IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: 20 September 2026

S. Zhuang
H. Wang
Huawei

19 March 2026

BGP Flow Specification Extensions for Network Congestion Management
draft-zhuang-idr-flowspec-extension-for-ncm-00

Abstract

BGP Flow Specification (FlowSpec) [RFC8955] and [RFC8956] has been proposed to distribute traffic filter policy (traffic filters and actions) via BGP [RFC4271]. Multiple applications have used BGP FlowSpec to distribute traffic filter policy. These applications include the following: mitigation of denial of service (DoS), enabling traffic filtering in BGP/MPLS VPNs, centralized traffic control of router firewall functions, and SFC traffic insertion.

Due to its powerful extensibility, FlowSpec can be easily used for network congestion management. This document describes how to use BGP FlowSpec to implement network congestion management.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD 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.

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 20 September 2026.

Copyright Notice

Copyright (c) 2026 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction

Explicit Congestion Notification (ECN) is an extension to the Internet Protocol (IP) and the Transmission Control Protocol (TCP), defined in RFC 3168. ECN allows for end-to-end notification of congestion control to avoid packet loss. ECN is an optional feature that may be used by two endpoints if the underlying network infrastructure supports it.

Typically, a TCP/IP network indicates a channel congestion by dropping packets. In the case of successful ECN negotiation, an ECN-aware router can set a flag in the IP header instead of dropping the packet to indicate that congestion is about to occur. The receiver of the packet then responds to the sender by reducing its transmission rate.

The ECN (Explicit Congestion Notification) field in an IP header (bits 6 and 7 of the Traffic Class/ToS byte) allows routers to signal network congestion by marking packets instead of dropping them. Defined in RFC 3168, it uses four values (00-11) to indicate ECN capability and congestion status, enabling end-to-end congestion control while reducing packet loss and retransmissions.

ECN Field Codepoints (Bits 6-7)

* 00 (Non-ECT): Non-ECN-Capable Transport. The packet does not support ECN.

* 01 or 10 (ECT): ECN-Capable Transport (0 or 1). The endpoints support ECN, and the path is not currently congested.

* 11 (CE): Congestion Experienced. The router is experiencing congestion and marks the packet, prompting the receiver to notify the sender to slow down.

When a router’s buffer exceeds a certain threshold, it changes the ECN field from ECT (01/10) to CE (11) rather than dropping the packet.

It occupies the last two bits of the Type of Service (ToS) field in IPv4 or the Traffic Class field in IPv6.

Fast-CNP (Congestion Notification Packet) is a mechanism for Remote Direct Memory Access (RoCEv2) networks that allows switches to directly send congestion notifications to the source node, bypassing the traditional, slower method of waiting for destination server feedback. By reducing latency in congestion signaling, Fast-CNP prevents severe buffer congestion and improves network performance.

This document describes how to use BGP FlowSpec to implement network congestion management.

2. Definitions and Acronyms

* BGP-FS: BGP Flow Specification
* CE: Congestion Experienced
* DPI: Deep Packet Inspection
* ECN: Explicit Congestion Notification
* ECT: ECN Capable Transport
* FlowSpec: Flow Specification
* FSv2: BGP Flow Specification Version 2
* SR: Segment Routing
* SR-MPLS: SR over the MPLS data plane
* SRv6: SR over the IPv6 data plane
* SAFI: Subsequent Address Family Identifier
* SID: Segment Identifier
* SRH: Segment Routing Header
* TE: Traffic Engineering
* USD: Ultimate Segment Decapsulation

3. The Flow Specification Encoding for Network Congestion Management

3.1. Type TBD1 - ECN

Encoding: <type (1 octet), [numeric_op, value]+>

Defines a list of {numeric_op, value} pairs used to match the 2-bit ECN field (see also [RFC3168] and [RFC8200]).

This component uses the Numeric Operator (numeric_op) described in Section 4.2.1.1 of [RFC8955]. The ECN component values MUST be encoded as single octet (numeric_op len=00).

The two least significant bits contain the ECN value. All other bits SHOULD be treated as 0.

An example of an ECN Flow Specification component encoding for: "all packets matching ECN {1 or 2}".

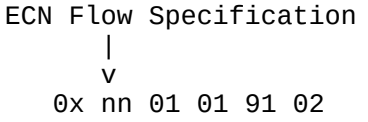

Figure 1: Example of ECN flow specification TLV encoding

Decoded:

Value		
0xnn	Type	Type nn - ECN Flow Specification component's Type
0x01	numeric_op	value size = 1, ==
0x01	value	1, ECN's value = 1
0x91	numeric_op	end-of-list, value size = 1, ==
0x02	value	2, ECN's value = 2

Figure 2: Decoded the example of ECN flow specification TLV encoding

3.2. ECN Marking (ecn-marking) Sub-Type TBD2

The ecn marking Extended Community instructs a system to modify the ECN bits in the IP header (Section 5 of [RFC3168]) of a transiting IP packet to the corresponding value encoded in the 6 least significant bits of the Extended Community value, as shown in Figure XXX.

The Extended Community is encoded as follows:

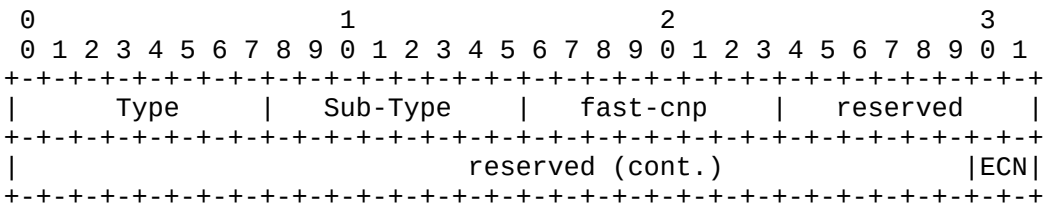

Figure 3: ECN Marking Extended Community Encoding

Type (1 octet): 0x80

Sub-Type (1 octet): TBD2

ECN: new ECN value for the transiting IP packet

reserved (46 bits): MUST be set to 0 on encoding and MUST be ignored during decoding

3.3. Extending the Traffic-Action (traffic-action) field

The traffic-action Extended Community consists of 6 octets of which only the 2 least significant bits of the 6th octet (from left to right) are defined by [RFC8955], as shown in Figure 4.

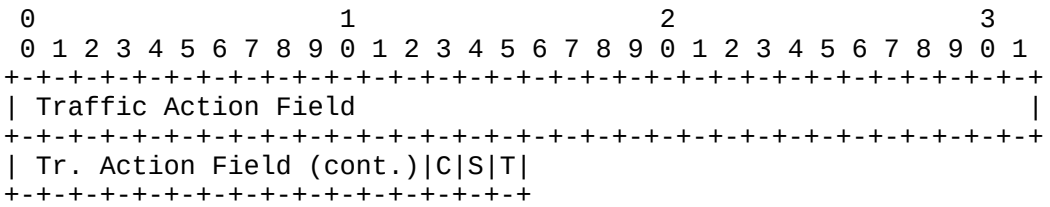

Figure 4: Traffic-Action Extended Community Encoding

The S and T bits are defined in [RFC8955], and this document defines the C bit as follows:

C Fast-CNP (bit 45): Enables the Fast CNP function (only effective when set).

3.4. Fast-CNP Action Extended Community Sub-Type TBD3

The Fast-CNP Action Extended Community instructs a system to send fast congestion notification packets to the source node.

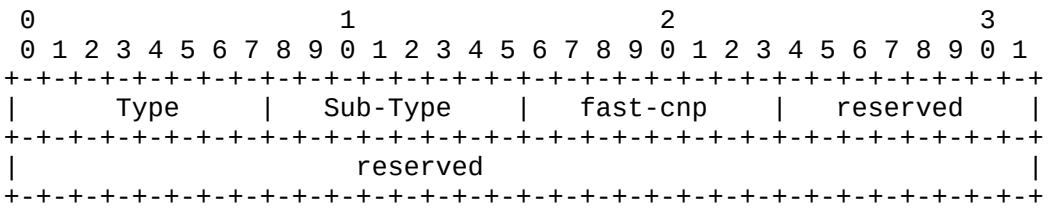

Figure 5: Fast-CNP Action Extended Community Encoding

Type (1 octet): 0x80

Sub-Type (1 octet): TBD3

fast-cnp (1 octet): Enables the Fast CNP function if set to 1 and disables the Fast CNP function if set to 0

reserved (5 octets): MUST be set to 0 on encoding and MUST be ignored during decoding

3.5. Marking Threshold Action Extended Community Sub-Type TBD4

The Marking Threshold Action Extended Community instructs a system to set a threshold for the target traffic.

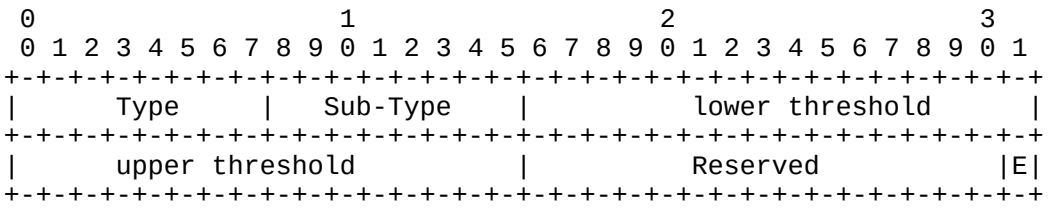

Figure 6: Marking Threshold Action Extended Community

Type (1 octet): 0x80

Sub-Type (1 octet): TBD4

lower threshold (2 octets): Lower threshold of the forwarding queue.

upper threshold (2 octets): Upper threshold of the forwarding queue.

reserved (7 bits): MUST be set to 0 on encoding and MUST be ignored during decoding.

E (1 bit): Indicates whether to mark ECN flag.

4. Use Cases

4.1. Case1: Detects the risk of traffic congestion and proactively sets the ECN flag

A traffic management device on the network analyzes traffic on the network, and finds that there is a risk of packet loss due to congestion in traffic sent from Sender to Receiver through R1, R2,..., and Rn. The management device sends a Flowspec route to the device R1. The Flowspec route carries characteristic information of target traffic by using a Flowspec NLRI, and carries an marking ECN action defined in this draft.

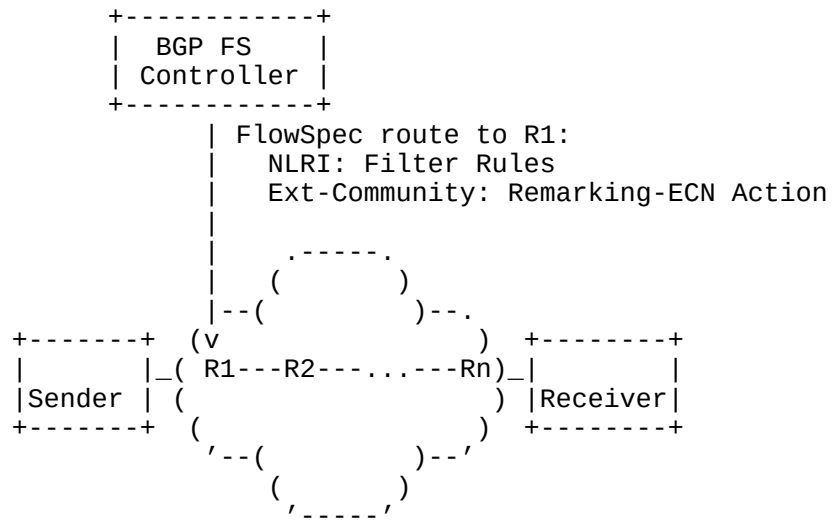

Figure 7: Detects the risk of traffic congestion and proactively sets the ECN flag

When R1 receives a FlowSpec route, it uses the NLRI in the FlowSpec route to match the target traffic and marks the target traffic with the ECN flag according to the instruction of the ECN-Marking action.

The traffic with the ECN flag is forwarded to the next hop. In this way, other devices along the path and the receiver can detect network congestion in a timely manner based on the ECN flag and make corresponding adjustments.

4.2. Case2: Detects traffic congestion and Sends Fast-CNP to Sender

When a traffic management device in the network detects that traffic forwarded from a Sender terminal to a Receiver terminal through the network path R1, R2,..., Rn may encounter a risk of congestion and packet loss, the management device sends a Flowspec route to a device on the network path, for example, R2. The NLRI of the Flowspec route carries the characteristic of the target traffic, and the Flowspec route carries the Fast-CNP action defined in this draft.

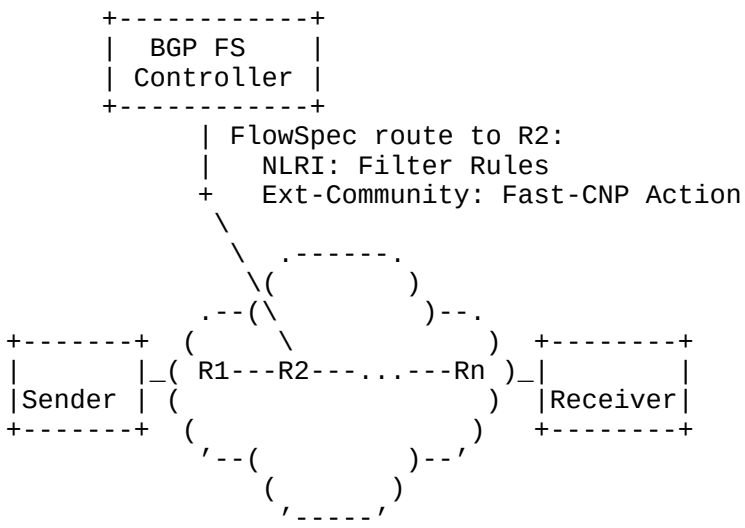

Figure 8: Detects traffic congestion and Sends Fast-CNP to Sender

After receiving the FlowSpec route, R2 uses the NLRI field in the FlowSpec route to match the target traffic. In addition, R2 identifies that the FlowSpec route carries the Fast-CNP action. According to the instruction of this action, R2 obtains the address of the sender terminal from the target traffic, constructs a Fast-CNP packet to be sent from R2 to the sender terminal, and sends the packet to the sender terminal.

After receiving the Fast-CNP packet, the sender terminal reduces the rate at which packets are sent to the receiver terminal.

4.3. Case3: Optimizing the traffic that contains congestion indication information

When the traffic management device in the network detects that traffic forwarding from a Sender terminal to a Receiver terminal through the network path R1, R2,..., Rn encounters congestion, the management device sends a Flowspec route to a device (for example, R1) on the network path. The NLRI of the Flowspec route carries the characteristic of the target traffic, where an ECN matching element is included, and the Flowspec route carries various possible traffic optimization actions.

The traffic optimization action may include redirecting to a next hop, setting an optimized DCSP value, or indicating a forwarding through a path with a higher priority etc,.

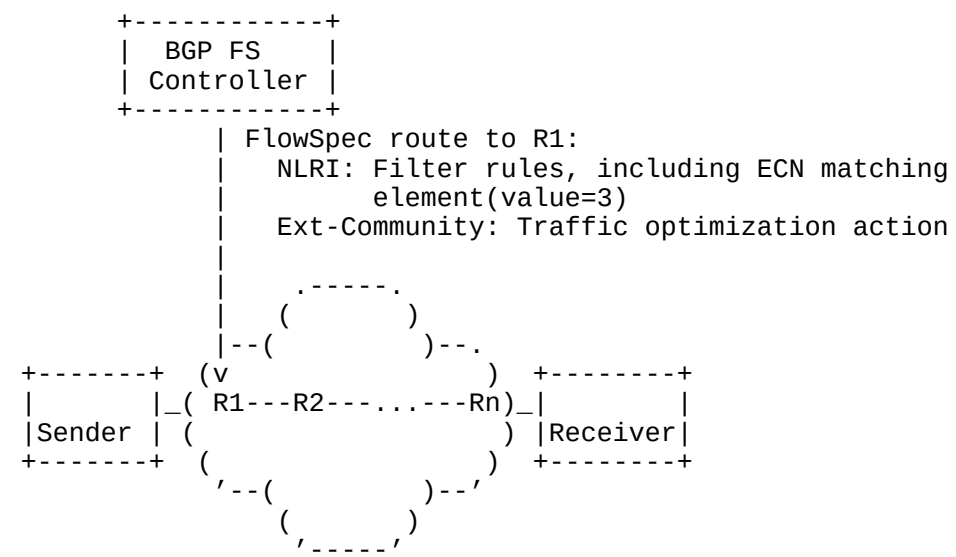

Figure 9: Optimizing the traffic that contains congestion indication information

After receiving the FlowSpec route, R1 matches the target traffic based on the NLRI field in the FlowSpec route. The ECN value in the FlowSpec route must be the same as that in the target traffic. After the target traffic is matched, R1 performs the traffic optimization action specified in the FlowSpec route on the target traffic. The traffic optimization action may include redirecting to a next hop, setting an optimized DCSP value, or indicating a forwarding through a path with a higher priority etc,.

4.4. Case4: Set a threshold for the target traffic and enable the ECN marking function.

When a management device on the network plans to optimize and manage target traffic on a device, the management device delivers a FlowSpec route to the network device (for example, R1 in the following figure). The FlowSpec route encapsulates the characteristics of the target traffic in the NLRI and uses the Marking Threshold Action community attribute defined in this document to set a threshold for the target traffic. In addition, the management device sets the E bit to instruct the network device to mark the target traffic with the ECN flag when the rate of the target traffic is within the threshold range.

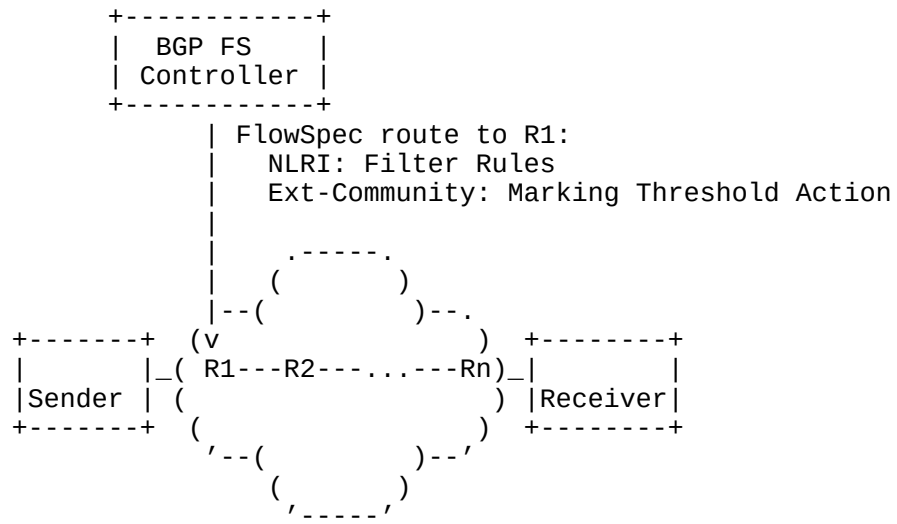

Figure 10: Detects the risk of traffic congestion and proactively sets the ECN flag

When receiving the FlowSpec route, the network device R1 matches the target traffic based on the NLRI field in the FlowSpec route, sets a threshold for the target traffic according to the instruction of the marking-threshold action defined in this document, and enables the function of marking the target traffic with the ECN flag.

The traffic with the ECN flag is forwarded to the next hop. In this way, other devices along the path and the receiver can detect network congestion in a timely manner based on the ECN flag and make corresponding adjustments.

5. IANA Considerations

TBD

6. Security Considerations

The security considerations of BGP [RFC4271] and BGP FlowSpec [RFC8955] [RFC8956] apply to this document.

7. Contributors

The following people made significant contributions to this document:

TBD

8. Acknowledgements

The authors would like to acknowledge the review and inputs from XXX (TBD).

9. References

9.1. Normative References

[I-D.ietf-idr-flowspec-redirect-ip]
Haas, J., Henderickx, W., and A. Simpson, "BGP Flow-Spec Redirect-to-IP Action", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-redirect-ip-06, 2 March 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-flowspec-redirect-ip-06>.

[I-D.ietf-idr-fsv2-ip-basic]
Hares, S., Eastlake, D. E., Dong, J., Yadlapalli, C., Maduschke, S., and J. Haas, "BGP Flow Specification Version 2 - for Basic IP", Work in Progress, Internet-Draft, draft-ietf-idr-fsv2-ip-basic-04, 16 March 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-fsv2-ip-basic-04>.

[I-D.ietf-idr-sr-policy-safi]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., and D. Jain, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-sr-policy-safi-13, 6 February 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-sr-policy-safi-13>.

[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>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, DOI 10.17487/RFC5701, November 2009, <https://www.rfc-editor.org/info/rfc5701>.

[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>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

[RFC8955] Loibl, C., Hares, S., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", RFC 8955, DOI 10.17487/RFC8955, December 2020, <https://www.rfc-editor.org/info/rfc8955>.

[RFC8956] Loibl, C., Ed., Raszuk, R., Ed., and S. Hares, Ed., "Dissemination of Flow Specification Rules for IPv6", RFC 8956, DOI 10.17487/RFC8956, December 2020, <https://www.rfc-editor.org/info/rfc8956>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

[RFC9252] Dawra, G., Ed., Talaulikar, K., Ed., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "BGP Overlay Services Based on Segment Routing over IPv6 (SRv6)", RFC 9252, DOI 10.17487/RFC9252, July 2022, <https://www.rfc-editor.org/info/rfc9252>.

[RFC9256] Filsfils, C., Talaulikar, K., Ed., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", RFC 9256, DOI 10.17487/RFC9256, July 2022, <https://www.rfc-editor.org/info/rfc9256>.

9.2. Informative References

[I-D.ietf-pce-segment-routing-policy-cp] Koldychev, M., Sivabalan, S., Sidor, S., Barth, C., Peng, S., and H. Bidgoli, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing (SR) Policy Candidate Paths", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-policy-cp-27, 4 April 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-segment-routing-policy-cp-27>.

[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

Shunwan Zhuang
Huawei
156 Beiqing Road
Beijing
100095
P.R. China
Email: zhuangshunwan@huawei.com

Haibo Wang
Huawei
156 Beiqing Road
Beijing
100095
P.R. China
Email: rainsword.wang@huawei.com