Workgroup: IDR Working Group
Internet-Draft:
draft-smn-idr-inter-domain-ibgp-00
Updates: 4364, 4456 (if approved)
Published: 9 January 2023
Intended Status: Standards Track
Expires: 13 July 2023
Authors: K. Szarkowicz, Ed. I. Means M. Nayman
Juniper Networks ATT Juniper Networks

Interconnecting domains with Multiprotocol IBGP

Abstract

This document relaxes the constraints specified in [RFC4364] and [RFC4456] allowing the building of Inter-domain L3VPN architecture with Multiprotocol internal BGP.

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 13 July 2023.

Table of Contents

1. Introduction

Service provides must often partition (or divide) the large network into smaller domains. This might be required for various reasons; for example:

- Separate geographic brown field networks: region 1, region 2, region 3 etc, for management or administrative purposes

- Avoid advertising unnecessary routes from domain 1 to domain 2 to improve network scale of PE (Provider Edge) nodes and RR (Route Reflector) per region

- Avoid advertising remote PE nodes loopback between regions, only DBR (Domain Boundary Router) nodes will advertise routes between regions using 'next-hop self' mechanism

The advantage of dividing the large network into smaller domains can be numerous, with important examples like:

- Per domain IGP (Interior Gateway Protocol) reduces blast radius during IGP errors or failures

- Per domain RR reduces the blast radius and BGP message exchange when RR fails

At the same time, dividing the network can be impactful and result in unwanted behavior for both the operator and its customers. For example, some BGP (Border Gateway Protocol) attributes, such as LOCAL_PREF, are not sent to the EBGP (external BGP) peers but are sent to IBGP (internal BGP) peers. Also, depending on the actual requirements, operators can selectively choose, if they keep originator NEXT_HOP attribute or change the NEXT_HOP attribute to

some local address. Further, Constrained Route Distribution ([RFC4684]) can be used to prevent DBR from sending VPN (Virtual Private Network) prefixes for VRFs (Virtual Routing and Forwarding instances) that are not locally attached to each region.

[RFC4364], in Section 10, describes three multi-domain L3VPN (Layer 3 Virtual Private Network) architectures - commonly referenced as Option 10A, Option 10B, and Option 10C - in the context of different AS (Autonomous Systems), therefore, these architectures use EBGP peerings between the domains. However, many operators might divide the network into multiple domains with one AS number used across these domains. This implies IBGP peers between domains. In multi-domain architecture there might be a need to modify the NEXT_HOP path attribute at the domain boundary. While this is the default behavior for EBGP ([RFC4271], Section 5.1.3.), it is not recommended behavior for IBGP ([RFC4456], Section 10, recommends keeping NEXT_HOP path attribute unmodified when reflecting the NLRIs - Network Layer Reachability Information - between IBGP peers).

This document relaxes these constraints specified in [RFC4271] and [RFC4364], allowing the building of Inter-domain L3VPN architectures with MP-IBGP (Multiprotocol internal BGP).

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. Inter-domain L3VPN Option 10A with MP-IBGP

Inter-domain L3VPN architecture based on so called Option 10A ([RFC4364], Section 10, "a)" bullet point) relies on multiple logical interfaces (typically, sub-interfaces with unique VLAN - Virtual Local Area Network - per sub-interface) and multiple single-hop external BGP (SH-EBGP) peerings (single peering per sub-interface) between ASBRs (autonomous system boundary router), in an architecture as outlined in Figure 1. Each SH-EBGP peering is responsible for exchanging unicast IPv4 (AFI/SAFI=1/1) or unicast IPv6 (AFI/SAFI=2/1) NLRIs for single L3VPN service. Essentially, in this architecture ASBRs consider each other as CE (Customer Edge) devices. RRs within each AS depicted in Figure 3 are optional - depending on the scalability requirements within each AS, multi-hop internal BGP (MH-IBGP) peerings could be directly established between PEs and ASBRs.

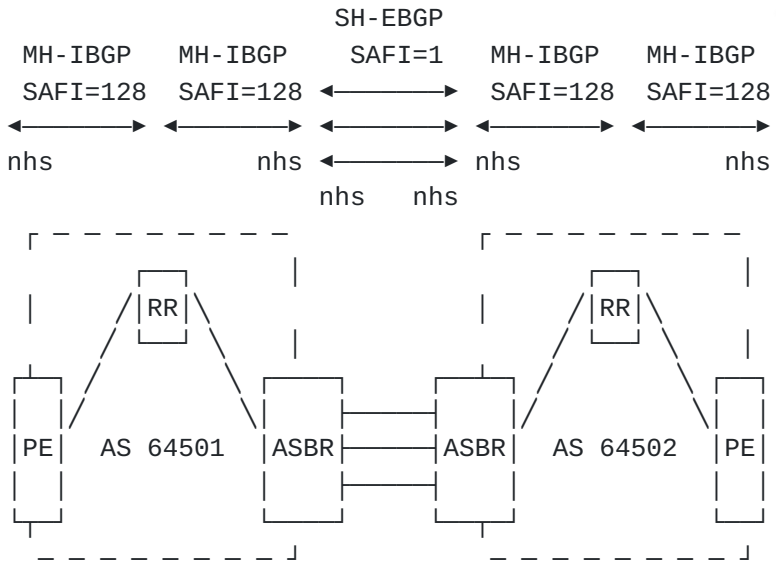

Figure 1: Inter-AS L3VPN Option 10A

This architecture does not require an end-to-end LSP (label switched path) leading from a packet's ingress PE in one AS to its egress PE in another AS, as the user packets exchanged between ASBRs are native IP (no MPLS - Multiprotocol Label Switching - encapsulation) packets. Hence, each ASBR has potentially multiple L3VPN service instances, and performs MPLS encapsulation/decapsulation. At the control plane level, each ASBR performs conversion between labeled unicast VPN-IPv4/VPN-IPv6 (SAFI=128) and unicast IPv4/IPv6 (SAFI=1) NLRIs. When these NLRIs are advertised by ASBR, NEXT_HOP attribute MUST be modified to self (nhs).

In the original context described in [RFC4364], domains are different ASs, therefore, multiple BGP peerings between two domains are EBGP. However, Option 10A concept can be applied not only to domains with different AS values, but as well as to domains with the same AS value, as depicted in Figure 2.

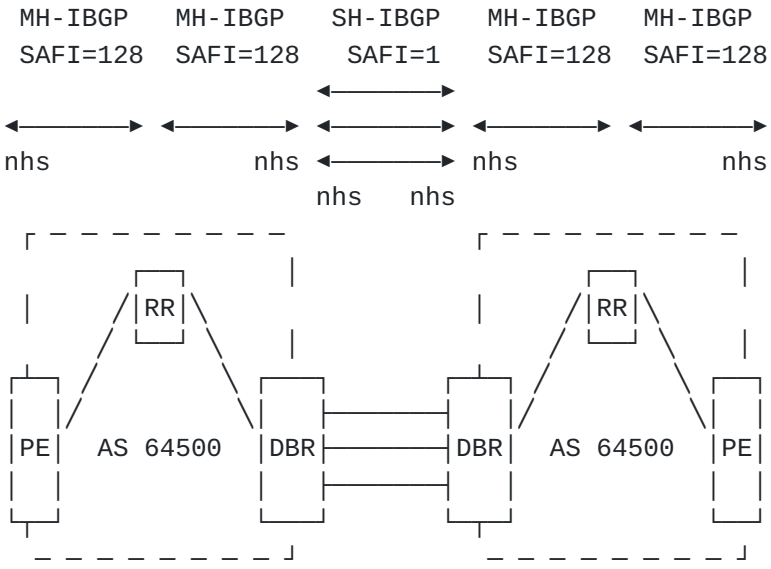

Figure 2: Inter-Domain L3VPN Option 10A using IBGP

The only difference compared to the original Inter-domain Option 10A is the peering between two domains: now, it is IBGP, and no longer EBGP. All other aspects of the architecture are unchanged.

3. Inter-domain L3VPN Option 10B with MP-IBGP

Inter-domain L3VPN architecture based on so called Option 10B ([RFC4364], Section 10, "b)" bullet point) relies on exchanging labeled unicast VPN-IPv4 (AFI/SAFI=1/128) or labeled unicast VPN-IPv6 (AFI/SAFI=2/128) NRLIs via direct SH-EBGP peering between ASBRs, in an architecture as outlined in Figure 3. RRs within each AS depicted in Figure 3 are optional - depending on the scalability requirements within each AS, MH-IBGP peerings could be directly established between PEs and ASBRs.

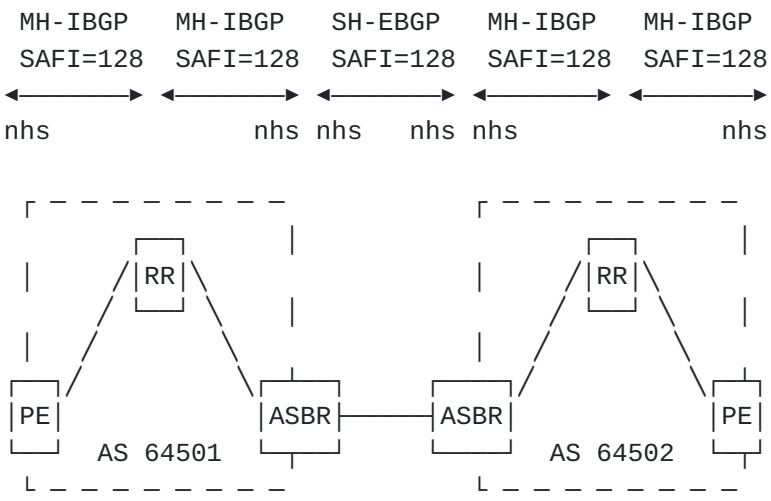

Figure 3: Inter-AS L3VPN Option 10B

This architecture requires an end-to-end LSP leading from a packet's ingress PE in one AS to its egress PE in another AS. Hence, at each ASBR, NEXT_HOP attribute MUST be modified to self (nhs), which results in new service label allocation, and programing of appropriate label forwarding entries in the data plane. On the ASBR-to-ASBR link between two ASs there is no additional 'labeled transport' (i.e., no LDP - Label Distribution Protocol, RSVP - Resource Reservation Protocol, SR - Segment Routing, ...) protocol - the packets are transmitted on the ASBR-to-ASBR link with single L3VPN service label.

In the original context described in [RFC4364], domains are different ASs, therefore, the BGP peering between two domains is EBGP. However, Option 10B concept can be applied not only to domains with different AS values, but also to domains with the same AS value, as depicted in Figure 4.

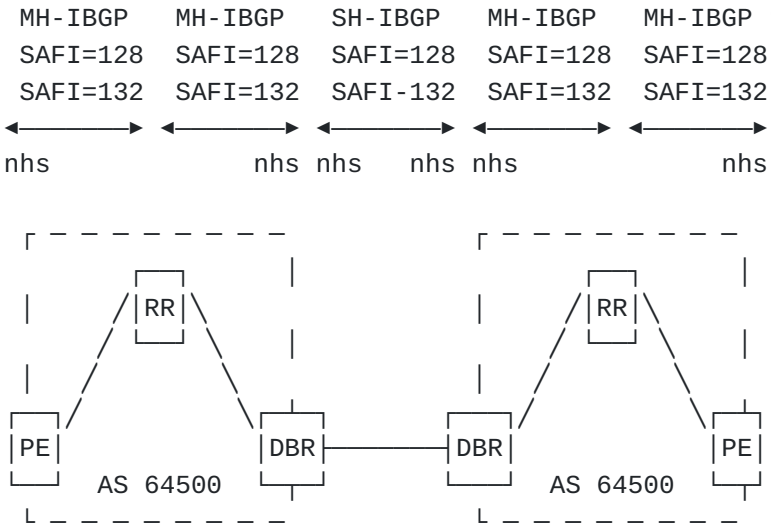

Figure 4: Inter-Domain L3VPN Option 10B using IBGP, with separate DBRs

The only difference compared to the original Inter-domain Option 10B is the peering between two domains: now, it is IBGP, and no longer EBGP. All other aspects of the architecture are unchanged. This implies that DBR becomes inline (on the path) RR for labeled VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs, and MUST change the NEXT_HOP attribute to self, when reflecting these NLRIs. This is not in accordance with [RFC4364], Section 10 recommendation that RR SHOULD NOT modify the NEXT_HOP attribute. therefore, this document updates [RFC4364] by defining the use case, where RR MUST modify the NEXT_HOP attribute, when reflecting NRLIs over IBGP peerings.

It is strongly advisable to control the exchange of labeled unicast VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs between domains via Constrained Route Distribution ([RFC4684]). Therefore, DBR-to-DBR SH-IBGP peering, in addition to SAFI=128, SHOULD include Route Target Constraint - RTC (SAFI=132) - as well, and DBRs SHOULD be provisioned to exchange between each other only desired RTCs. Please note, RTC MAY be used inside of each domain, too, to control route distribution within domains.

When using IBGP, instead of EBGP, small variation of the architecture can be achieved, by collapsing two separate DBRs to single, collapsed DBR, as depicted in Figure 5.

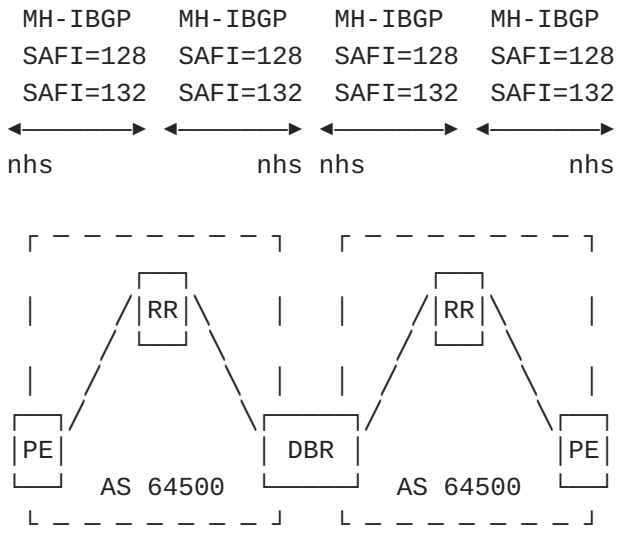

Figure 5: Inter-Domain L3VPN Option 10B using IBGP, with collapsed DBR

Similarly to the previous example, DBR MUST change the NEXT_HOP attribute to self, when reflecting labeled unicast VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs, and DBR SHOULD use RTC (SAFI=132) to control the exchange of labeled unicast VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs between domains. RTC MAY be used inside of each domain.

4. Inter-domain L3VPN Option 10C with MP-IBGP

Inter-domain L3VPN architecture based on so called Option 10C ([RFC4364], Section 10, "c)" bullet point) relies on exchanging labeled unicast VPN-IPv4 (AFI/SAFI=1/128) or labeled unicast VPN-IPv6 (AFI/SAFI=2/128) NLRIs via MH-EBGP peering between domains, without changing the NEXT_HOP attribute, and exchanging labeled unicast IPv4 or labeled unicast IPv6 (SAFI=4) host routes (PE loopbacks) via direct SH-EBGP peering between ASBRs, changing the NEXT_HOP attribute at the domain boundaries, in an architecture as outlined in Figure 6. As in previous architectures, RRs within each AS depicted in Figure 6 are optional. However, given the fact that

one of the main objectives of Option 10C architecture is to offload ASBRs from the task of maintaining/distributing labeled unicast VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs, without RR these NLRIs would need to be distributed via direct MH-EBGP peerings between PEs from different domains. Such approach makes the design very impractical and not scalable, therefore, in Option 10C RRs SHOULD be deployed, and MH-EBGP peerings to distribute labeled unicast VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs between domains SHOULD be established between RRs.

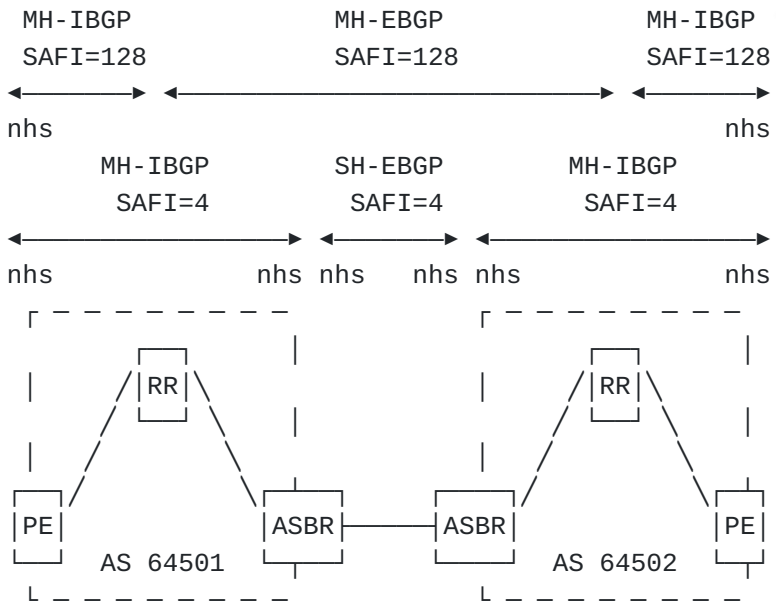

Figure 6: Inter-AS L3VPN Option 10C

This architecture requires an end-to-end LSP leading from a packet's ingress PE in one AS to its egress PE in another AS. Hence, at each ASBR, NEXT_HOP attribute for labeled unicast IPv4 or labeled unicast IPv6 (SAFI=4) NLRI MUST be modified to self (nhs), which results in new transport label allocation, and programming of appropriate label forwarding entries in the data plane. In the packets traversing ASBR-to-ASBR link between two ASs, similar to the links within each AS, there is additional transport label at the top of the label stack in addition to the L3VPN service label. This transport label is exchanged via BGP peering with SAFI=4.

In the original context described in [RFC4364], domains are different autonomous systems, therefore, the BGP peerings (both for SAFI=4 and SAFI=128) between two domains are EBGP. However, Option 10C concept can be applied not only to domains with different AS values, but as well to domains with the same AS value, as depicted in Figure 7.

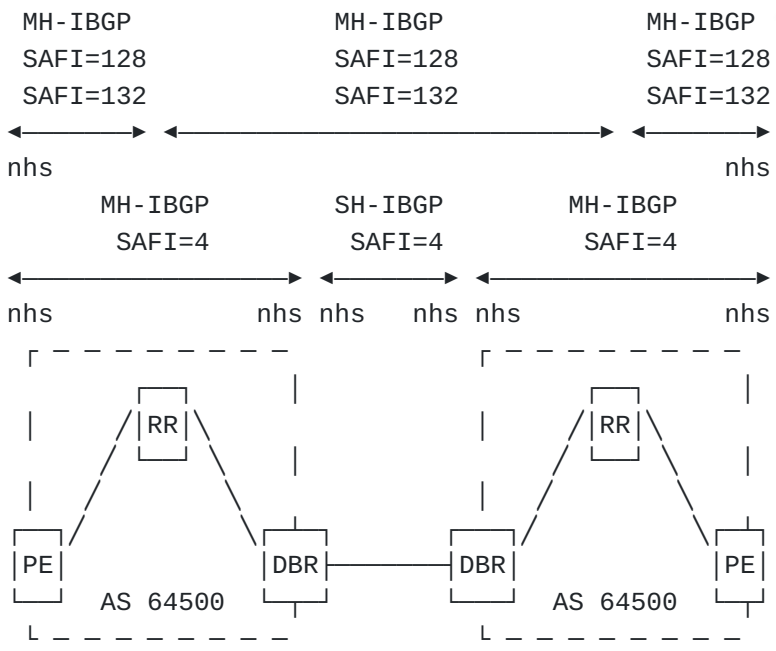

Figure 7: Inter-Domain L3VPN Option 10C using IBGP, with separate DBRs

Again, the only difference compared to the original Inter-domain Option 10C is the peering between two domains: now, it is IBGP, and no longer EBGP, for both single-hop BGP peering used to exchange labeled unicast IPv4 or labeled unicast IPv6 (SAFI=4) host routes (PE loopbacks), as well as multi-hop BGP peering used to exchange labeled unicast VPN-IPv4 (AFI/SAFI=1/128) or labeled unicast VPN-IPv6 (AFI/SAFI=2/128) NLRIs. All other aspects of the architecture are unchanged. This implies that domain boundary router (DBR) becomes inline (on the path) RR for labeled unicast IPv4 or labeled unicast IPv6 (SAFI=4) NLRIs, and MUST change the NEXT_HOP attribute to self, when reflecting these NLRIs. Again, this is not in accordance with [RFC4364], Section 10 recommendation that RR SHOULD NOT modify the NEXT_HOP attribute. therefore, this document updates [RFC4364] by defining the use case, where RR MUST modify the NEXT_HOP attribute.

As in Option 10B scenario, it is strongly advisable to control the exchange of labeled unicast VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs between domains via Constrained Route Distribution ([RFC4684]). Therefore, MH-IBGP peering between RRs in different domains, in addition to SAFI=128, SHOULD include RTC (SAFI=132), and RRs SHOULD be provisioned to exchange between each other only desired RTCs. Please note, RTC MAY be used inside of each domain, too, to control route distribution within domains.

When using IBGP, instead of EBGP, a small variation of the architecture can be achieved, by collapsing two separate DBRs to single, collapsed DBR, as depicted in Figure 8.

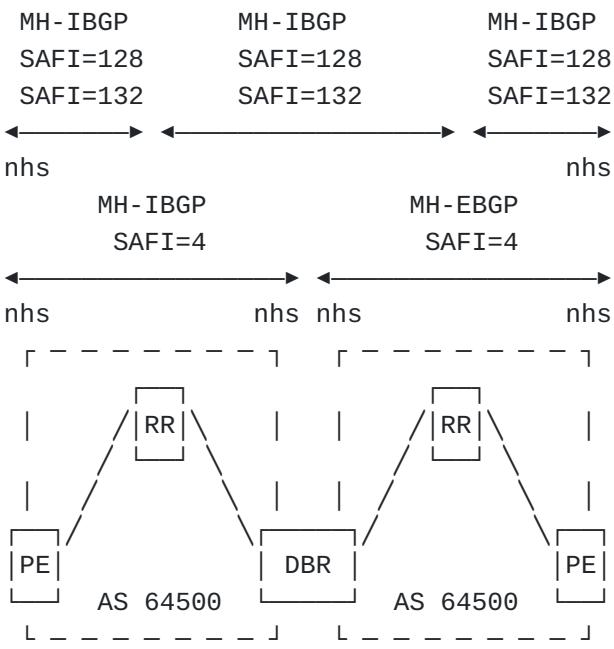

Figure 8: Inter-Domain L3VPN Option 10C using IBGP, with collapsed DBR

Similarly to the previous example, DBR MUST change the NEXT_HOP attribute to self, when reflecting labeled unicast IPv4 or labeled unicast IPv6 (SAFI=4) NLRIs, and RR SHOULD use RTC (SAFI=132) to control the exchange of labeled unicast VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs between domains. RTC MAY be used inside of each domain.

5. IANA Considerations

This memo includes no request to IANA.

6. Security Considerations

To be added later.

7. References

7.1. Normative References

[RFC2119] Bradner, S. and RFC Publisher, "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>.

[RFC8174] Leiba, B. and RFC Publisher, "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>.

7.2. Informative References

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

[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>.

[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>.

Appendix A. Acronyms and Abbreviations

AFI: Address Family Identifier

AS: Autonomous System

ASBR: Autonomous System Boundary Router

BGP: Border Gateway Protocol

CE: Customer Edge

DBR: Domain Boundary Router

EBGP: External Border Gateway Protocol

IBGP: Internal Border Gateway Protocol

IGP: Interior Gateway Protocol

IP: Internet Protocol

IPv4: Internet Protocol version 4

IPv6: Internet Protocol version 6

L3VPN: Layer 3 Virtual Private Network

LDP: Label Distribution Protocol

LSP: Label Switched Path

MH-IBGP: Multi-hop Internal Border Gateway Protocol

MP-IBGP: Multiprotocol Internal Border Gateway Protocol

MPLS: Multiprotocol Label Switching

nhs: next-hop self

NLRI: Network Layer Reachability Information

PE: Provider Edge

RR: Router Reflector

RSVP: Resource Reservation Protocol

RTC: Route Target Constraint

SAFI: Subsequent Address Family Identifier

SH-EBGP: Single-hop External Border Gateway Protocol

SR: Segment Routing

VLAN: Virtual Local Area Network

VPN: Virtual Private Network

VRF: Virtual Routing and Forwarding

Acknowledgements

To be added later

Contributors

To be added later

Authors' Addresses

Krzysztof G. Szarkowicz (editor)
Juniper Networks
Wien

Austria

Email: kszarkowicz@juniper.net

Israel Means
ATT
2212 Avenida Mara
Chula Vista, CA 91914
United States of America

Email: israel.means@att.com

Moshiko Nayman
Juniper Networks
18 Buckingham Dr
Manalapan, NJ 07726
United States of America

Email: mnayman@juniper.net