Network Working Group
Internet-Draft
Intended status: Informational
Expires: 30 August 2026

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

26 February 2026

QP-based SRv6 Load Balancing Deployment
draft-lll-srv6ops-qp-aware-srv6-lb-00

Abstract

This document describes the use of Segment Routing over IPv6 (SRv6) path selection based on Queue Pair (QP) in Intelligent Computing Wide Area Network (WAN) for Data Center Interconnection (DCI), optimizing load balancing for predictable workloads.

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 30 August 2026.

Copyright Notice

Table of Contents

1. Introduction

The proliferation of RDMA technology in Intelligent Computing Data Center (DC) fabrics has revolutionized high-performance computing, distributed storage, and machine learning workloads.

These workloads generate large, predictable flows that demand ultra-low latency, high bandwidth, and precise congestion control to ensure optimal performance. Traditional networking methods, like hash-based Equal-Cost Multi-Path (ECMP) load balancing, struggle with insufficient entropy due to the low diversity of RDMA (specifically RDMA over Converged Ethernet v2, abbreviated as RoCEv2) [IBTA-SPEC] flow identifiers. This often results in fabric hotspots, network congestion, and performance degradation.

The transmission process of RoCEv2 messages in intelligent computing Wide Area Network (WAN) used for Data Center Interconnection (DCI) is the same as inside the DC, and it will also generate elephant streams, which leads to fabric hotspots, network congestion, and performance degradation.

Segment Routing over IPv6 (SRv6) [RFC8986] provides flexible traffic engineering by supporting policy-based programmability and explicit path steering. SRv6 policy enables deterministic path steering and fine-grained traffic control for RoCEv2 flows, ensuring predictable performance.

This document details SRv6 path selection based on Queue Pair (QP) to optimize load balancing for predictable RoCEv2 flows in intelligent computing WAN by ensuring all packets within a QP follow the same path.

2. Terminology

The following terms are used in this document:

* QP (Queue Pair): A communication endpoint in RDMA architecture, identified by a 24-bit or 32-bit value.

* BTH (Base Transport Header): The RDMA transport header containing QP information.

* SRv6 Policy: An ordered list of segments (SIDs) representing a path through the SRv6 network.

* SL (Segment List): An ordered list of SIDs in an Segment Routing Header (SRH) [RFC8754].

* ECMP (Equal-Cost Multi-Path): A routing technique for load-balancing traffic across multiple best-path routes.

2.1. 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.

3. Problem Statement

Traditional ECMP load balancing faces several challenges with RoCEv2 flow:

* Insufficient Entropy: The relatively static 5-tuple of RoCEv2 flows provides limited entropy, which leads to poor load balancing of elephant flows.

* Persistent Hotspots: When multiple elephant flows hash to the same path, they create persistent hotspots that cannot be resolved without manual intervention or flow termination.

* Poor Failure Convergence: When a link fails, all flows previously using that link are rehashed to remaining paths. This sudden influx of elephant flows can overwhelm the remaining links, causing secondary congestion.

* Lack of Application Awareness: ECMP operates purely on packet header fields without understanding the application-level semantics of QPs.

4. QP-based SRv6 Path Selection

By encoding an ordered list of segments in the packet header, SRv6 (Policy) allows the ingress device to directly steer RoCEv2 workload traffic through the fabric.

FlowSpec, as a traffic scheduling tool, can guide RoCEv2 flows to different SRv6 policies based on their characteristics (such as Dest QP), and forward them along different paths. QP-based FlowSpec protocol extensions are beyond the scope of this document.

QP-to-SRv6 Policy Mapping:

* Upon ingress, the RoCEv2 packet is parsed to extract the destination QP identifier from its Base Transport Header (BTH).

* When multiple SRv6 policies exist, the destination QP is mapped to a corresponding SRv6 Policy via a pre-configured mapping table. The mapping table can rely on local configuration or the flowspec mechanism.

Enhanced Hash-Based Segment List (SL) Scheduling:

* For the selected SRv6 Policy (which may contain multiple SLs), an enhanced hash algorithm, using the QP as a key input, deterministically selects one specific SL for use.

* The chosen SRv6 SL is applied to the RoCEv2 packet, which is then forwarded accordingly.

5. Deployment Illustration

A typical WAN topology for DCI is shown in the figure below.

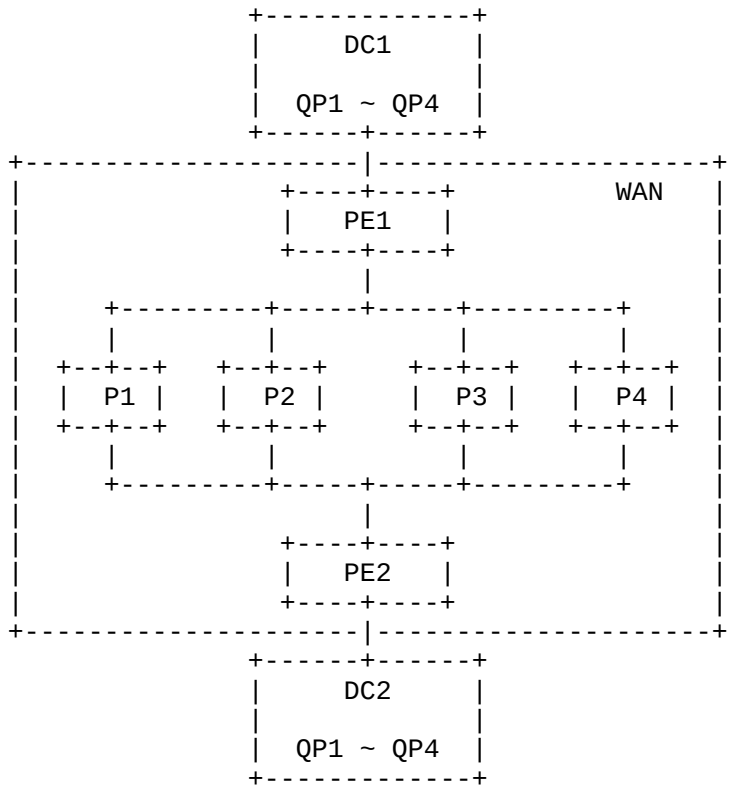

Figure 1: Reference Topology

The topology consists of two Provider Edge (PE) devices, and each of the PEs is connected to four Provider (P) devices and one DC.

In this example, there are 2 DCs, in which four QPs can be established to transmit RoCEv2 workloads.

In the above topology, there are four paths that pass through WAN from DC1 to DC2.

all paths is below:

* *Path1*: PE1 -> P1 -> PE2

* *Path2*: PE1 -> P2 -> PE2

* *Path3*: PE1 -> P3 -> PE2

* *Path4*: PE1 -> P4 -> PE2

5.1. SRv6 Policy Provisioning

During the Day-0 cluster fabric bring-up, the topology is provisioned with SRv6 SIDs on the PE and P devices. These SIDs are statically configured, making them independent of any dynamic routing protocol state.

The PE1 could create two SRv6 Policies with PE2 as the endpoint. Each SRv6 Policy contain two SLs. The following is provisioned:

* SRv6 Policy 1 (low-latency):

  - *SL1*: PE1 -> P1 -> PE2

  - *SL2*: PE1 -> P2 -> PE2

* SRv6 Policy 2 (high-bandwidth):

  - *SL1*: PE1 -> P3 -> PE2

  - *SL2*: PE1 -> P4 -> PE2

5.2. QP-based SRv6 Path Orchestration

The fabric is now orchestrating four AI workloads. During this orchestration, the collective communication among DCs necessitates periodic data transmission from DC1 to DC2.

Between DC1 to DC2, each AI workload is divided into a separate QP, and QPs are QP1, QP2, QP3, and QP4.

During AI job computation, firstly, RoCEv2 packets are redirected to different SRv6 policies based on QP to achieve coarse-grained traffic classification and isolation; secondly, within a single policy, QP is used as a hash key for SL selection, distributing multiple QP flows evenly across multiple candidate paths (SLs) contained in that policy to achieve fine-grained load balancing.

5.2.1. QP-based SRv6 Policy Mapping

Assume that the AI training task traffic carried by each QP has different requirements for link quality. The traffic of QP1 and QP2 requires a low-latency path (Policy 1), while the traffic of QP3 and QP4 requires a high-bandwidth path (Policy 2). On PE1, the QP-to-SRv6 Policy Mapping Table is created as shown below:

QP Range	SRv6 Policy Name
QP1, QP2	Policy 1
QP3, QP4	Policy 2

Table 1: Mapping Table

* During AI job computation, for each AI job, DC1 creates a RoCEv2 packet destined for DC2.

* Based on the destination QP contained in each RoCEv2 packet, each RoCEv2 packet received by PE1 is mapped to its corresponding SRv6 Policy according to the table 1.

5.2.2. QP-based SL Orchestration

In selected SRv6 Policy for each RoCEv2 packet, QP-based hash algorithm is used to select one specific SL for fowarding the RoCEv2 packet, as shown below:

* QP1 Packet: SRv6 Policy 1 -> SL1 (PE1->P1->PE2)

* QP2 Packet: SRv6 Policy 1 -> SL2 (PE1->P2->PE2)

* QP3 Packet: SRv6 Policy 2 -> SL1 (PE1->P3->PE2)

* QP4 Packet: SRv6 Policy 2 -> SL2 (PE1->P4->PE2)

PE1 will encapsulate each RoCEv2 packet with an outer IPv6 header and SRH using the selected SL, and then forward it to the appropriate link.

The PE1->P1 link carries the traffic of QP1, the PE1->P2 link carries the traffic of QP2, the PE1->P3 link carries the traffic of QP3, and the PE1->P4 link carries the traffic of QP4.

6. Operational Considerations

In ingress device, the control plane must support QP range to SRv6 Policy mapping by protocol extension or local configuration. For non-RoCEv2 traffic, the system MUST revert to the standard five-tuple hash for SL selection.

The ingress devices require deep packet inspection capability to parse BTH headers, programmable hash engines with configurable input fields, sufficient TCAM/SRAM for QP classification mapping tables, and support for multiple active SRv6 policies with multiple SLs.

When network congestion or failure occurs, operators can flexibly configure QP range to SRv6 Policy mapping strategies on the ingress device to guide RoCEv2 flows to the appropriate path.

7. Security Considerations

Malicious actors could spoof QP values to bypass mapping policies, cause hash collisions, or exhaust specific network paths. Mitigations may include cryptographic validation of RoCEv2 packets, and QP whitelisting/blacklisting.

QP values may reveal application-level information, so QP values SHOULD be anonymized or encrypted.

The additional packet processing (such as parsing BTH headers) could be exploited for Denial of Service (DoS) attacks; therefore, implementations MUST support graceful degradation mechanisms (such as rate limiting) under attack.

8. IANA Considerations

This document has no IANA actions.

9. References

9.1. Normative References

[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>.

[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>.

[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>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

9.2. Informative References

[IBTA-SPEC] InfiniBand Trade Association, "InfiniBand Architecture Specification", InfiniBand Architecture Specification Volume 1-2, Release 1.6, December 2023, <https://www.infinibandta.org/ibta-specification/>.

Authors' Addresses

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

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

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