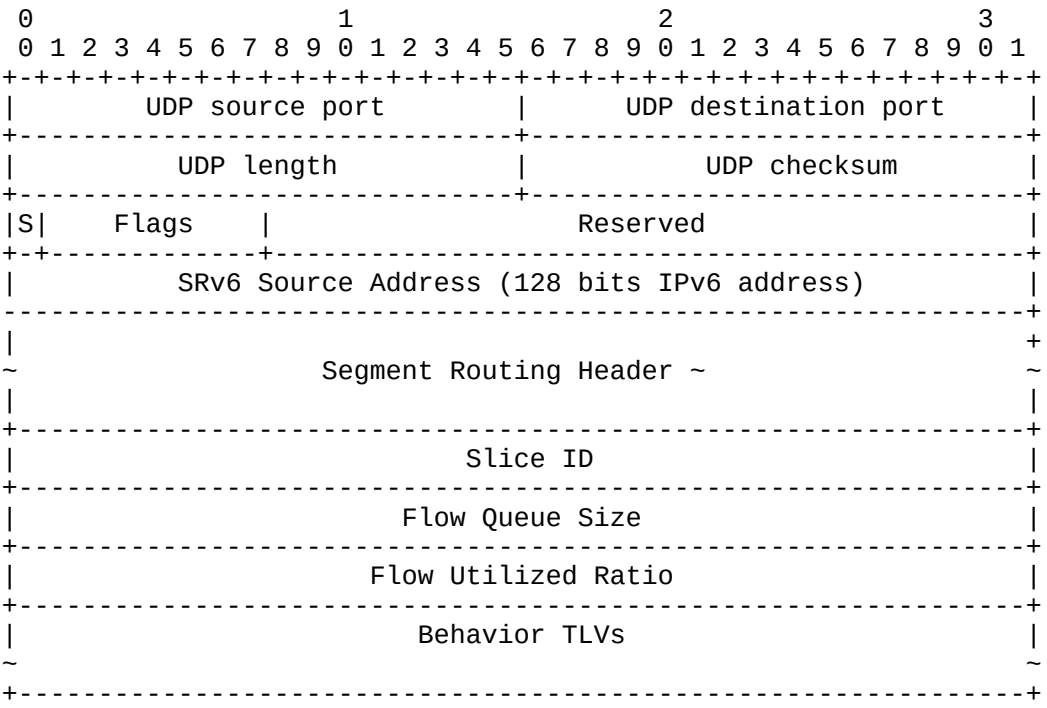

Figure 3: Congestion Notification in UDP

Where:

UDP Destination port: A new port indicates the congestion notification packet.

4.3. Behavior TLV

The format for the behavior TLV is defined as follows:

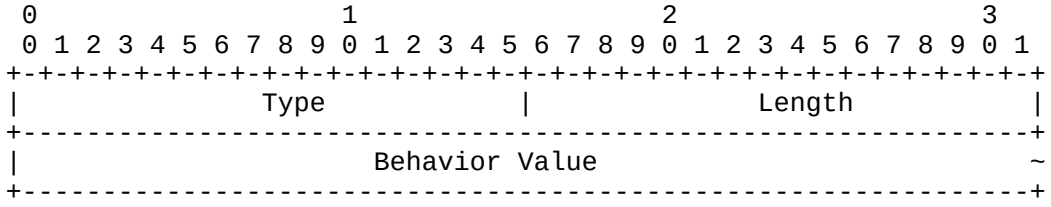

Figure 4: Behavior TLV

Where:

Type: 2 octet value that indicates the behavior type.

Length: 2 octet value that indicates the length of the following Behavior Value.

Behavior Value: Variable-length value that indicates the action of user queue to take in response to congestion.

This document defines two behaviors, namely Type 1 and Type 2. The specific formats for type 1 and type 2 are as follows.

The format for type 1 behavior TLV:

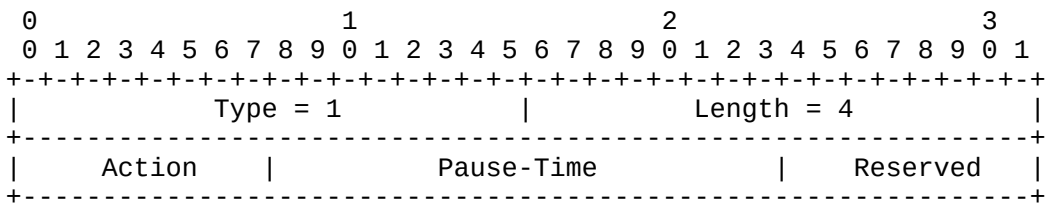

Figure 5: Type 1 Behavior TLV

Where:

Type = 1: indicates the action for all user queues.

Action: 1 octet value that indicates the action of one user queue to take in response to congestion. The first two bits represent specific actions, 00 for no action, 01 for pausing traffic, and 10 for reducing traffic. The last six bits represent the relative number for reducing traffic, and the reference value of the relative number is defined by the user according to specific scenarios.

Pause-Time: 2 octet value that indicates the time for the corresponding action to be execute, in units of microseconds. When the action execution exceeds the Pause-Time, traffic MUST be immediately recovered.

Reserved: Reserved Field.

The format for type 2 behavior TLV:

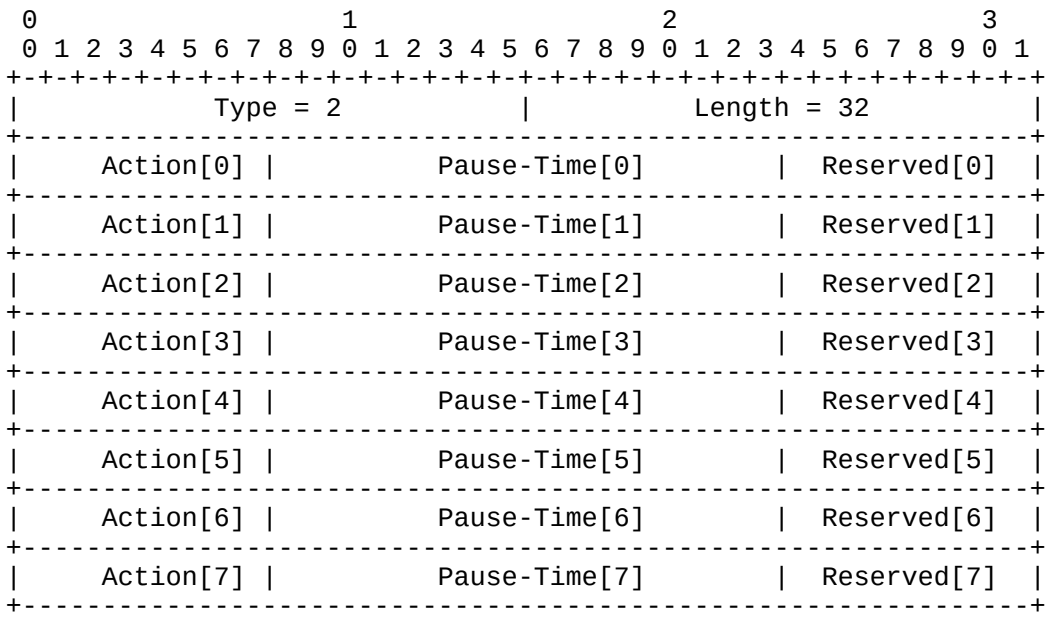

Figure 6: Type 2 Behavior TLV

Where:

Type = 2: indicates the action for the 8 user queues.

5. SRv6 congestion notification running process

The SID configuration of each node in the figure is as follows: End.X SIDs of nodes R1 to R5 are A::1:1,A::2:1,A::3:1,A::4:1,A::5:1, and the slice ID corresponding to each SID is 1. The VPN SID of the R5 node is A::5:F.

The running process of each node is as follows:

1. The data packet sent by R1 is encapsulated with the SRv6 Policy. The SID list is {A::2:1, A::3:1, A::4:1, A::5:F}. The source address is A::1::, and the destination address is A::2:1.

2. Packets are forwarded based on the SID list of SRv6 paths. The destination address of the packet forwarded by R4 is replaced with A::5:F. However, the queue of slice 1 on the outbound interface of R4 is congested. The queue size of slice 1 is 200 Mb. When the current usage exceeds the threshold(75%), the congestion message needs to be advertised to the previous hop of the SRv6 path.

3. R4 encapsulates the ICMPv6 packet with the source address A::4:1 and the destination address A::3:1 into the ICMPv6 packet. The ICMPv6 packet carries the source address A::1:: of the original SRv6 packet and the entire SRH header, and encapsulates the congestion information into the ICMPv6 packet. The following information includes the slice ID of 1, the size of the congested queue is 200 Mb, the usage is 75%, and the recommended congestion control behavior, for example, the sending rate is reduced by 30%.

4. R3 receives the congestion notification packet, checks that the destination address is local, and searches, according to the SID list in the SRH, for the forwarding queue or slice corresponding to the local SID, or for the forwarding queue or slice corresponding to the PSID. In this example, the information about the slice 1 already carried does not need to be searched again. Reduce the local sending rate based on the situation that the 200 Mb congestion queue is 75% occupied and the buffer capacity of the local queue. In this example, reduce the sending rate by 30%.

5. If R3 cannot perform congestion control after receiving the congestion notification packet, for example, the local queue buffer capacity is insufficient, R3 continues to send the congestion notification packet to the previous hop of the SRv6 path. The packet is encapsulated in an ICMPv6 packet and the source address remains to be R4’ A::4:1. The destination address is changed to A::2:1. In addition, other information carried in the packet remains unchanged, but the congestion control action can be adjusted. As the path for the congestion notification becomes longer, the congestion may deteriorate. For example, the sending rate can be reduced by at least 50%.

6. Security Considerations

This document does not introduce any new security considerations.

7. IANA Considerations

This document requests IANA to allocate a new ICMP message type and UDP port.

8. 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/rfc/rfc8754>.

[RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <https://www.rfc-editor.org/rfc/rfc4443>.

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

[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/rfc/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/rfc/rfc8174>.

Authors' Addresses

Yisong Liu
China Mobile
Beijing
China
Email: liuyisong@chinamobile.com

Shuping Peng
Huawei
Beijing
China

Email: pengshuping@huawei.com

Changwang Lin
New H3C Technologies
Beijing
China
Email: linchangwang.04414@h3c.com