Network Working Group
Internet Draft
Intended status: Standards Track
Expires: September 15, 2025

J. Li
China Mobile
C. Lin
New H3C Technologies

March 15, 2025

BGP-Link State (BGP-LS) Advertisement of Flow Queue
draft-li-idr-bgpls-flow-queue-01

Abstract

In the traffic congestion management of routers, queueing technology is generally used to classify and transmit traffic. Therefore, queue information indirectly reflects the congestion status of the router. This document makes the necessary expansion of the BGP-LS mechanism to send queue information to the network controller in a scalable way, enabling the network controller to understand the router's traffic congestion status and make appropriate adjustments to the network traffic.

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 September 15, 2025.

Copyright Notice

Table of Contents

1. Introduction

In complex network forwarding processes, when the maximum resources supported by network equipment cannot meet the resources required for normal forwarding, a phenomenon known as traffic congestion occurs, causing negative impacts such as latency and jitter in traffic transmission, leading to a decline in network service quality.

As networks become increasingly complex, traffic congestion within the network is inevitable. To alleviate congestion, a scheduler policy needs to be formulated to determine the processing order of packet forwarding. This scheduler policy generally adopts queue technology, using a queue algorithm to classify traffic and a precedence algorithm to send traffic. Several queue algorithms exist, such as First-In First-Out (FIFO) queue, Priority Queue (PQ), and Custom Queue (CQ), each addressing specific network traffic issues. In other words, queue information can reflect the congestion status of network nodes.

When congestion occurs in the network, network managers need to promptly detect and respond to it by collecting traffic queue information from network nodes and reporting it to the network controller. When the network controller receives queue information

indicating congestion at network nodes, it can execute some actions like traffic engineering. However, specific actions are not designated in this document.

This document specifies the BGP-LS mechanism with necessary expansions for scalable transmission of queue information to the network controller. Queue information is sent using a separate type of BGP-LS NLRI, allowing flexible updates of queue information without affecting the based link state information.

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.

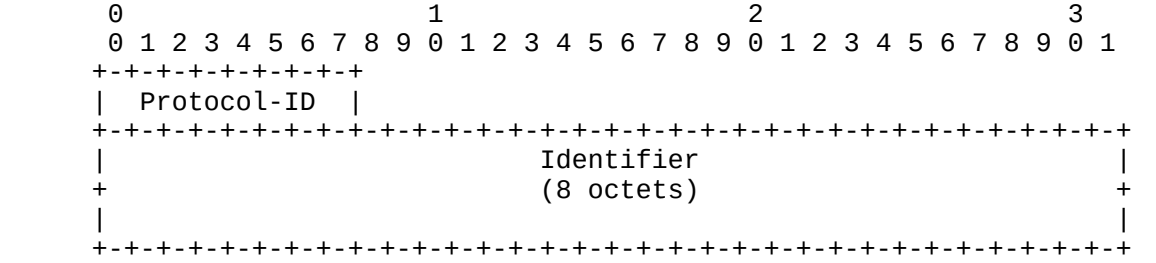
2. BGP-LS Extensions for Queue Information Advertisement

BGP-LS [RFC9552] defines the mechanisms to advertise Node, Link, and Prefix Link-State NLRI types and their associated attributes via BGP.

This document proposes expanding BGP-LS to carry queue information. Each queue operates on the corresponding interface of a network node (NN) for traffic forwarding on that interface. Queue information related to a link is advertised through a new BGP-LS NLRI and associated BGP-LS attributes. The queue ID and the link ID associated with the queue are carried using a new BGP-LS Queue Link NLRI, and the queue attributes of the related link are conveyed via relevant BGP-LS attribute TLVs.

2.1. BGP-LS Queue Link NLRI

A new BGP-LS NLRI type called "Queue-link" is defined to advertise Queue-specific link information. The NLRI-Type is to be assigned by IANA (TBD1). Its format is shown as below:



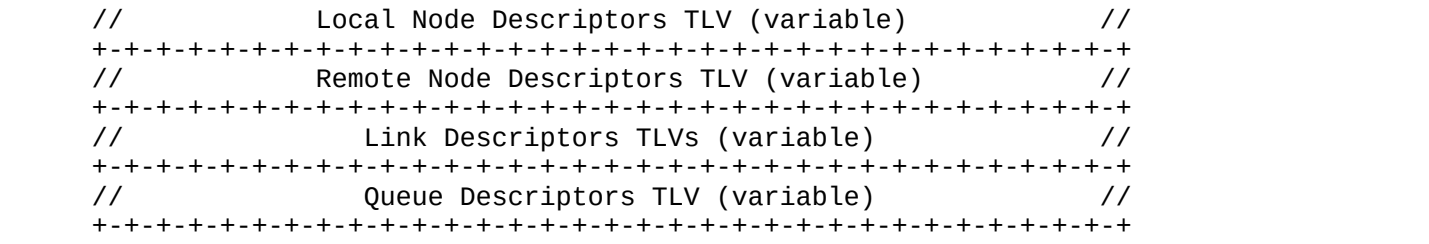


Figure 1: Encoding of NRP-Link NLRI

The encoding and semantics of Protocol-ID, Identifier, Local Node Descriptors and Remote Node Descriptors are the same as defined in [RFC9552]. If interface and neighbor addresses, either IPv4 or IPv6, are present, then the interface/neighbor address TLVs MUST be included in the Link Descriptors. If the link borrows the address from another interface, then both the Link Local/Remote Identifiers and the interface/neighbor address TLVs MUST be included.

The Queue Descriptors TLV includes descriptions of the Queue which the link is associated with. This is a mandatory TLV for Queue-Link NLRIs. Its type is to be assigned by IANA (TBD2). The length of this TLV is variable. The value contains one or more Queue Descriptor sub-TLSs defined in Section 2.1.1.

2.1.1. Queue Descriptors Sub-TLVs

In this document, one Queue Descriptors sub-TLV is defined as below:

Sub-TLV Code Point	Description	Length
TBD3	Queue ID	4

Table 1: Queue Descriptor Sub-TLVs

Queue ID: A 32-bit link-interface-wide unique identifier, which is used to identify a Queue that the link is associated with.

The Queue ID sub-TLV is mandatory in the Queue Descriptors. There MUST be only one instance of Queue ID sub-TLV present in the Queue Descriptors.

2.2. Queue Attribute TLVs

The Queue Attribute TLVs are a set of TLVs which may be encoded in the BGP-LS Attribute associated with a Queue-Link NLRI.

The following Queue Attribute TLVs can be defined in this document.

TLV Code Point	Description	Length
TBD4	Queue Depth	4
TBD5	Queue Precedence	1
TBD6	Queue Length	4

Table 2: Queue Attribute TLVs

2.2.1. Queue Depth TLV

A new Queue Attribute TLV is defined to carry the Queue Depth information, which indicates the upper limit of the queue capacity. The form of the Queue Depth TLV is as below:

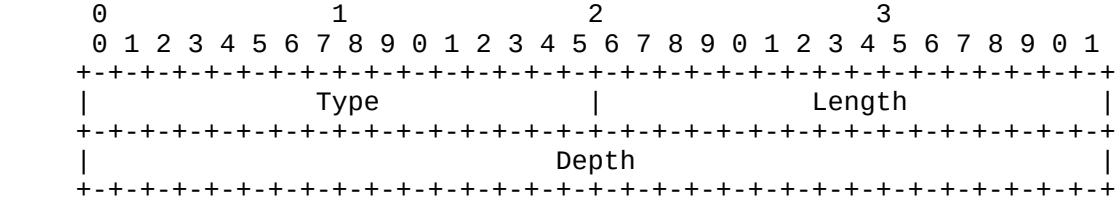


Figure 2: Encoding of Queue Depth TLV

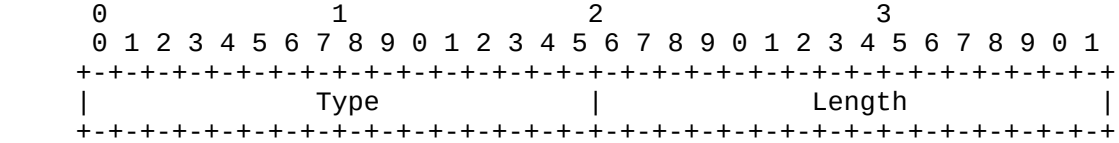
Type (16 bits): TBD4.

Length (16 bits): Length of the value field in octets. The value is 4.

Depth: Indicates the upper limit of the queue capacity.

2.2.2. Queue Precedence TLV

When there are multiple queues on a specified link, the Queue Precedence TLV is used to determine the order in which the queues are scheduled. The format of the Queue Precedence TLV is as below:



```
|   Precedence  |
+-+-+-+-+-+-+-+-+
```

Figure 3: Encoding of Queue Precedence TLV

Type (16 bits): TBD5.

Length (16 bits): Length of the value field in octets. The value is 1.

Precedence: The Precedence of the Queue. The larger the value, the higher the priority.

2.2.3. Queue Length TLV

A new Queue Attribute TLV is defined to carry the Queue Length information, which indicates the number of packets actually waiting to be forwarded in the current queue. The form of the Queue Length TLV is as below:

```
 0               1               2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Type              |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Queue Length                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Encoding of Queue Length TLV

Type (16 bits): TBD6.

Length (16 bits): Length of the value field in octets. The value is 4.

Queue Length: The number of packets actually waiting to be forwarded in the current queue. If the queue length exceeds the queue depth, packets will be discarded.

3. Operational Considerations

It SHOULD implement the functionality to allow the operator to globally control whether to send link-related queue information. It SHOULD also allow the operator to control whether to send queue information for a specified link.

It MUST implement the functionality to allow the operator to configure the queue length threshold, where the queue information is

only triggered and sent after reaching this threshold, ensuring that queue information is sent only during actual link congestion and reducing the frequency of advertisements.

After reaching the queue length threshold, it MUST allow the operator to configure the queue length variation threshold, where updated queue information is sent only when the queue length variation reaches this threshold, to avoid frequent advertisements.

4. Security Considerations

This document introduces no additional security vulnerabilities in addition to the ones as described in [RFC9552].

5. IANA Considerations

There is a need to request IANA to assign a new code point for "Queue-Link NLRI" under the "BGP-LS NLRI Types" Registry.

Type	NLRI Type	Reference
TBD1	Queue Link NLRI	This document

Table 3: Queue NLRI Type

There also is a need to request IANA to assign the following new code points for under the "BGP-LS NLRI and Attribute TLVs" Registry.

TLV Code Point	Description	Reference
TBD2	Queue Descriptors	This document
TBD3	Queue ID	This document
TBD4	Queue Depth	This document
TBD5	Queue Precedence	This document
TBD6	Queue Length	This document

Table 4: Queue related TLV/Sub-TLVs

6. References

6.1. 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>.

[RFC9552] Talaulikar, K., Ed., "Distribution of Link-State and Traffic Engineering Information Using BGP", RFC 9552, DOI 10.17487/RFC9552, December 2023, <https://www.rfc-editor.org/info/rfc9552>.

Author’s Address

Jinming Li
China Mobile
32 Xuanwumen West Street
Beijing
Xicheng District, 100053
China
Email: lijinming@chinamobile.com

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