IDR Working Group
Internet Draft
Intended status: Standards Track
Expires: September 02, 2026

Y. Liu
China Mobile
C. Lin
New H3C Technologies
J. Li
China Mobile

March 02, 2026

BGP Flow Specification Filtered by Destination-QP
draft-lll-idr-flowspec-filter-qp-00

Abstract

BGP Flowspec mechanism (BGP-FS) [RFC8955] [RFC8956] propagates both traffic Flow Specifications and Traffic Filtering Actions by making use of the BGP NLRI and the BGP Extended Community encoding formats.

This document specifies a new BGP-FS component type named Destination-QP(Destination Queue Pair) to support filtering by Destination-QP.

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 02 September 2026.

Copyright Notice

Table of Contents

1. Introduction

BGP Flowspec mechanism (BGP-FS) [RFC8955] [RFC8956] propagates both traffic Flow Specifications and Traffic Filtering Actions by making use of the BGP NLRI and the BGP Extended Community encoding formats.

In modern AI training clusters, especially those based on RoCEv2 and RDMA for high-performance inter-GPU communication, traffic exhibits distinct characteristics that differ significantly from traditional Internet flows. AI training communication typically consists of long-lived, high-throughput, and delay-sensitive and jitter-sensitive flows, such as those generated by collective communication operations like AllReduce. These flows are often bound to long-term Queue Pair (QP) [IB-SPEC] instances, with relatively stable five-tuple fields, making them prone to path polarization and uneven link utilization when scheduled solely by five-tuple-based hashing. Such static mapping fails to adapt to the dynamic communication patterns and strict performance requirements of AI workloads, leading to localized congestion, degraded training throughput, and reduced cluster efficiency.

For these reasons, five-tuple-only flow scheduling is no longer sufficient for AI-oriented lossless networks. Instead, scheduling mechanisms based on a combination of five-tuple and QP information have become necessary. By introducing QP-level identification into

the flow distribution logic, networks can achieve finer-grained traffic steering, better load balancing across equal-cost multi-path (ECMP) groups, and improved isolation between communication streams.

As shown in Figure 1, the controller uses BGP Flow-Spec to distribute QP routes to the Ingress PE according to QP. Based on the QP value, traffic is redirected to different forwarding paths. Methods for redirecting traffic to different paths can include redirection to IP [draft-ietf-idr-flowspec-redirect-ip] or redirection to SRv6 [draft-ietf0-idr-srv6-flowspec-path-redirect], among others. Specific methods are beyond the scope of this document.

At the forwarding level, when the DC sends packets, it carries the corresponding QP for traffic belonging to the same task.

The Ingress PE looks up the QP routes distributed via BGP Flow-Spec based on the QP of the traffic and forwards the packets according to the QP routes.

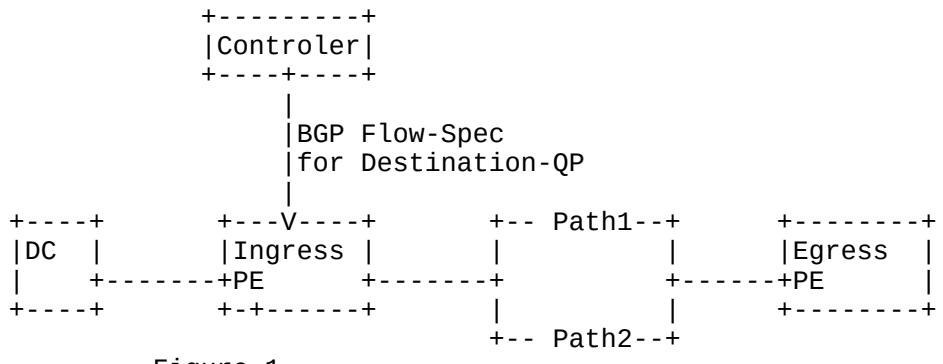

Figure 1

This document specifies a new BGP-FS component type named Destination-QP(Destination Queue Pair) to support filtering by Destination-QP.

2. 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 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Terminology

FS: Flow Specification

QP: Queue Pair

4. FlowSpec Filtered by Destination-QP

[draft-ietf-idr-fsv2-ip-basic] defines the Components in the IP Basic TLV. This document proposes a new Component for Destination-QP information.

The following new component type is defined.

* Destination-QP

Type TBD - Destination-QP

Length: variable

Component Value format: [numeric_op, value]+

Each Destination-QP value is 4 octets.

Per section 10 of [RFC8955] , If a receiving BGP speaker cannot support this new Flow Specification component type, it MUST discard the NLRI value field that contains such unknown components. Since the NLRI field encoding (Section 4 of [RFC8955]) is defined in the form of a 2-tuple <length, NLRI value>, message decoding can skip over the unknown NLRI value and continue with subsequent remaining NLRI.

5. Use case

The BGP agent specifies the traditional 5-tuple and new defined Destination-QP as matching criteria.

As shown in Figure 2, for traffic with a Destination-QP value of 1001, redirect it to Path1; for traffic with a Destination-QP value of 1001, redirect it to Path2.

BGP-FS Route 1:

FS Filters

Destination: 203.0.113.0/24

source address: 198.51.100.0/24

protocol: UDP

destination port: 4791 (RoCEv2 protocol)

source port: 10001

Destination-QP value: 1001 (the newly defined in this document.)

FS Action:
Redirect Flow to Path1 (The specific format is not discussed in this document.)

BGP-FS Route 2:
FS Filters
Destination: 203.0.113.0/24

source address: 198.51.100.0/24

protocol: UDP

destination port: 4791 (RoCEv2 protocol)

source port: 10001

Destination-QP value: 2001 (the newly defined in this document.)

FS Action:
Redirect Flow to Path2 (The specific format is not discussed in this document.)

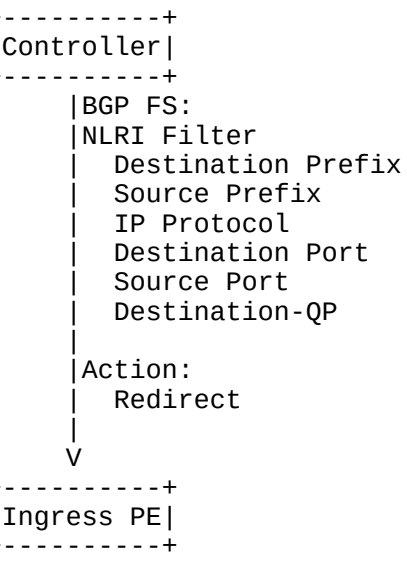

Figure 2

The Ingress PE receives the Flow-Spec route and installs it into the forwarding plane. Upon receiving AI data traffic, it redirects the traffic to the corresponding Path for forwarding based on the traffic 5-tuple and Destination-QP parameters.

In this document, the example uses QP, destination address, and source address as filtering criteria for flow-spec to redirect traffic to different paths.

Destination QP	Dest Prefix	Source Preifx	Redirect
1001	203.0.113.0/24	198.51.100.0/24	Path1
2001	203.0.113.0/24	198.51.100.0/24	Path2

6. Security Considerations

No new security issues are introduced to the BGP protocol by this specification.

7. IANA Considerations

[draft-ietf-idr-flowspec-v2-04] defines the Types for IP Filters.

This document requested to assign a new type code point from "Non-IP Types for IP Filters" registry for Destination-QP.

Non-IP Types for IP Filters SubTLV

type	Definition
64 -	Parts of SID
65 -	MPLS Match 1: Label in Label stack
66 -	MPLS Match 2: EXP bits in top Label
TBD -	Destination-QP This document
67-249	unassigned (reserved for now)
250-	Filter Error handling
251-255	Reserved

8. References

8.1. Normative References

[IB-SPEC]InfiniBand Trade Association. InfiniBand Routing and Forwarding. Architecture Supplement, 2023.

[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://www.ietf.org/archive/id/draft-ietf-idr-flowspec-v2-04.txt>.

[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>.

[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>.

8.2. Informative References

Authors' Addresses

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

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

Jinming Li
China Mobile
China
Email: lijinming@chinamobile.com