IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 3, 2022

W. Wang
A. Wang
China Telecom
H. Wang
Huawei Technologies
G. Mishra
Verizon Inc.
S. Zhuang
J. Dong
Huawei Technologies

September 30, 2021

Route Distinguisher Outbound Route Filter (RD-ORF) for BGP-4
draft-wang-idr-rd-orf-08

Abstract

This draft defines a new Outbound Route Filter (ORF) type, called the Route Distinguisher ORF (RD-ORF). The described RD-ORF mechanism is applicable when the VPN routes from different VRFs are exchanged via one shared BGP session(e.g. routers in a single-domain connect via Route Reflector).

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 April 3, 2022.

Table of Contents

1. Introduction

[I-D.wang-idr-vpn-routes-control-analysis] analysis the scenarios and necessaries for VPN routes control in the shared BGP session. This draft analizes the existing solutions and their limitations for these scenarios, proposes the new RD-ORF solution to meet the requirements that described in section 8 of [I-D.wang-idr-vpn-routes-control-analysis].

Now, there are several solutions can be used to alleviate these problem:

- Route Target Constraint (RTC) as defined in [RFC4684]
- Address Prefix ORF as defined in [RFC5292]
- PE-CE edge peer Maximum Prefix
- Configure the Maximum Prefix for each VRF on edge nodes

However, there are limitations to existing solutions:

1) Route Target Constraint

RTC can only filter the VPN routes from the uninterested VRFs, if the "trashing routes" come from the interested VRF, filter on RTs will erase all prefixes from this VRF.

2) Address Prefix ORF

Using Address Prefix ORF to filter VPN routes need to pre-configuration, but it is impossible to know which prefix may cause overflow in advance.

3) PE-CE edge peer Maximum Prefix

This mechanism can only protect the edge between PE-CE, it can't be deployed within PE that peered via RR. Depending solely on the edge protection is dangerous, because if only one of the edge points being comprised/error-configured/attacked, then all of PEs within domain are under risk.

4) Configure the Maximum Prefix for each VRF on edge nodes

When a VRF overflows, it stops the import of routes and log the extra VPN routes into its RIB. However, PEs still need to parse the BGP updates. These processes will cost CPU cycles and further burden the overflowing PE.

This draft defines a new ORF-type, called the Route Distinguisher ORF (RD-ORF). Using RD-ORF mechanism, VPN routes can be controlled based on RD. This mechanism is event-driven and does not need to be pre-configured. When a VRF of a router overflows, the router will find out the RD of excessive VPN routes in this VRF, and send a RD-ORF to its BGP peer that carrys the RD. If a BGP speaker receives a RD-ORF entry from its BGP peer, it will filter the VPN routes it tends to send according to the entry.

RD-ORF is applicable when the VPN routes from different VRFs are exchanged via one shared BGP session.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] .

3. Terminology

The following terms are defined in this draft:

- RD: Route Distinguisher, defined in [RFC4364]
- ORF: Outbound Route Filter, defined in [RFC5291]
- AFI: Address Family Identifier, defined in [RFC4760]
- SAFI: Subsequent Address Family Identifier, defined in [RFC4760]
- EVPN: BGP/MPLS Ethernet VPN, defined in [RFC7432]
- RR: Router Reflector, provides a simple solution to the problem of IBGP full mesh connection in large-scale IBGP implementation.
- VRF: Virtual Routing Forwarding, a virtual routing table based on VPN instance.

4. Operation process of RD-ORF mechanism on sender

The operation of RD-ORF mechanism on each device is independent, each of them makes a local judgement to determine whether it needs to send RD-ORF to its peers.

When the RD-ORF mechanism is triggered, the device must send an alarm information to network operators.

4.1. Intra-domain Scenarios and Solutions

For intra-AS VPN deployment, there are three scenarios:

- RD is allocated per VPN/per PE, each VRF only import one RT(see Section 4.1).
- RD is allocated per VPN/per PE. Multiple RTs are associated with such VPN routes, and be imported into different VRFs in other devices(see Section 4.2).
- RD is allocated per VPN, each VRF imports one/multiple RTs(see Section 4.3).

The following sections will describe solutions to the above scenarios in detail.

4.1.1. Scenario-1 and Solution (Unique RD, One RT)

In this scenario, RD is allocated per VPN or per PE, each VRF only import one RT. We assume the network topology is shown in Figure 1.

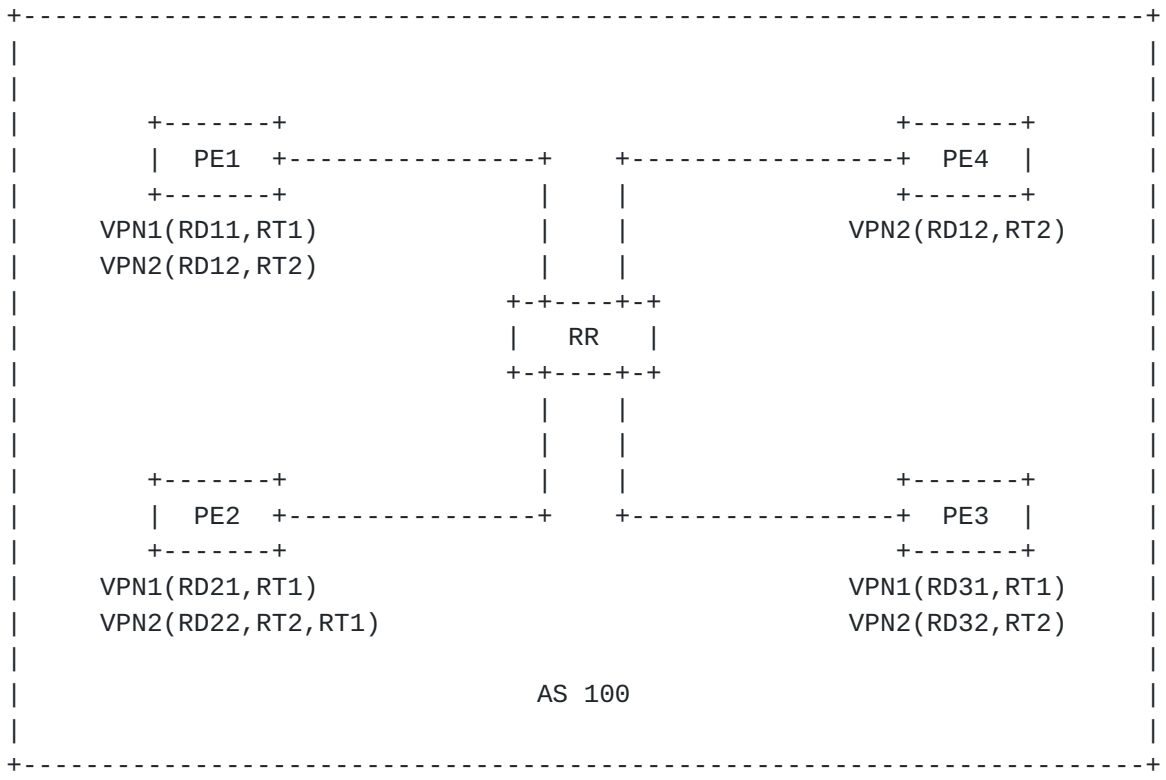

Figure 1 Network Topology of Scenario-1

When PE3 sends excessive VPN routes with RT1, while both PE1 and PE2 import VPN routes with RT1, the process of excessive VPN routes will influence performance of VRFs on PEs. PEs and RR should have some mechanisms to identify and control the advertisement of excessive VPN routes.

On PE1, each VRF has a set threshold, we assume it is 80% of Maximum Prefix of VRF. When the number of VPN1 VRF routing entries reaches the threshold, PE1 will start monitoring the RD carried by the received VPN routing entries. Once the number of VPN routing entries exceed the prefix limit, PE1 will calculate the RD and its source PE received the most times during this period, the result is RD31 from PE3, which is associated with RT1. Then, PE1 will locally discards the VPN routes carry RD31 which come from PE3 in VRF1.

Due to there is no other VRFs on it to import the VPN routes with RT1. after local processing, PE1 will generate a BGP ROUTE-REFRESH message contains a RD-ORF entry, and send to RR. RR will withdraw and stop to advertise such excessive VPN routes to PE1.

On PE2, the local processing is the same as PE1. Due to there has other VRF on it to import the VPN routes with RT1, PE2 triggers the RD-ORF message to RR(RD field is set to RD31) only when all the VRFs that import RT1 are overflowed.

4.1.2. Scenario-2 and Solution (Unique RD, Multiple RTs)

In this scenario, RD is allocated per VPN or per PE. Multiple RTs are associated with such VPN routes, and be imported into different VRFs in other devices. We assume the network topology is shown in Figure 2.

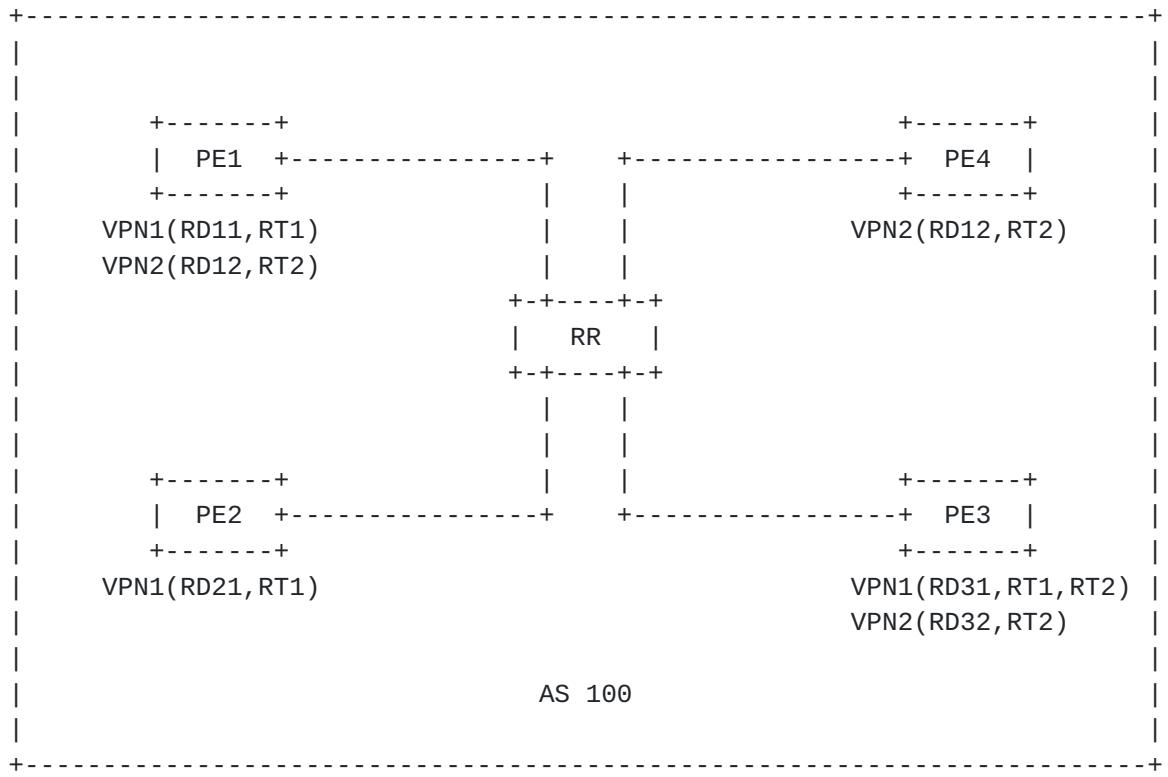

Figure 2 Network Topology of Scenario-2

When PE3 sends excessive VPN routes with RT1 and RT2, while both PE1 and PE2 import VPN routes with RT1, and PE1 also imports VPN routes with RT2, the process of excessive VPN routes will influence performance of VRF on PEs. PEs and RR should have some mechanisms to identify and control the advertisement of excessive VPN routes.

In this senario, both VRF1 and VRF2 import VPN route carries RT2, which contains RD31.

On PE1, if it overflows, it will know that the RD of excessive VPN routes is RD31 during the local processing, which come from PE3 and associated with RT1 and RT2. There are different VRFs on PE1 import

the VPN routes respectively with RT1 and RT2. If PE1 trigger the RD-ORF message when VRF1 overflows, it cannot receive the VPN routes with RT2 from PE3. The local determination of the PE can be used to inhibit the PE from sending RD-ORF entries. PE1 will not trigger the RD-ORF message until all VPNs that import VPN routes with RD31 are overflowed. When RD-ORF mechanisms is triggered, PE1 will discard the overflowed VPN routes locally and send RD-ORF entry to RR, and RR withdraws and stops to advertise such excessive VPN routes to PE1.

On PE2, due to there is only one VRF imports VPN routes with RT1. If it overflows, it will trigger RD-ORF(RD31) mechanisms. RR will withdraw and stop to advertise such excessive VPN routes to PE2.

4.1.3. Scenario-2 and Solution (Universal RD)

In this scenario, RD is allocated per VPN. One/Multiple RTs are associated with such VPN routes, and be imported into different VRFs in other devices. We assume the network topology is shown in Figure 3.

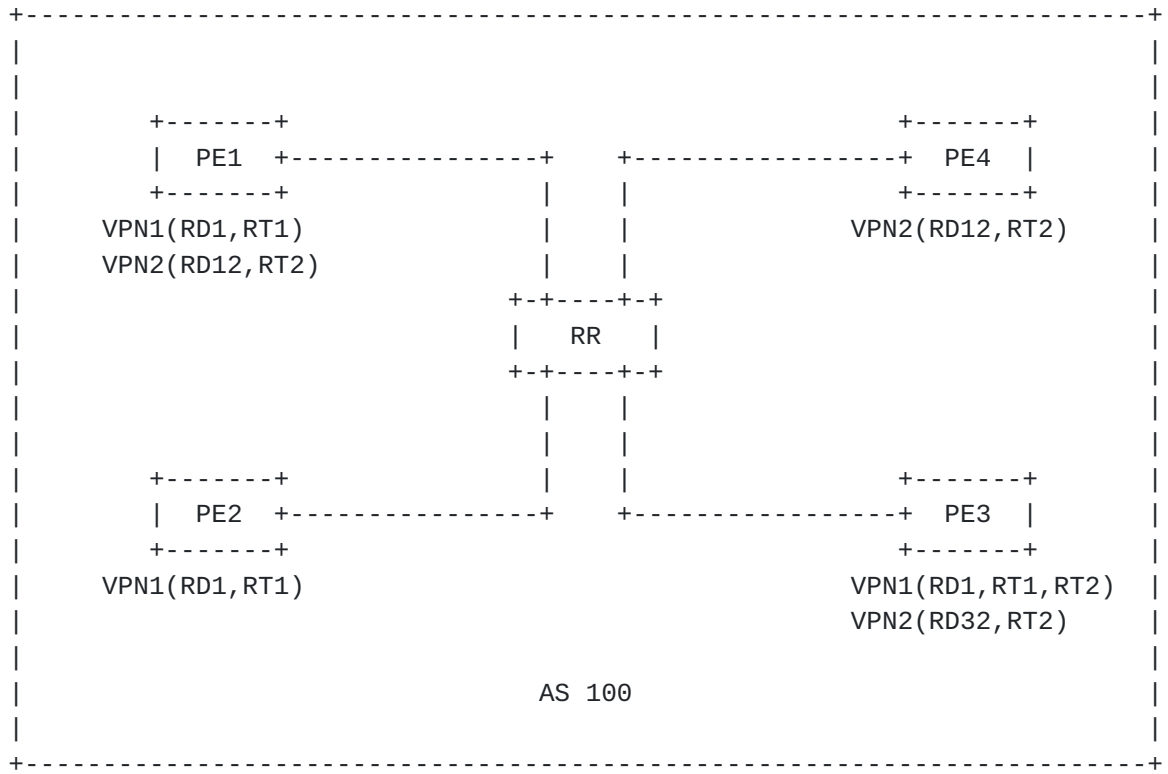

Figure 3 Network Topology of Scenario-3

When PE3 sends excessive VPN routes with RD1 and attached RT1 and RT2, while both PE1 and PE2 import VPN routes with RT1, the process of excessive VPN routes will influence performance of VRF on PEs.

PEs and RR should have some mechanisms to identify and control the advertisement of excessive VPN routes.

Based on previous principle, when PE2 overflows and PE1 not. PE2 triggers the RD-ORF message(RD1, comes from PE3). RR will withdraw and stop to advertise such excessive VPN routes to PE2. The communication between PE2 and PE1 for VPN1 will not be influenced.

5. Operation process of RD-ORF mechanism on receiver

The receiver of RD-ORF entries may be a RR or PE. As it receives the RD-ORF entries, it will checks <AFI/SAFI, ORF-Type, Sequence, Route Distinguisher> to find if it already existed in its ORF-Policy table. If not, the receiver will add the RD-ORF entries into its ORF-Policy table; otherwise, the receiver will discard it. Before the receiver send a VPN route, it will check its ORF-Policy table whether there is a related RD-ORF entry or not. If not, the receiver will send this VPN route; otherwise, the receiver will stop sending that VPN route to its peer.

6. Withdraw of RD-ORF entries

When the RD-ORF mechanism is triggered, the alarm information will be generated and sent to the network operators. Operators should manually configure the network to resume normal operation. Due to devices can record the RD-ORF entries sent by each VRF, operators can find the entries needs to be withdrawn, and trigger the withdraw process as described in [RFC5291] manually. After returning to normal, the device sends withdraw ORF entries to its peers who have previously received ORF entries.

7. RD-ORF Encoding

In this section, we defined a new ORF type called Route Distinguisher Outbound Route Filter (RD-ORF). The ORF entries are carried in the BGP ROUTE-REFRESH message as defined in [RFC5291]. A BGP ROUTE-REFRESH message can carry one or more ORF entries. The ROUTE-REFRESH message which carries ORF entries contains the following fields:

- AFI (2 octets)
- SAFI (1 octet)
- When-to-refresh (1 octet): the value is IMMEDIATE or DEFER
- ORF Type (1 octet)
- Length of ORF entries (2 octets)

A RD-ORF entry contains a common part and type-specific part. The common part is encoded as follows:

- Action (2 bits): the value is ADD, REMOVE or REMOVE-ALL
- Match (1 bit): the value is PERMIT or DENY
- Reserved (5 bits)

RD-ORF also contains type-specific part. The encoding of the type-specific part is shown in Figure 4.

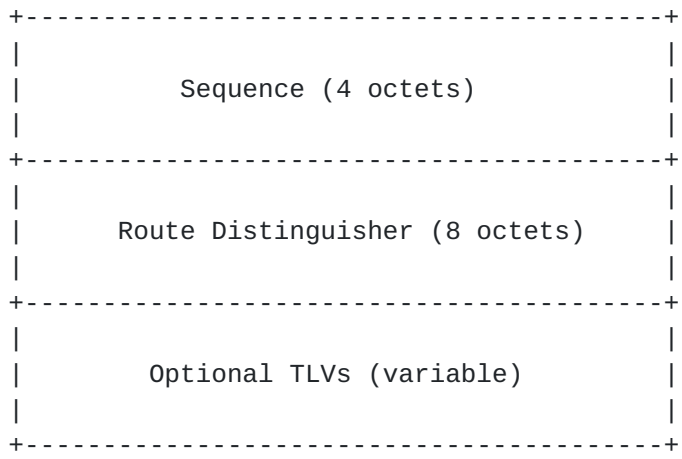

Figure 4: RD-ORF type-specific encoding

- Sequence: identifying the order in which RD-ORF is generated
- Route Distinguisher: distinguish the different user routes. The RD-ORF filters the VPN routes it tends to send based on Route Distinguisher.
- Optional TLVs: carry the potential additional information to give the extensibility of the RD-ORF mechanism.

Note that if the Action component of an ORF entry specifies REMOVE-ALL, the ORF entry does not include the type-specific part. The Sequence can uniquely identifies an RD-ORF entry. All VRFs share the sequence field, and the corresponding sequence of RD-ORF sent by each VRF will be recorded on the device.

When the BGP ROUTE-REFRESH message carries RD-ORF entries, it must be set as follows:

- The ORF-Type MUST be set to RD-ORF.

- The AFI MUST be set to IPv4, IPv6, or Layer 2 VPN (L2VPN).
- If the AFI is set to IPv4 or IPv6, the SAFI MUST be set to MPLS-labeled VPN address.
- If the AFI is set to L2VPN, the SAFI MUST be set to BGP EVPN.
- The Match field MUST be equal to DENY.

7.1. Source PE TLV

Source PE TLV is defined to identify the source of the VPN routes. Using source PE and RD to filter VPN routes together can achieve more refined route control. The source PE TLV contains the following types:

- In single-domain or Option C cross-domain scenario, NEXT_HOP attribute is fixed during routing transmission, so it can be used as source address.

  Type = 1, Length = 4 octets, value = NEXT_HOP.

  Type = 2, Length = 16 octets, value = NEXT_HOP.

- In Option B or Option AB cross-domain scenario, NEXT_HOP attribute may be changed by ASBRs and cannot be used as the source address. The originator can be traced by the Route Origin Community in BGP (as defined in Section 5 of [RFC4360]).

  Type = 3, Length = 6 octets, value = the value field of Route Origin Community.

8. Security Considerations

A BGP speaker will maintain the RD-ORF entries in an ORF-Policy table, this behavior consumes its memory and compute resources. To avoid the excessive consumption of resources, [RFC5291] specifies that a BGP speaker can only accept ORF entries transmitted by its interested peers.

9. IANA Considerations

This document defines a new Outbound Route Filter type - Route Distinguisher Outbound Route Filter (RD-ORF). The code point is from the "BGP Outbound Route Filtering (ORF) Types". It is recommended to set the code point of RD-ORF to 66.

This document also define a RD-ORF TLV type under "Border Gateway Protocol (BGP) Parameters", three TLV types are defined:

Registry	Type	Meaning
IPv4 Source PE TLV	1	IPv4 address for source PE.
IPv6 Source PE TLV	2	IPv6 address for source PE.
ROC Source PE TLV	3	Route Origin Community for Source PE.

Figure 5: IANA Allocation for newly defined TLVs

10. Acknowledgement

Thanks Robert Raszuk, Jim Uttaro, Jakob Heitz, Jeff Tantsura, Rajiv Asati, John E Drake, Gert Doering, Shuanglong Chen, Enke Chen and Srihari Sangli for their valuable comments on this draft.

11. Normative References

[I-D.ietf-bess-evpn-inter-subnet-forwarding]
Sajassi, A., Salam, S., Thoria, S., Drake, J. E., and J. Rabadan, "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-15 (work in progress), July 2021.

[I-D.wang-idr-vpn-routes-control-analysis]
Wang, A., Wang, W., Mishra, G. S., Wang, H., Zhuang, S., and J. Dong, "Analysis of VPN Routes Control in Shared BGP Session", draft-wang-idr-vpn-routes-control-analysis-04 (work in progress), September 2021.

[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>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[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>.

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <https://www.rfc-editor.org/info/rfc5291>.

[RFC5292] Chen, E. and S. Sangli, "Address-Prefix-Based Outbound Route Filter for BGP-4", RFC 5292, DOI 10.17487/RFC5292, August 2008, <https://www.rfc-editor.org/info/rfc5292>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

Authors' Addresses

Wei Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: weiwang94@foxmail.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangaj3@chinatelecom.cn

Haibo Wang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: rainsword.wang@huawei.com

Gyan S. Mishra
Verizon Inc.
13101 Columbia Pike
Silver Spring MD 20904
United States of America

Phone: 301 502-1347
Email: gyan.s.mishra@verizon.com

Shunwan Zhuang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: zhuangshunwan@huawei.com

Jie Dong
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: jie.dong@huawei.com