Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 17 August 2026

Y. Liu
ZTE Corporation
C. Lin
New H3C Technologies

13 February 2026

BGP Flow-Spec Redirect to SR Segment List Action
draft-ll-idr-flowspec-redirect-sidlist-00

Abstract

BGP Flow Specification (Flow-spec) provides a mechanism for distributing traffic filtering and policy-based forwarding rules. Existing works enables traffic steering into an SR Policy. However, in Artificial Intelligence (AI) network scenarios, "elephant flows" may require deterministic forwarding over specific segment lists within an SR Policy candidate path.

This document specifies a new BGP Flow-spec Redirect action that identifies a specific segment list within an SR Policy.

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 17 August 2026.

Copyright Notice

Table of Contents

1. Introduction

BGP Flow Specification (Flow-spec) [RFC8955] [RFC8956] is an extension to BGP that allows for the dissemination of traffic flow specification rules. BGP Flow-spec Version 2 (FSv2) is defined in [I-D.ietf-idr-flowspec-v2].

[RFC9256] details the concepts of Segment Routing (SR) Policy and steering into an SR Policy. An SR Policy consists of one or more candidate paths, each comprising one or more segment lists. As specified in [RFC9256], if a set of segment lists is associated with the active path of the policy, then the steering is per flow and weighted-ECMP (W-ECMP) based according to the relative weight of each segment list.

Traffic generated by AI training and sample transmission exhibits elephant flow characteristics. Compared to ordinary traffic, elephant flows are characterized by high per-flow bandwidth, and long duration. When the traffic carried by an SR Policy contains elephant flows, distributing traffic across segment lists randomly based on weights may cause elephant flows to be assigned to inappropriate segment lists, leading to network load imbalance, e.g., collocation with other large flows on shared links.

Currently BGP-FS supports steering traffic into an SR Policy. [I-D.ietf-idr-ts-flowspec-srv6-policy] enables steering into an SR Policy using the address in the Redirect-to-IP action [I-D.ietf-idr-flowspec-redirect-ip] as the endpoint and the color in the Color Extended Community [RFC9012] as the color of the target SR Policy. [I-D.ietf0-idr-srv6-flowspec-path-redirect] proposes a scheme that steers traffic into an SR Policy through a Binding SID. [I-D.li-idr-flowspec-sr-policy] defines a FSv2 Redirect to SR Policy action directly using Policy Color and endpoint as the identifier.

This document extends BGP-FS to allow traffic to be redirected to a specific segment list within an SR Policy candidate path.

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.

1.2. Terminology

This document uses terminology from [RFC8955], [RFC9256], and [RFC9857].

* SR Policy: Segment Routing Policy, identified by <headend, color, endpoint>.

* Candidate Path: A specific path option for an SR Policy, with an associated preference value.

* Segment List: A specific SID stack that realizes a path; may have an associated weight for load balancing.

* Segment List ID: A 32-bit identifier assigned to a segment list, unique within the scope of a candidate path [RFC9857].

* Elephant Flow: A long-lived, high-bandwidth traffic flow typical in AI/ML training data transfers and storage replication flows.

2. Redirect to Segment List Action

This action instructs the headend to redirect matching traffic into a specific segment list of a specific candidate path of an SR Policy.

This action is encoded as a Flow Specification v2 (FSv2) Action SubTLV carried in a Wide Community container [I-D.ietf-idr-wide-bgp-communities].

2.1. Action SubTLV Format

The Redirect to Segment List action is encoded as a single FSv2 Action SubTLV.

Sub-Type: TBD1 (2 octets)
Length: 2 octets, indicating the total number of octets of the Value field.
Value: Variable, formatted as follows:

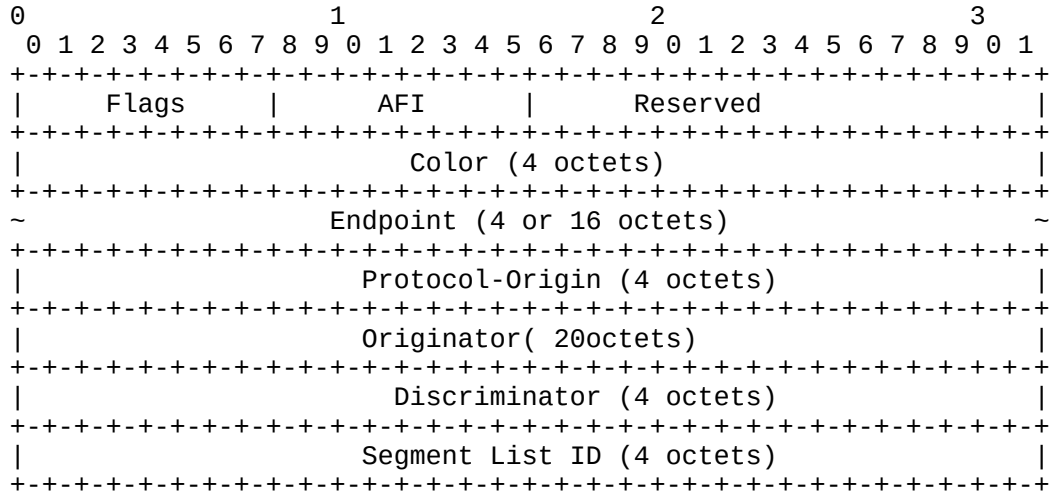

Figure 1: Redirect to Segment List Action SubTLV Format

2.1.1. Flags Field

The Flags field is 1 octet and defined as follows:

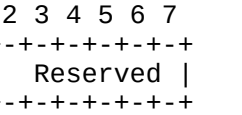

Figure 2

F (Fallback Enabled - 1 bit):
When set (1), the headend MUST perform the fallback procedure described in Section 3.2 if the designated segment list or candidate path becomes invalid or ceases to be the active candidate path. When clear (0), the headend MUST NOT fall back.

R (Recovery - 1 bit):
When set (1), indicates that if the designated segment list later recovers, the headend SHOULD revert to pinning the traffic flow onto it. When clear (0), the headend SHOULD NOT automatically revert.

Reserved (6 bits):
MUST be set to 0 on transmission and ignored on receipt.

2.1.2. SR Policy Identification

Color (4 octets):
The Color value of the target SR Policy, an unsigned non-zero 32-bit integer.

AFI (2 octets):
Address Family Identifier of the Endpoint. Values are taken from IANA's "Address Family Numbers" registry. This field determines the length of the Endpoint field:
- 1 (IPv4): Endpoint = 4 octets
- 2 (IPv6): Endpoint = 16 octets
Other AFI values are reserved and MUST NOT be sent; if received, the entire Action SubTLV MUST be treated as malformed (Section 4).

Endpoint (variable):
The endpoint address of the SR Policy, encoded according to the AFI field.

2.1.3. Candidate Path Identification

Protocol-Origin (4 octets):
The Protocol-Origin value of the candidate path. (See [RFC9256] Section 2.3.)

Originator (20 octets):
The Originator of the candidate path as specified in [RFC9256] Section 2.4.

Discriminator (4 octets):
The Discriminator of the candidate path, as defined in [RFC9256] Section 2.5.

2.1.4. Segment List Identification

Segment List ID (4 octets):
The identifier of the segment list within the candidate path. The value zero is reserved and MUST NOT be used; if received, the Action SubTLV MUST be treated as malformed.

3. Operational Considerations

3.1. Validation and Installation

Upon receipt of a Flow Specification route containing this Action SubTLV, the headend MUST perform the following steps in order:

1. Resolve the SR Policy using <Color, Endpoint>.

2. Locate the candidate path within that SR Policy that exactly matches <Protocol-Origin, Originator, Discriminator>.

3. Verify that the candidate path is the active candidate path of the SR Policy (as in [RFC9256] Section 2.9) .

4. Locate the segment list within that candidate path whose Segment List ID matches the value in the SubTLV.

5. Verify that the segment list is valid as per [RFC9256] Section 5.

If any of these steps fail, the Action SubTLV is considered invalid, and the Flow Specification rule continues to be processed as if this Action SubTLV were not present. An implementation SHOULD log the validation failure.

If all steps succeed, the headend installs a forwarding rule that redirects matching traffic exclusively to the designated segment list. Traffic SHOULD NOT be load-balanced to other segment lists within the same candidate path.

3.2. Fallback Behavior

After successful installation, if the F-Flag is set and any of the following events occur:

* The designated segment list becomes invalid,

* The designated candidate path becomes invalid,

* The designated candidate path is no longer the active candidate path of the SR Policy.

Then the headend MUST revert to ordinary SR Policy forwarding behavior as defined in [RFC9256].

An implementation SHOULD generate an alert to indicate that explicit path pinning is no longer in effect and the reason for fallback.

3.3. Recovery Behavior

If the R-Flag is set and the designated segment list later recovers (becomes valid again) while the designated candidate path remains active, the headend SHOULD revert to pinning the traffic flow onto the recovered segment list.

An implementation MAY apply a configurable revert timer to avoid flapping. During the reversion process, traffic MUST NOT be dropped. The transition MAY be performed via make-before-break.

If the R-Flag is clear, the headend SHOULD NOT automatically revert. A controller MAY withdraw and re-advertise the Flow Specification route to re-establish pinning.

4. Error Handling

A Redirect to SR Policy Segment List SubTLV is considered malformed if:

* The Length field does not match the expected size derived from the AFI and fixed fields.

* The AFI is not 1 (IPv4) or 2 (IPv6).

* The Segment List ID is zero.

* The Originator field does not comply with the encoding rules (e.g., non-zero high-order bits for an IPv4 address).

* The Protocol-Origin field has non-zero high-order 24 bits.

Malformed SubTLVs MUST be handled according to [RFC7606] as treat-as-withdraw. The implementation SHOULD log the error.

5. IANA Considerations

IANA is requested to assign a new Action Sub-Type value from the "BGP FSv2 Action types" registry (established by [I-D.ietf-idr-flowspec-v2]) for the action defined in this document.

Sub-Type	Name	Reference
TBD1	Redirect to SR Policy Segment List	[this document]

6. Security Considerations

The security considerations of [RFC8955], [RFC9256], and [I-D.ietf-idr-flowspec-v2] apply.

The ability to pin traffic to a specific candidate path and segment list introduces additional risk. A malicious controller could direct all elephant flows to a single segment list, causing link overload and congestion collapse. Operators SHOULD:

* Implement per-session policies to restrict which BGP peers are allowed to send BGP-FS routes with this action.

* Limit the rate of BGP-FS updates to protect headend resources.

7. Normative References

[I-D.ietf-idr-flowspec-v2]
Hares, S., Eastlake, D. E., Yadlapalli, C., and S. Maduschke, "BGP Flow Specification Version 2", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-v2-04, 28 April 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-flowspec-v2-04>.

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S., and P. Jakma, "BGP Community Container Attribute", Work in Progress, Internet-Draft, draft-ietf-idr-wide-bgp-communities-12, 17 March 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-wide-bgp-communities-12>.

[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>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[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>.

[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>.

[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>.

[RFC9857] Previdi, S., Talaulikar, K., Ed., Dong, J., Gredler, H., and J. Tantsura, "Advertisement of Segment Routing Policies Using BGP - Link State", RFC 9857, DOI 10.17487/RFC9857, October 2025, <https://www.rfc-editor.org/info/rfc9857>.

8. Informative 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-04, 2 September 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-flowspec-redirect-ip-04>.

[I-D.ietf-idr-ts-flowspec-srv6-policy]
Wenying, J., Liu, Y., Zhuang, S., Mishra, G. S., and S. Chen, "Traffic Steering using BGP FlowSpec with SR Policy", Work in Progress, Internet-Draft, draft-ietf-idr-ts-flowspec-srv6-policy-08, 1 February 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-ts-flowspec-srv6-policy-08>.

[I-D.ietf0-idr-srv6-flowspec-path-redirect]
Van de Velde, G., Patel, K., Li, Z., Zhuang, S., and H. Chen, "Flowspec Indirection-id Redirect for SRv6", Work in Progress, Internet-Draft, draft-ietf0-idr-srv6-flowspec-path-redirect-14, 28 December 2025, <https://datatracker.ietf.org/doc/html/draft-ietf0-idr-srv6-flowspec-path-redirect-14>.

[I-D.li-idr-flowspec-sr-policy]
Li, Z. and liusong, "BGP Flowspec Redirects to SR Policy", Work in Progress, Internet-Draft, draft-li-idr-flowspec-sr-policy-02, 14 October 2025, <https://datatracker.ietf.org/doc/html/draft-li-idr-flowspec-sr-policy-02>.

Authors' Addresses

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

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