

Network Working Group  
Internet Draft  
Intended status: Standards Track  
Expires: January 02, 2026

R. Wang  
China Mobile  
C. Lin  
New H3C Technologies  
R. Zhuang  
F. Qin  
Q. Zhang  
W. Wang  
China Mobile  
July 02, 2025

Distribution of Device Discovery Information in NVMe Over RoCEv2  
Storage Network Using BGP  
draft-wang-idr-bgp-nof-nlri-03

## 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 January 02, 2026.

## Copyright Notice

Copyright (c) 2025 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction.....	2
1.1. Requirements Language.....	3
2. Distribution of Device Discovery Information Using BGP.....	3
3. BGP Extensions.....	6
3.1. TLV Format.....	6
3.2. NoF NLRI.....	7
3.3. Device Discovery NLRI.....	8
3.3.1. IPv4 Address TLV.....	9
3.3.2. IPv6 Address TLV.....	9
3.3.3. Role Type TLV.....	10
3.3.4. Service Protocol TLV.....	10
3.3.5. Device Status TLV.....	11
3.3.6. Status Changing Reason TLV.....	12
3.3.7. More Device Info TLVs.....	13
3.4. Device Zone NLRI.....	13
3.5. Operations.....	14
4. Security Considerations.....	14
5. IANA Considerations.....	14
6. References.....	14
6.1. Normative References.....	14
6.2. Informative References.....	15
Authors' Addresses.....	16

## 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.

When a host or storage subsystem is directly connected to a switch, the device reports its information to the switch using device discovery protocol like LLDP. 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.

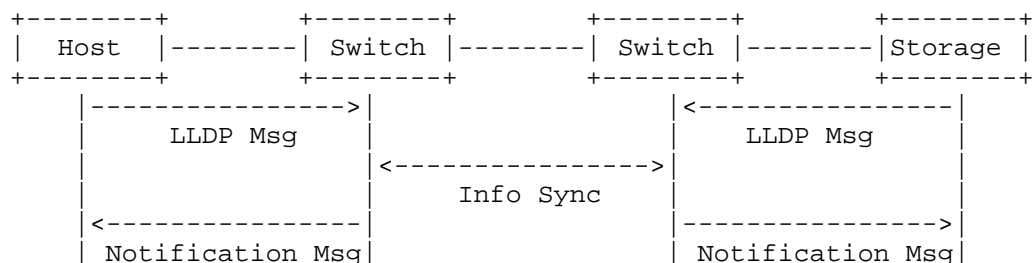
### 1.1. Requirements Language

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.

[ODCC-2020-05016] describes a solution for device discovery in NVMe over RoCEv2 storage network.



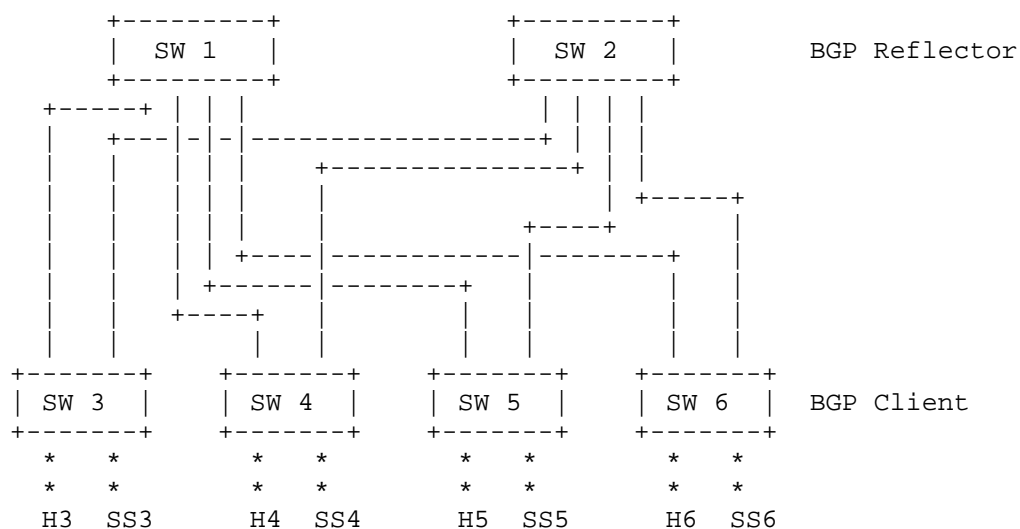
When any host or storage subsystem is connected with an access switch, it periodically sends LLDP messages to the access switch, including its own information and subscription of interested devices. The access switch receives the LLDP messages and maintains the states of directly connected devices. If the state of any device changes, such as going online or offline, the access switch will notify the other directly connected devices which have subscribed.

However, the devices on the other access switches may also be interested 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. If that storage subsystem failed, that host, which has NVMe connection with the failed storage subsystem, needs to be notified as soon as possible. Then the host will quickly disconnect from the storage subsystem and switch over to the redundant service. Therefore, the access switches are required to distribute device discovery information among them. So that the host can get the required information from the directly connected switch.

[ODCC-2020-05016] specifies the definitions of LLDP messages and notification messages between access switches and hosts or storage subsystems, but leaves the information synchronization method undefined.

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 multiple reflectors are recommended.



SW: Switch  
H: Host  
SS: Storage Subsystem  
--: BGP Connection  
\*: Access Link

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.

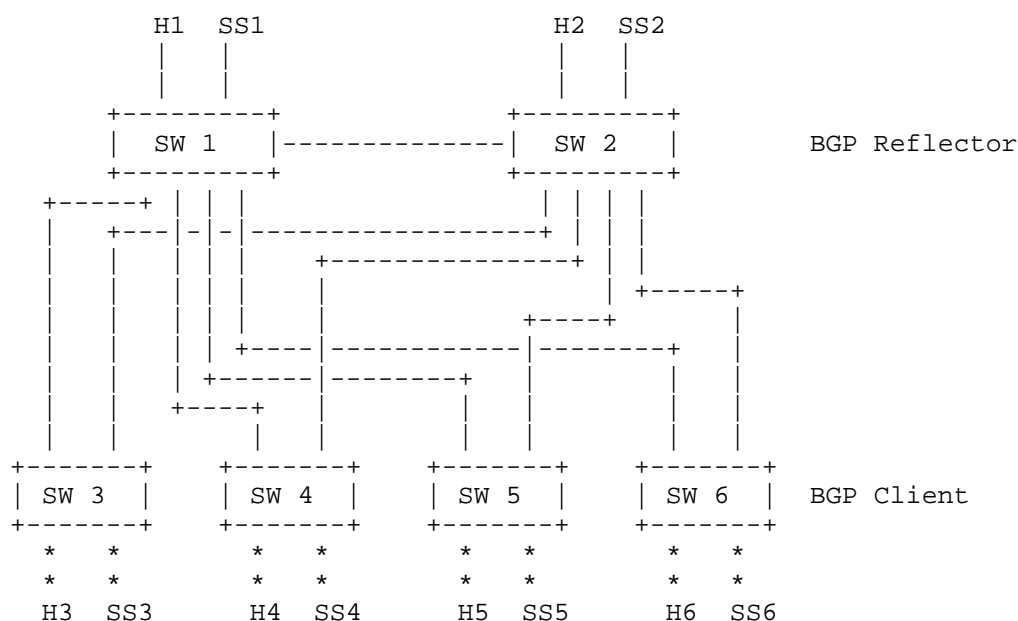


Figure 2 Access Switches Serve as Reflectors

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 [RFC4271]. 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.

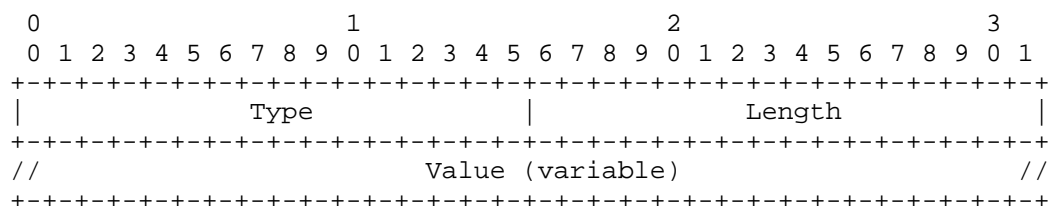


Figure 3: TLV Format

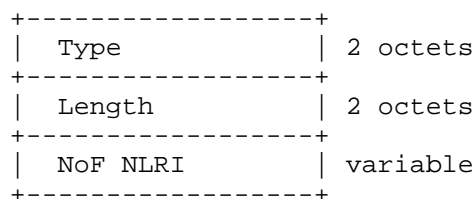
The Length field defines the length of the value portion in octets (thus, a TLV with no value portion would have a length of zero). The TLV is not padded to 4-octet alignment. Unrecognized types MUST be preserved and propagated.

### 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:

- o Type: the type of NoF NLRI.
- o Length: the length of the rest of the NLRI in octets, not including the Type field or itself.
- o 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:

- o Router ID: the Router ID of the access switch which originates this NLRI, usually the same as the BGP Identifier.
- o Mac Address: the Mac Address of a connected device.
- o Port Name Length: the length of the following Port Name field in octets.
- o Port Name: the name of the connecting port, to distinguishing different ports which share the same Mac Address.

- o 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, 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.

+-----+	
Type	2 octets
+-----+	
Length	2 octets
+-----+	
IPv4 Address	4 octets
+-----+	

where:

- o Type: 1.
- o Length: 4.
- o 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.

+-----+	
Type	2 octets
+-----+	
Length	2 octets
+-----+	
IPv6 Address	16 octets
+-----+	

where:

- o Type: 2.
- o Length: 16.
- o 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.

+-----+		
	Type	2 octets
+-----+		
	Length	2 octets
+-----+		
	Role Type	1 octets
+-----+		

where:

- o Type: 3.
- o Length: 1.
- o 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. Service Protocol TLV

The format of the Service Protocol TLV is shown in the following figure.

Type	2 octets
Length	2 octets
Protocol Type	1 octets
Protocol Version	2 octets
Protocol Port	2 octets
Protocol Identifier Length	1 octets
Protocol Identifier	variable octets

where:

- o Type: 4.
- o Length: the length of the rest of the TLV in octets.
- o Protocol Type: the type of the service protocol. The following values are defined.
  - \* 0: NVMe over RoCEv2.
- o Protocol Version: the version of the service protocol.
- o Protocol Port: the port number used by the service protocol. The value 0 indicates the default or well-known port number.
- o Protocol Identifier Length: the length of the following Protocol Identifier field in octets.
- o Protocol Identifier: the device identifier used by the service protocol.

### 3.3.5. Device Status TLV

The format of the Device Status TLV is shown in the following figure.

Type	2 octets
Length	2 octets
Device Status	4 octets

where:

- o Type: 5.
- o Length: 4.
- o Device Status: the current status of the device. The following values are defined.
  - \* 0: offline.
  - \* 1: online.

### 3.3.6. Status Changing Reason TLV

The format of the Device Status TLV is shown in the following figure.

Type	2 octets
Length	2 octets
Status Changing Reason	4 octets

where:

- o Type: 6.
- o Length: 4.
- o Status Changing Reason: the reason of the device status changing. The following values are defined.
  - \* 0: normal.
  - \* 1: link failure.
  - \* 2: PFC storm.

- \* 3: access network failure.
- \* 4: zone change.
- \* 5: configuration change.
- \* 6: lldp age-out.

### 3.3.7. 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 Type		1 octets
+-----+		
IP Address		4 or 16 octets
+-----+		
Zone Name Length		2 octets
+-----+		
Zone Name		variable
+-----+		

where:

- o Router ID: the Router ID of the access switch which originates this NLRI, usually the same as the BGP Identifier.
- o IP Type: indicating the type of IP Address. The following values are defined.
  - \* 0: IPv4.
  - \* 1: IPv6.
- o IP Address: the IPv4 or IPv6 Address of a connected device.

- o Zone Name Length: the length of the following Zone Name field in octets.
- o 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. Security Considerations

TBD

### 5. IANA 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>>.
- [ODCC-2020-05016] Open Data Center Committee, "NVMe over RoCEv2 Network Control Optimization Technical Requirements and Test Specifications", 2020.

Authors' Addresses

Ruixue Wang  
China Mobile  
China  
Email: wangruixue@chinamobile.com

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

Rui Zhuang  
China Mobile  
China  
Email: zhuangruiyjy@chinamobile.com

Fengwei Qin  
China Mobile  
China  
Email: qinfengwei@chinamobile.com

Qi Zhang  
China Mobile  
China  
Email: zhangqiyjy@chinamobile.com

Wenxuan Wang  
China Mobile  
China  
Email: wangwenxuan@chinamobile.com

