Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 26 December 2025

K. Szarkowicz, Ed.
M. Nayman, Ed.
Juniper Networks
I. Means
AT&T

24 June 2025

Interconnecting domains with IBGP
draft-smn-idr-inter-domain-ibgp-07

Abstract

This document relaxes the recommendations specified in BGP/MPLS IP Virtual Private Networks (VPNs) and BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP) allowing the building of Inter-domain L3VPN architecture with 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 26 December 2025.

Table of Contents

1. Introduction

Service provides must often partition (or divide) a large network into smaller IGP (Interior Gateway Protocol) domains which are interconnected via BGP (Border Gateway Protocol) only. 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 IGP 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 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 - restricted to the use cases, where the domains are distinct BGP domains and use different AS (Autonomous System) numbers, therefore, these architectures use EBGP peerings between the domains. However, many operators might segment the network into multiple IGP domains while maintaining a single BGP domain, with one AS number used across the IGP domains. This is especially the case when migrating an existing network into multiple IGP domains (brownfield deployment). Operationally, it is often too complex to migrate separated domains to new AS numbers. As the result, it implies IBGP peers are used between IGP 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).

In a network scenario where domains are interconnected using IBGP, and the same BGP AS number is used across multiple domains, the Accumulated IGP AIGP metric (as specified in [RFC7311]) remains preserved across domain boundaries. This preservation is due to IBGP sessions maintaining internal BGP attributes, including AIGP, ensuring the exptected path selection based on accumulated internal IGP metrics.

This document relaxes these recommendations specified in [RFC4364] and [RFC4456], allowing Inter-domain L3VPN architectures stitching multiple IGP domains with IBGP.

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 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 SHOULD be used. However, in small scale domains (for example, small access rings with few PEs), RR function could be placed on ASBRs, where multi-hop internal BGP (MH-IBGP) peerings are directly established between PEs and ASBRs.

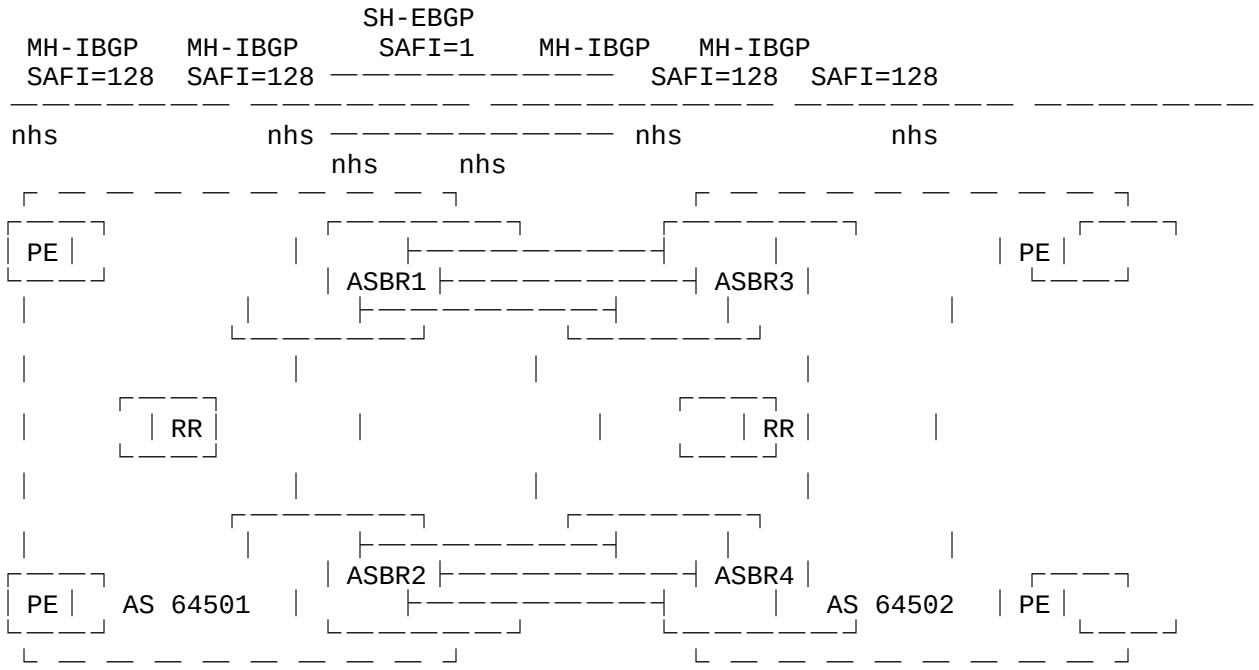

Figure 1: Inter-AS L3VPN Option 10A

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, which is typical PE function. At the control plane level, each ASBR performs conversion between 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).

When reflecting routes from DBRs in Domain A to Domain B, the CLUSTER_LIST attribute SHOULD be enhanced during SAFI transitions. When Domain A routes learned within SAFI-128 are advertised over SAFI-1 to the DBR node in Domain B, the locally configured CLUSTER_ID on the DBR node MUST be appended to the CLUSTER_LIST. This ensures a clear path history, which is important for preventing routing loops. An OPTIONAL technique is to use region-specific community tagging. Region A (DBR1) tags all routes received from its local RR with a region-specific community attribute, using a specific value as the region identifier (e.g., name COMM-DOMAIN-A with a value 1234:4321) to all routes received from its local RR. This tagging helps identify routes originating from Region A. Before advertising these routes to Region B (DBR3), DBR1 ensures that the COMM-DOMAIN-A community is attached. Upon receiving routes from DBR3, DBR1 imports all routes but rejects any with the COMM-DOMAIN-A community. This prevents the re-advertisement of the same routes back into Region A, thereby avoiding routing loops.

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

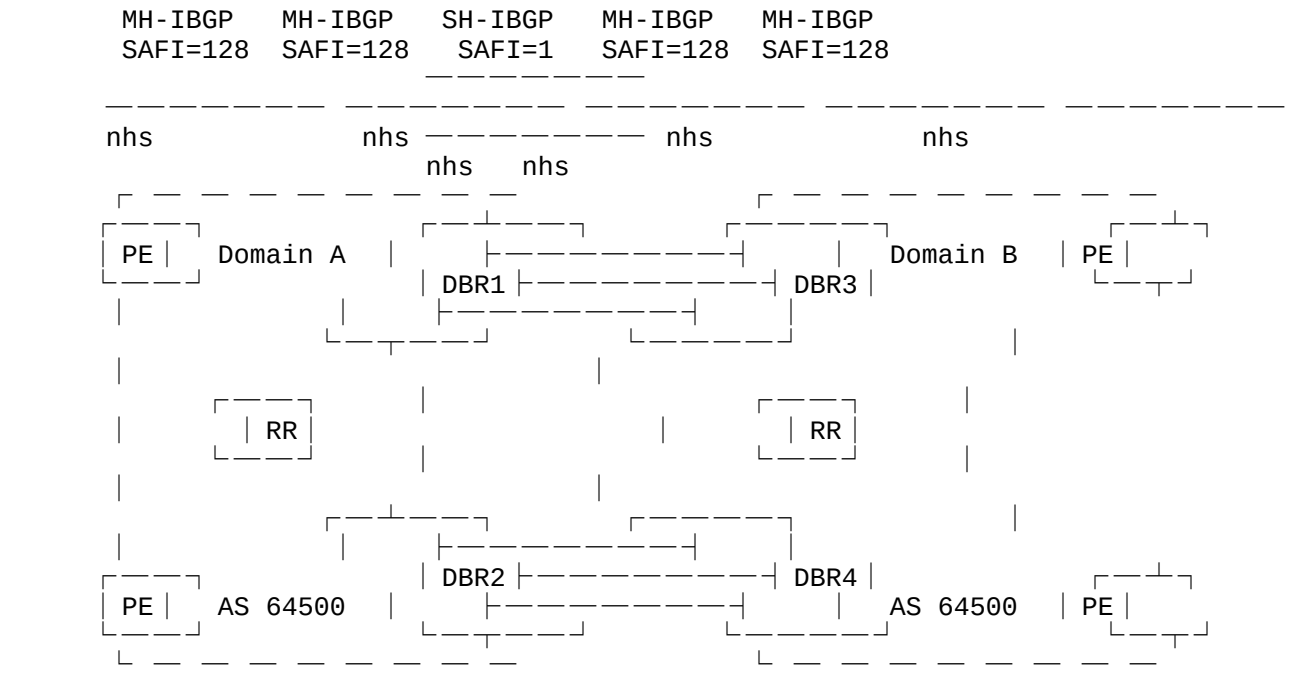

Figure 2: Inter-Domain L3VPN Option 10A using IBGP

The main differences, compared to the original Inter-domain Option 10A, are:

* the BGP peering between two IGP domains are now IBGP (SAFI=1), and no longer EBGP (SAFI=1)

* DBRs become PEs with all BGP peerings (global and inside VRFs) using the same AS number

* To prevent routing loops, DBRs in each domain will add a domain-specific community to routes before exporting them to the other domain. They will block any routes received from the other domain that contain the same community.

Other aspects of the architecture are similar.

3. Inter-domain L3VPN Option 10B with IBGP

Inter-domain L3VPN architecture based on so called Option 10B ([RFC4364], Section 10, "b)" bullet point) relies on exchanging VPN-IPv4 (AFI/SAFI=1/128) or 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 SHOULD be used. However, in small scale domains (for example, small access rings with few PEs), RR function could be placed on ASBRs, where multi-hop internal BGP (MH-IBGP) peerings are directly established between PEs and ASBRs.

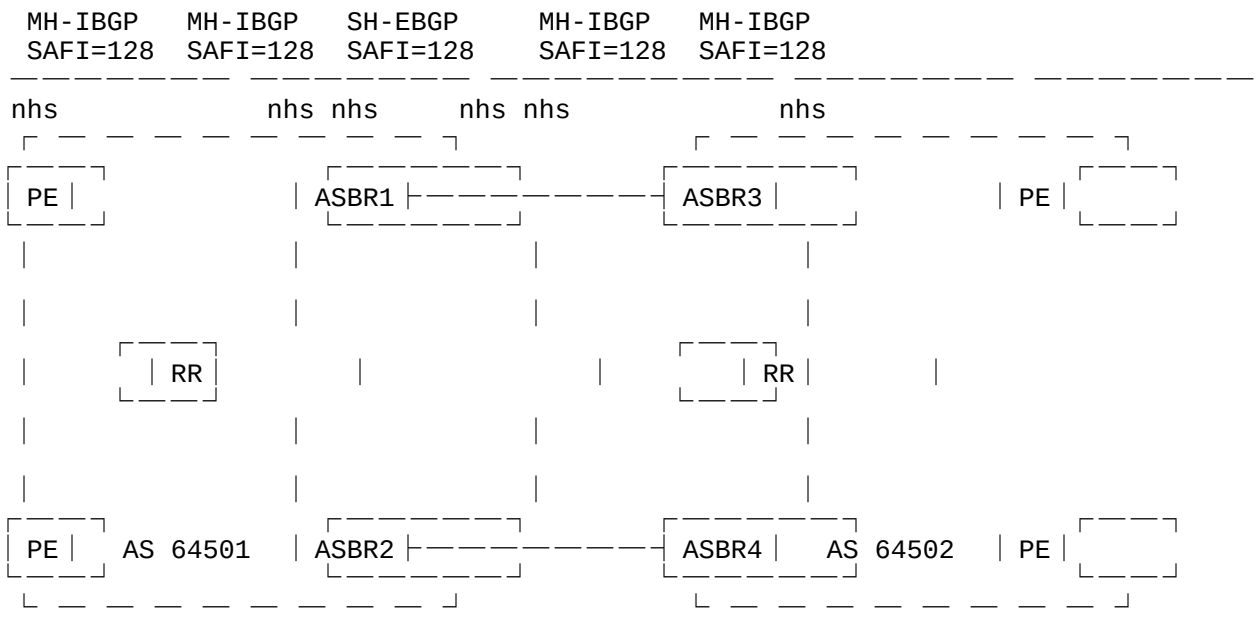

Figure 3: Inter-AS L3VPN Option 10B

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 BGP domains with different ASs, therefore, the BGP peering between two BGP domains is EBGP. However, Option 10B concept can be applied not only to BGP domains with different AS numbers, but also to IGP domains with the same AS number, as depicted in Figure 4.

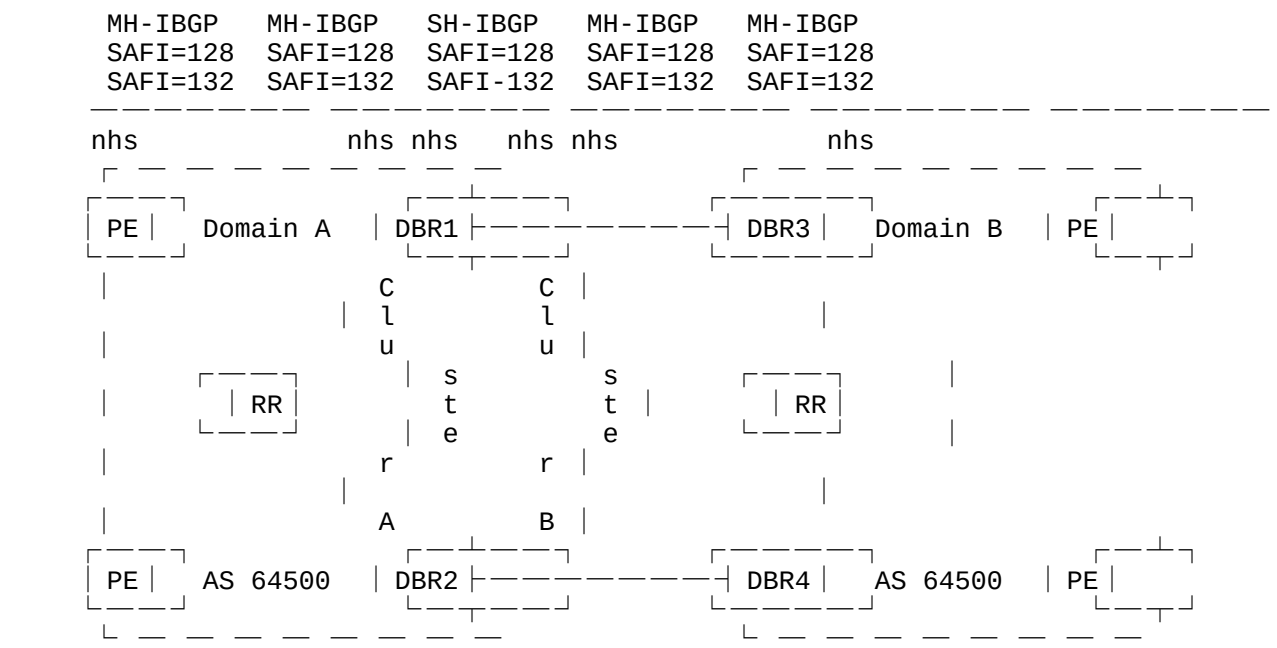

Figure 4: Inter-Domain L3VPN Option 10B using IBGP, with separate DBRs

Similar to the Inter-Domain Option 10A case, main differences, when comparing to the original Inter-domain Option 10B are:

* the BGP peering between two IGP domains are now IBGP (SAFI=128), and no longer EBGP (SAFI=128)

* DBRs become on-the-path route reflectors for SAFI=128

This implies that DBR with RR function MUST change the NEXT_HOP attribute to self, when reflecting these NLRIs. This is not in accordance with [RFC4456], Section 10 recommendation that RR SHOULD NOT modify the NEXT_HOP attribute, therefore, this document relaxes the recommendation from [RFC4456] 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 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 IGP domain, too, to control route distribution within IGP domains.

Important aspect of the inter-domain conenctivity, when the domains are interconnected via multiple interconnection points, as depicted in Figure 4, is loop prevention. In classic Inter-AS Option 10B, with different AS numbers used in each BGP domain, and EBGP peerings between BGP domains, loop prevention is ensured by rejecting updates containig local AS number in the AS_PATH attribute. In the use case with multiple IGP domains and single BGP domain, each DBR is on-the-path RR, thus is associated with a CLUSTER_ID. One option for CLUSTER_ID allocation is, that each DBR is configured with a unique CLUSTER_ID. Another option for CLUSTER_ID allocation is that each DBR pair in the IGP domain uses unique CLUSTER_ID, as depicted in Figure 4. Using the first CLUSTER_ID allocation scheme, there is a risk of a BGP routing loop occurring, since all of the DBRs are reflecting prefixes as RRs in the same AS with next-hop self. To prevent the DBRs of the same IGP domain from accepting updates from each other, they MUST use the same CLUSTER_ID. In this case, a DBR will discard a prefix update that has the same CLUSTER_ID as itself in order to prevent routing information loops in BGP. For example, DBR1 and DBR2 configured with one CLUSTER_ID, while DBR3 and DBR4 have another single CLUSTER_ID. This mechanism, loop prevention based on CLUSTER_LIST filtering, is described in [RFC4456], Section 10 - Avoiding Routing Information Loops.

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.

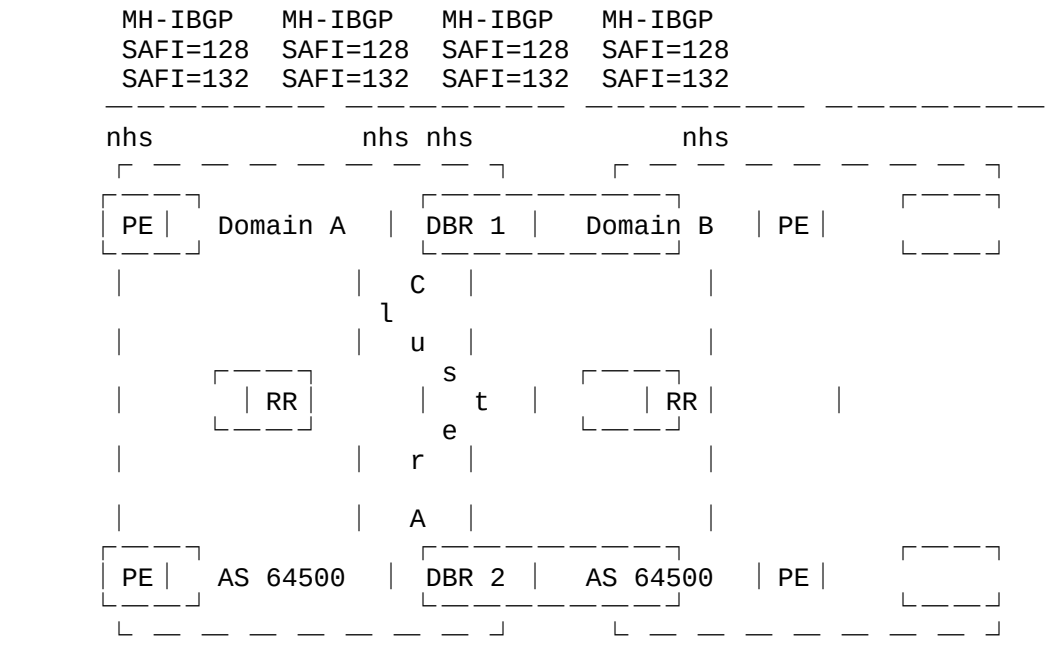

Figure 5: Inter-Domain L3VPN Option 10B using IBGP, with collapsed DBR

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

4. Inter-domain L3VPN Option 10C with IBGP

Inter-domain L3VPN architecture based on so called Option 10C ([RFC4364], Section 10, "c)" bullet point) relies on exchanging VPN-IPv4 (AFI/SAFI=1/128) or VPN-IPv6 (AFI/SAFI=2/128) NLRIs via MH-EBGP peering between BGP 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 BGP domain boundaries, in an architecture as outlined in Figure 6. As in previous architectures, RRs within each AS depicted in Figure 6 SHOULD be used. One of the main objectives of Option 10C architecture is to offload ASBRs from the task of maintaining/distributing 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 BGP 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 VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs between BGP domains SHOULD be established between RRs.

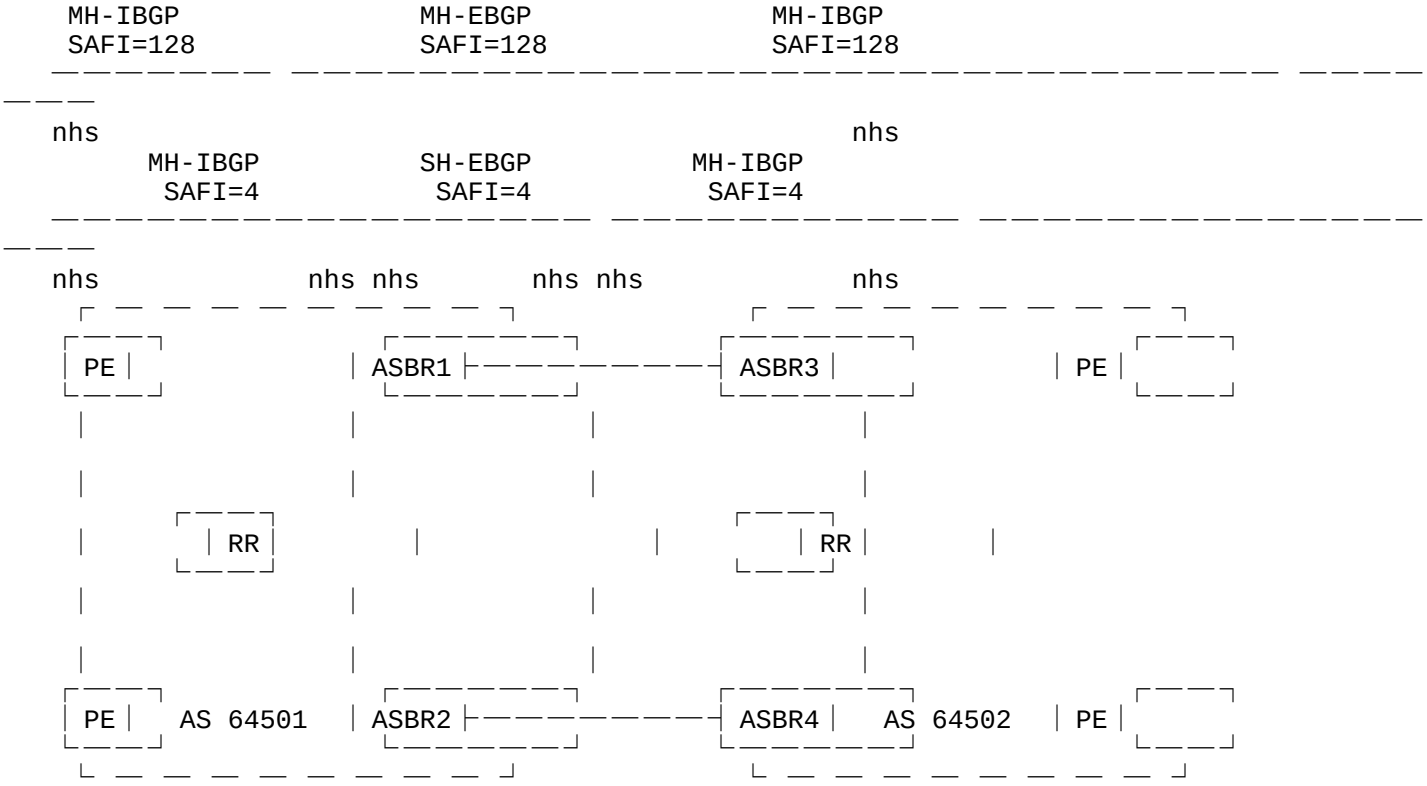

Figure 6: Inter-AS L3VPN Option 10C

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 BGP domains with different AS numbers, therefore, the BGP peerings (both for SAFI=4 and SAFI=128) between two BGP domains are EBGP. However, Option 10C concept can be applied not only to BGP domains with different AS numbers, but as well to IGP domains with the same AS number, as depicted in Figure 7.

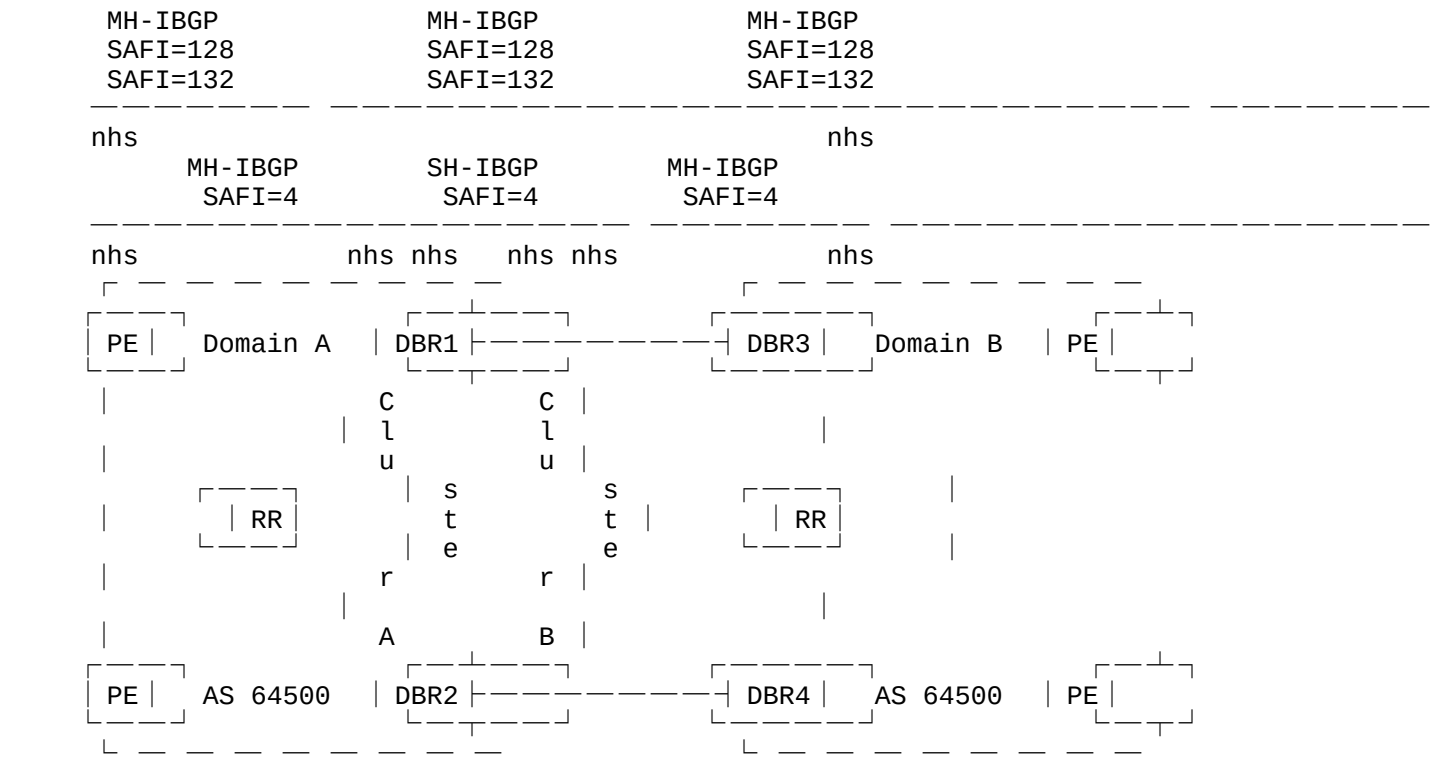

Figure 7: Inter-Domain L3VPN Option 10C using IBGP, with separate DBRs

Again, the differences compared to the original Inter-domain Option 10C are:

* the peerings between two IGP domains are now 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 VPN-IPv4 (AFI/SAFI=1/128) or VPN-IPv6 (AFI/SAFI=2/128) NLRIs

* DBRs become on-the-path route reflectors for SAFI=4

Remaining aspects of the architecture are similar. This implies that IGP 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 relaxes the recommendation from [RFC4456] by defining the use case, where RR MUST modify the NEXT_HOP attribute, when reflecting NRLIs over IBGP peerings

As in Option 10B scenario, it is strongly advisable to control the exchange of VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs between domains via Constrained Route Distribution ([RFC4684]). Therefore, MH-IBGP peering between RRs in different IGP 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 IGP 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.

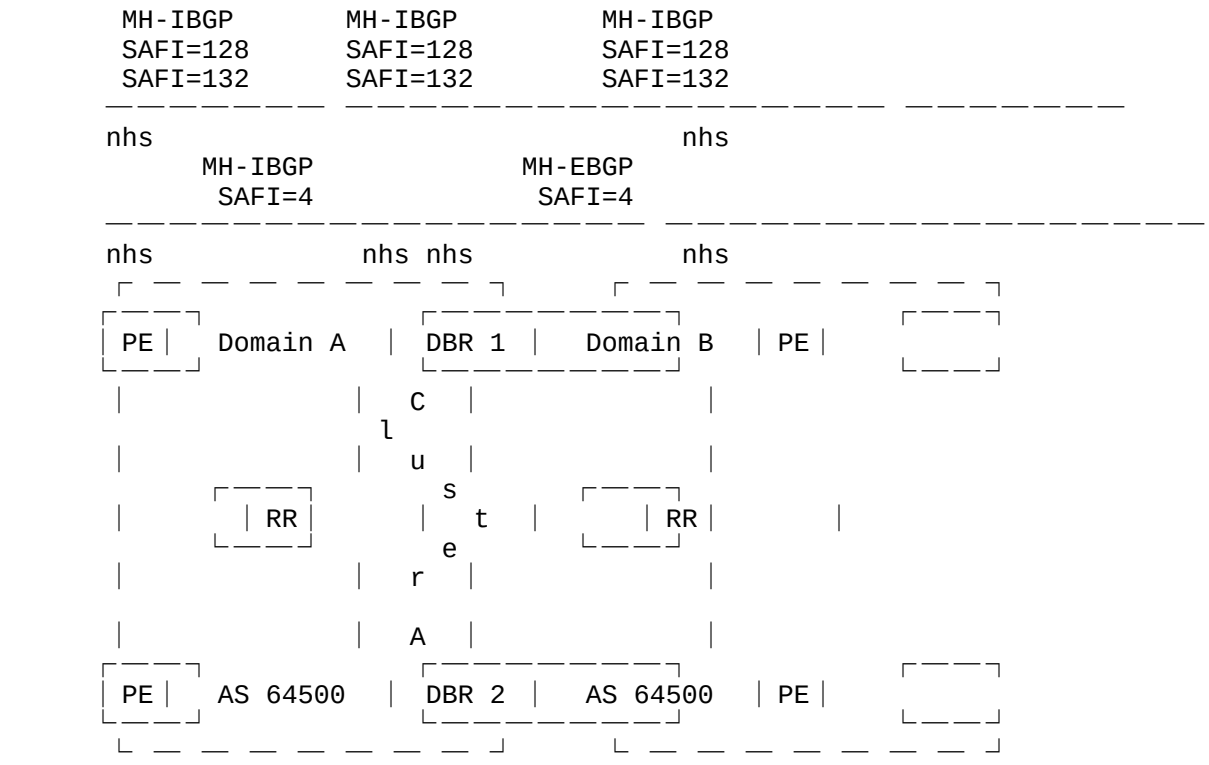

Figure 8: Inter-Domain L3VPN Option 10C using IBGP, with collapsed DBR

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 VPN-IPv4/VPN-IPv6 (SAFI=128) NLRIs between IGP domains. RTC MAY be used inside of each domain.

5. IANA Considerations

This memo includes no request to IANA.

6. Security Considerations

As a general rule, traffic spoofing between domains must be prevented. In Option A, the interfaces connecting the two domains do not transport labeled traffic; they only carry regular IP traffic. Therefore, well-established IP mechanisms, like filtering unwanted traffic, can be used to prevent spoofing.

BGP options B and C are more vulnerable due to possible unauthorized labeled forwarding or label spoofing, especially when engaging in peering arrangements with third-party DBRs. ([RFC4364]), Section 6: "Maintaining Proper Isolation of VPNs", addresses the requirements for proper isolation of VPNs. If MPLS is being used as the tunneling technology, this means that a DBR MUST NOT accept a labeled packet from any adjacent DBR unless the following two conditions hold:

* the label at the top of the label stack was actually distributed by a DBR to that adjacent DBR, and,

* the DBR can determine that use of that label will cause the packet to leave the backbone before any labels lower in the stack will be inspected, and before the IP header will be inspected.

Defining a concerete solution to address above isolation requirements is out of scope for this document. However, a possible approach to mitigate these concerns, for both Option B and Option C, is described in [I-D.kaliraj-bess-bgp-sig-private-mpls-labels], with Section 6.1 providing specific examples. With this method, distinct context table per BGP peer (or group of peers) is established. This table only contains labels advertised to specific BGP peer (group of peers). Interfaces facilitating DBR connections, where Option B or Option C is implemented, SHOULD be associated with this newly introduced context table. Consequently, the label lookup for packets received over the interface from specific DBR happens within the context of this table. Packets with labels not advertised to DBR, fail the lookup and are dropped.

Additionally, for Options B and C, an MPLS filter can be implemented, specifying inner IPv4 or IPv6, under one or more labeled packets. The number of labels in a filter is crucial for differentiating the transport and enabling the filtering of spoofed traffic originating from unwanted source IPv4 or IPv6 addresses.

7. References

7.1. 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>.

[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>.

7.2. Informative References

[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>.

[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>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

[I-D.kaliraj-bess-bgp-sig-private-mpls-labels] Vairavakkalai, K., Jeganathan, J. M., Ramadenu, P., and I. Means, "BGP Signaled MPLS Namespaces", Work in Progress, Internet-Draft, draft-kaliraj-bess-bgp-sig-private-mpls-labels-09, 9 November 2024, <https://datatracker.ietf.org/doc/html/draft-kaliraj-bess-bgp-sig-private-mpls-labels-09>.

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

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

The authors would like to thank Robert Raszuk, and Bruno Decraene for their reviews of this document and for providing valuable comments.

Contributors

To be added later

Authors' Addresses

Krzysztof G. Szarkowicz (editor)
Juniper Networks
Wien
Austria
Email: kszarkowicz@juniper.net

Moshiko Nayman (editor)
Juniper Networks
18 Buckingham Dr
Manalapan, NJ 07726
United States of America
Email: mnayman@juniper.net

Israel Means
AT&T
2212 Avenida Mara
Chula Vista, CA 91914
United States of America
Email: israel.means@att.com