Network Working Group C. Lin
Internet-Draft M. Chen
Intended status: Standards Track H. Li
Expires: May 12, 2022 H3C
November 8, 2021

Distribution of Device Discovery Information in NVMe Over RoCEv2 Storage Network Using BGP

draft-lin-idr-bgp-nof-nlri-00

Abstract

This document proposes a method of distributing device discovery information in NVMe over RoCEv2 storage network using the BGP routing protocol. A new BGP Network Layer Reachability Information (NLRI) encoding format, named NoF NLRI, is defined.

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 May 12, 2022.

Copyright Notice

Table of Contents

1. Introduction 2
1.1. Requirements Language 3
2. Distribution of Device Discovery Information Using BGP . . . 3
3. BGP Extentions 5
3.1. TLV Format 5
3.2. NoF NLRI 6
3.3. Device Discovery NLRI 7
3.3.1. IPv4 Address TLV 8
3.3.2. IPv6 Address TLV 8
3.3.3. Role Type TLV 9
3.3.4. Online/Offline Status TLV 9
3.3.5. More Device Info TLVs 10
3.4. Device Zone NLRI 10
3.5. Operations 11
4. IANA Considerations 11
5. Security Considerations 11
6. References 11
6.1. Normative References 11
6.2. Informative References 12
Authors' Addresses 12

1. Introduction

As data center networks keep growing, the performance of communication methods needs to accelerate. At present, NVMe over RoCEv2 is becoming a popular solution of storage network based on Ethernet. In such network, a host accesses to an NVMe storage subsystem via Ethernet Fabric with RoCEv2 protocol.

In the traditional way, the discovery of hosts and storage subsystems is achieved by manual configurations. However the manual way is difficult for management and maintenance. In addition, the reaction speed is slow when a device goes online or offline, making it hard to realize hot-plug and failover. To solve these problems, automatic discovery method should be deployed.

LLDP is generally used to achieve the discovery task when a host or storage subsystem is directly connected to a switch. Then, the device discovery information is distributed to others switches in the fabric. Finally, other devices get the information from the switches which they directly connect with.

This document proposes a new method of distributing device discovery information among switches in NVMe over RoCEv2 storage network using the BGP routing protocol [RFC4271].

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. Distribution of Device Discovery Information Using BGP

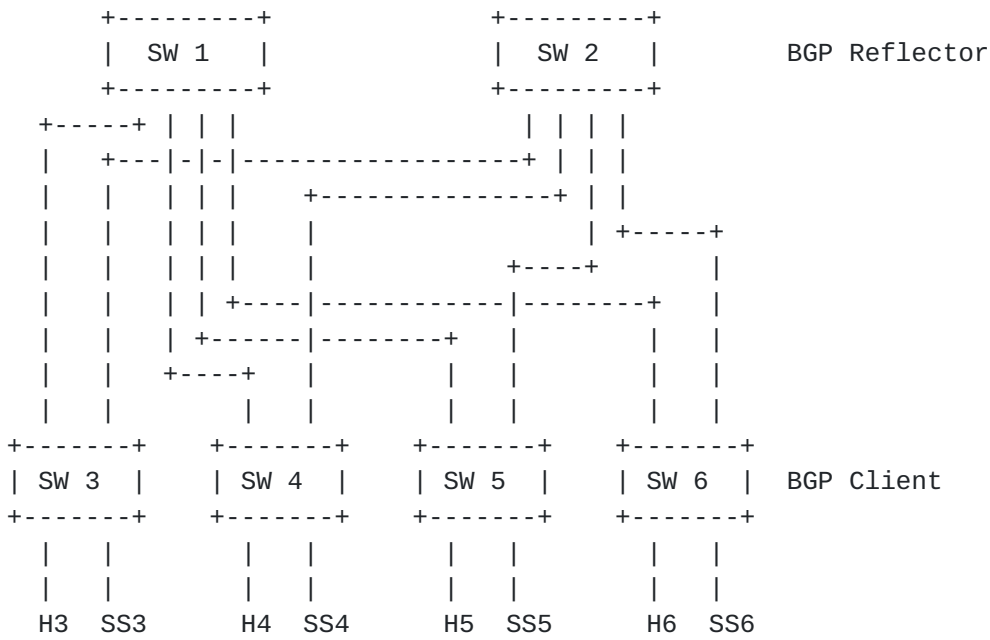
In hierarchical topology, a host or storage subsystem is usually connected to a switch at access layer. In Clos topology, a host or storage subsystem is usually connected to a "Leaf" switch. To keep terminology uniform, in this document the switches which the hosts and storage subsystems directed connect with will be referred to as the access switches.

When any host or storage subsystem is connected with an access switch, it periodically sends LLDP messages to the access switch. According to the received LLDP messages, the access switch maintains the states of directly connected devices. If the state of any device changes, such as going online or offline, the access switch will announce the other devices connected with it. However, the devices on the other access switches may also be concerned with the device discovery information, especially in a large-scale storage network. For example, when a storage subsystem is newly connecting to an access switch, a host located in another access switch needs to know that it gets online. Then the host will establish connection with the storage subsystem, and transmit data through NVMe over RoCEv2. Therefore, the access switches are required to distribute device discovery information among them.

In this document the distribution of device discovery information among access switches is achieved by using BGP. All the access switches are BGP speakers, and the device discovery information is exchanged as BGP routes among them.

In order to reduce the number of BGP connections, the application of BGP Route Reflectors [RFC4456] is recommended. Figure 1 shows an example of BGP connections with route reflectors. SW 1 and SW 2 serve as reflectors, and SW 3, SW 4, SW 5 and SW 6 are their clients. When a client sends a BGP route, which contains device discovery information, to a reflector, the reflector will reflect the route to the other clients. Therefore, all the access switches work as

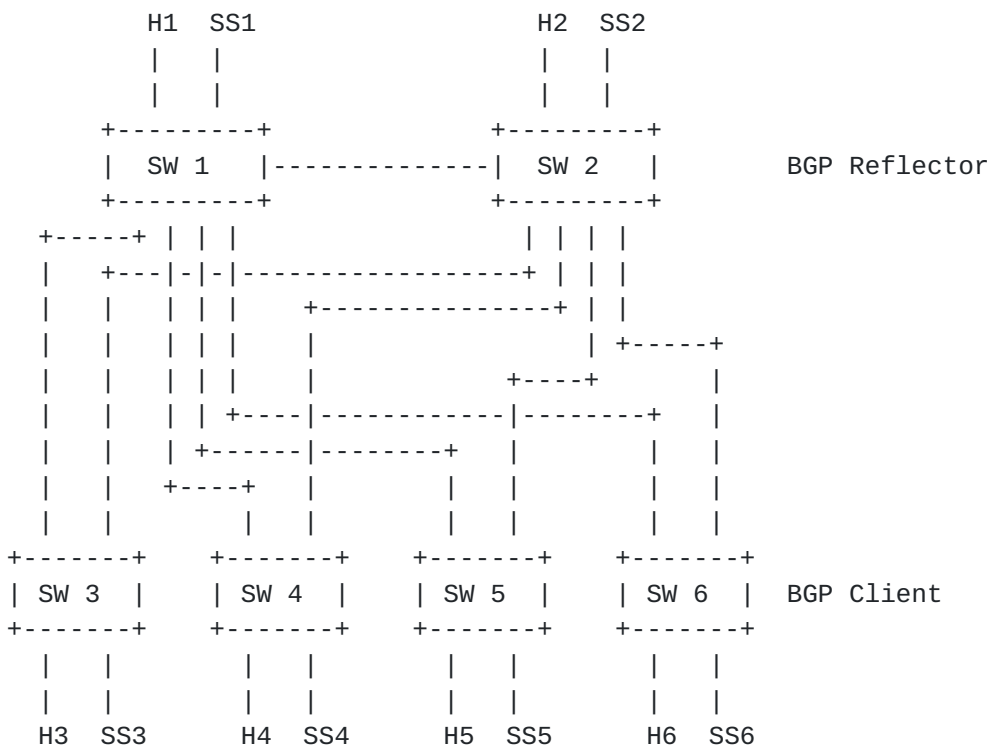
clients, and each of them only needs to establish BGP connections to the reflectors, rather than establishing BGP connections between each other. In this example, there are two reflectors, SW 1 and SW 2, which run as a hot standby for each other. It is also fine to deploy only one reflector in the network. However, to improve availability, deploying more than one reflectors are recommended.



SW: Switch
H: Host
SS: Storage Subsystem

Figure 1 BGP Connections with Route Reflectors

In Figure 1, the reflector switches are not directly connected with hosts or storage subsystems, and they are not access switches. Figure 2 shows another example, in which case two of the access switches serve as BGP route reflectors. The main difference with Figure 1 is that the reflectors, SW 1 and SW 2, also need to establish BGP connections between each other. If any device directly connected with the reflector goes online or offline, the reflector not only sends the device discovery information to its clients, but also sends information to the other reflectors.



SW: Switch
H: Host
SS: Storage Subsystem

Figure 2 Access Switches Serve as Reflectors

This document mainly focus on the distribution method of device discovery information among access switches. The interaction between access switch and host, or the interaction between access switch and storage subsystem, is beyond the scope of this document.

3. BGP Extentions

This document describes a mechanism by which device discovery information can be distributed using the BGP routing protocol. This is achieved using a new BGP Network Layer Reachability Information (NLRI) encoding format, named NoF NLRI.

3.1. TLV Format

Information in the NoF NLRI is encoded in Type/Length/Value triplets. The TLV format is shown in Figure 3.

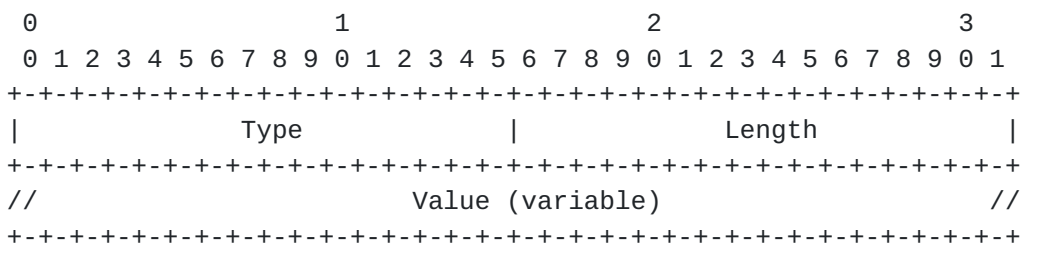


Figure 3: TLV Format

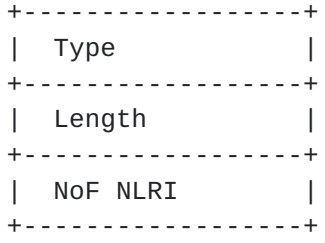
The Length field defines the length of the value portion in octets (thus, a TLV with no value portion would have a length of zero). The TLV is not padded to 4-octet alignment. Unrecognized types MUST be preserved and propagated.

3.2. NoF NLRI

New AFI and SAFI are defined for the NoF NLRI: the NoF AFI/SAFI (values to be assigned by the IANA).

In order for two BGP speakers to exchange NoF NLRI, they MUST use BGP Capabilities Advertisement to ensure that they are both capable of properly processing such NLRI. This is done as specified in [RFC4760].

The format of the NoF NLRI is shown in the following figure.



where:

- Type: the type of NoF NLRI.
- Length: the length of the rest of the NLRI in octets, not including the Type field or itself.
- NoF NLRI: carrying the device discovery information in NVMe over Fabric networks.

BGP NoF NLRI for both IPv4 and IPv6 networks can be carried over either an IPv4 BGP session or an IPv6 BGP session. If an IPv4 BGP session is used, then the next hop in the MP_REACH_NLRI SHOULD be an

IPv4 address. Similarly, if an IPv6 BGP session is used, then the next hop in the MP_REACH_NLRI SHOULD be an IPv6 address. Usually, the next hop will be set to the local endpoint address of the BGP session. The next-hop address MUST be encoded as described in [RFC4760].

The Device Discovery NLRI and Device Zone NLRI are currently defined in this document. More types of NLRI will be included in the future version.

Type	NoF NLRI Type
1	Device Discovery NLRI
2	Device Zone NLRI

3.3. Device Discovery NLRI

The Device Discovery NLRI is used to carry the discovery information of directly connected devices. The format of the Device Discovery NLRI is shown in the following figure.

Router ID	4 octets
Mac Address	6 octets
Port Name Length	2 octets
Port Name	variable
Device Info	variable

where:

- Router ID: the Router ID of the access switch which originates this NLRI, usually the same as the BGP Identifier.

- Mac Address: the Mac Address of a connected device.

- Port Name Length: the length of the following Port Name field in octets.

- Port Name: the name of the connecting port, to distinguishing different ports which share the same Mac Address.

- Device Info: the specific information of the connected device and its connecting port, which are identified by the above Mac Address and Port Name fields.

The Device Discovery NLRI carries the information of a device which is identified by the Router ID of the access switch and the Mac Address and Port Name of the connected port.

For the purpose of BGP route key processing, only the Router ID, Mac Address, MAC Address, Port Name Length, and Port Name fields are considered to be part of the prefix in the NLRI.

The Device Info field may contain the following TLVs.

3.3.1. IPv4 Address TLV

The format of the IPv4 Address TLV is shown in the following figure.

where:

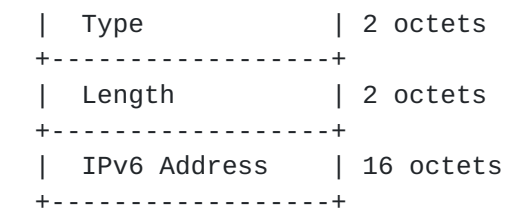
- Type: 1.

- Length: 4.

- IPv4 Address: the IPv4 Address of the connecting port.

3.3.2. IPv6 Address TLV

The format of the IPv6 Address TLV is shown in the following figure.

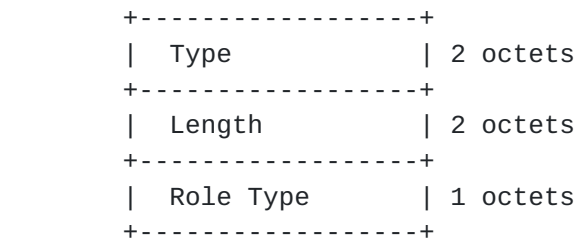


where:

- Type: 2.
- Length: 16.
- IPv6 Address: the IPv6 Address of the connecting port.

3.3.3. Role Type TLV

The format of the Role Type TLV is shown in the following figure.

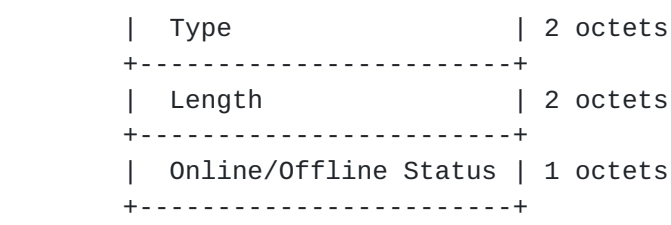


where:

- Type: 3.
- Length: 1.
- Role Type: the role of the device. The following values are defined.
  - 1: storage subsystem.
  - 2: host.
  - 3: the device can serve as both a host and a storage subsystem.

3.3.4. Online/Offline Status TLV

The format of the Online/Offline Status TLV is shown in the following figure.



where:

- Type: 4.
- Length: 1.
- Online/Offline Status: indicating the device is online or offline. The following values are defined.
  - 0: offline.
  - 1: online.

3.3.5. More Device Info TLVs

More Device Info TLVs will be included in the future version of this document.

3.4. Device Zone NLRI

In storage networks, hosts and storage subsystems are generally divided into several zones. Only the devices in the same zone are allowed to discover and communicate with each other.

The Device Zone NLRI is used to distribute the zone configuration of a device. The format of the Device Zone NLRI is shown in the following figure.

```
+------------------+
|  Router ID       | 4 octets
+------------------+
|  IP Address      | 4 or 16 octets
+------------------+
|  Zone Name Length| 2 octets
+------------------+
|  Zone Name       | variable
+------------------+
```

where:

- Router ID: the Router ID of the access switch which originates this NLRI, usually the same as the BGP Identifier.
- IP Address: the IPv4 or IPv6 Address of a connected device.
- Zone Name Length: the length of the following Zone Name field in octets.
- Zone Name: the name of the zone which the connected device belongs to.

3.5. Operations

The source of the NoF NLRI can be a dedicated module which receive LLDP messages and maintain the states of directly connected devices. For the originator of an NoF NLRI route, BGP receives information from relevant module, encapsulates the information into an NoF NLRI route, and sends the route to other peers. For the receiver of an NoF NLRI route, BGP extracts the NoF NLRI from the route and sends the information to relevant module.

The NoF NLRI field may be treated as an opaque hexadecimal string, depending on the implementation.

4. IANA Considerations

TBD.

5. Security Considerations

TBD.

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>.

[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>.

[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>.

[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>.

6.2. Informative References

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

Authors' Addresses

Changwang Lin
H3C

Email: linchangwang.04414@h3c.com

Mengxiao Chen
H3C

Email: chen.mengxiao@h3c.com

Hao Li
H3C

Email: lihao@h3c.com