

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 18, 2013

J. Dong
Z. Li
Huawei Technologies
October 15, 2012

A Framework for L3VPN Performance Monitoring
draft-dong-l3vpn-pm-framework-00

Abstract

This document specifies the framework and mechanisms for the application of performance monitoring (PM) to BGP/MPLS IP Virtual Private Networks (L3VPN).

Requirements Language

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 RFC 2119 [RFC2119].

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 <http://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 18, 2013.

Copyright Notice

Copyright (c) 2012 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	3
2. Overview and Concepts	3
2.1. VRF-to-VRF Tunnel	3
3. Control Plane	4
3.1. VPN Membership Auto-Discovery	4
3.2. VRF-to-VRF Label Allocation	4
4. Data Plane	4
4.1. Additional Label for Ingress VRF Identification	4
4.2. Replace the VPN Label with VT Label	5
5. L3VPN Performance Monitoring	5
6. IANA Considerations	6
7. Security Considerations	6
8. Acknowledgements	6
9. References	6
9.1. Normative References	6
9.2. Informative References	6
Authors' Addresses	7

1. Introduction

Level 3 Virtual Private Network (L3VPN) [RFC4364] service is widely deployed to provide enterprise VPN, Voice over IP (VoIP), video, mobile backhaul, etc. services. Most of these services are sensitive to the packet loss and delay. The capability to measure and monitor performance metrics for packet loss, delay, as well as related metrics is essential for meeting the Service Level Agreement (SLA). This measurement capability also provides operators with greater visibility into the performance characteristics of the services in their networks, and provides diagnostic information in case of performance degradation or failure and helps for fault localization.

To perform the measurement of packet loss, delay and other metrics on a particular VPN traffic flow, the egress PE needs to identify the ingress VRF sending the VPN packets. As specified in the [L3VPN-PM-ANA] document, such flow identification is a big challenge for existing L3VPN.

This document specifies the framework and mechanisms for the application of performance monitoring in L3VPN.

2. Overview and Concepts

Based on the mechanisms in [RFC4364], for a particular VPN prefix, the directly connected PE allocates the same VPN label to all the remote PEs which maintain VPN Routing and Forwarding Tables (VRFs) of that VPN. Thus performance monitoring can not be performed on the egress PE, since it is not able to identify the source VRF of the received VPN packets.

As analyzed by [L3VPN-PM-ANA], to perform the packet loss or delay measurement on a specific VPN flow, it is critical for the egress PE to identify the unique VRF, i.e. to establish the Point-to-Point connection between the two VRFs. Once the Point-to-Point connection is built up, current measurement mechanisms may be applied to L3VPN. A new concept "VRF-to-VRF Tunnel" is introduced in the following section to establish such Point-to-Point connection.

2.1. VRF-to-VRF Tunnel

In order to perform performance monitoring in L3VPN, a point-to-point connection between any two VRFs of a particular VPN needs to be established. This guarantees that the egress PE could identify the ingress VRF of the received VPN traffic, thus it could measure the packet loss and delay between the ingress and egress VRFs. Such point-to-point VPN connection between an ingress VRF and an egress

VRF is called "VRF-to-VRF Tunnel (VT)".

3. Control Plane

This section describes the control plane mechanisms needed for L3VPN performance monitoring.

3.1. VPN Membership Auto-Discovery

Before establishing the Point-to-Point connections between VRFs, each PE needs to know all the remote PEs participating in the same VPN. This can be achieved by the membership auto-discovery procedure. Some mechanisms similar to the membership auto-discovery in VPLS [RFC4761] and L2VPN [RFC6074] can be used.

3.2. VRF-to-VRF Label Allocation

After obtaining the VPN membership information, each PE needs to allocate MPLS labels to identify the VRF-to-VRF tunnel between the local VRF and the remote VRFs, such labels are called VT labels. For each local VRF, the egress PE SHOULD allocate different VT labels for each remote VRF in PEs belonging to the same VPN. This way, the egress PE could identify the VPN flow received from different ingress VRFs, and the packet loss and delay measurement could be performed between each ingress VRF and the local VRF.

4. Data Plane

This section introduces two new MPLS label stack encapsulations when VT label applies.

4.1. Additional Label for Ingress VRF Identification

When a VPN data packet needs to be sent, firstly the VPN label obtained from the BGP VPN route of the destination address prefix is pushed onto the label stack. The VT label allocated by the egress VRF should then be pushed onto the label stack to identify the Point-to-Point connection between the sending and receiving VRF. Lastly, the MPLS tunnel label is pushed onto the label stack. The TTL and COS value in the VPN label entry should be copied to the TTL and COS fields of the VT label respectively. This way, one additional label is carried in the label stack compared with L3VPN data plane in [RFC4364].

When the VPN data packet arrives at the egress PE, the outermost tunnel label is popped, then the egress PE could use the VT label to

identify the ingress VRF of the packet.

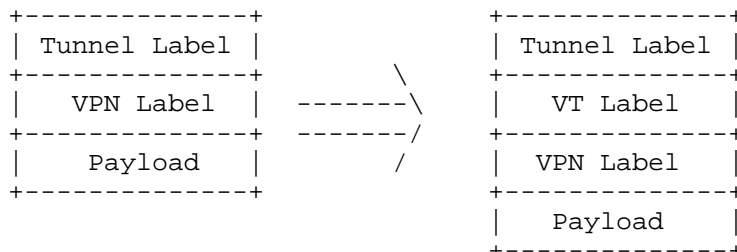


Fig.1 Additional Label for Ingress VRF Identification

4.2. Replace the VPN Label with VT Label

Since the VT label identifies the connection between the ingress VRF and egress VRF, it could also be used to identify the egress VRF table in which the VPN prefix lookup should be performed. Thus when encapsulating the VPN data packets, the ingress PE could simply replace the VPN label with the VT label, then push the tunnel label. The TTL and COS value of the VPN label entry should be copied to the TTL and COS field of the VT label respectively. This way the depth of the MPLS label stack is unchanged. Though this would require the egress PE to perform VPN prefix lookup in the egress VRF table before the packet can be forwarded to a specific CE, such lookup procedure is also required when per-instance VPN label allocation mechanism is used.

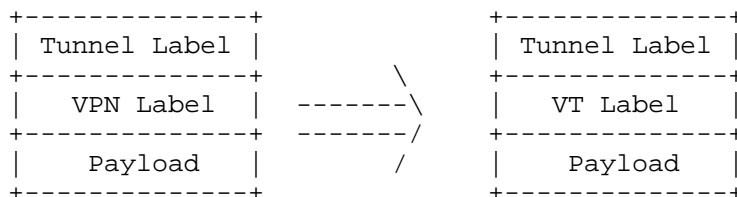


Fig.2 Replace the VPN Label with VT Label

5. L3VPN Performance Monitoring

Since the challenge of identifying the ingress VRF is resolved in section 4, the procedures for the packet loss and delay measurement as defined in [RFC6374] can be utilized for L3VPN performance monitoring. The main difference between performance monitoring of L3VPN and MPLS is the format of identifiers in the Loss Measurement (LM) and Delay Measurement (DM) messages. Specifically, for L3VPN, the source and destination addresses of the LM and DM messages should be set to the concatenation of the Route Distinguisher (RD) of the

particular VRF and the IP address of the ingress and egress PE respectively.

6. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

7. Security Considerations

TBD

8. Acknowledgements

TBD

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

9.2. Informative References

- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.

Authors' Addresses

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

Email: jie.dong@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 19, 2013

D. Rao
J. Mullooly
R. Fernando
Cisco
October 16, 2012

Layer-3 virtual network overlays based on BGP Layer-3 VPNs
draft-drao-bgp-l3vpn-virtual-network-overlays-00

Abstract

Virtual network overlays are being designed and deployed in various types of networks, including data center networks. These network overlays serve several purposes including flexible network virtualization, increased scale, multi-tenancy, and mobility. Such overlay networks may be used to provide both Layer-2 and Layer-3 network services to hosts at the network edge. New encapsulations are being defined and standardized to support these virtual networks. These encapsulations are primarily based on IP, such as VxLAN and NvGRE.

BGP based Layer-3 VPNs, as specified in RFC 4364, provide an industry proven and well-defined solution for supporting Layer-3 virtual network services. RFC 4364 mechanisms use MPLS labels to provide the network virtualization capability in the data plane. This document specifies a simple mechanism to use the new IP-based virtual network overlay encapsulations, while continuing to leverage the BGP based Layer-3 VPN control plane techniques and extensions.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 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	3
1.1. Requirements Language	3
2. Virtual Network Identifier	3
2.1. Virtual Network Identifier Specification	4
2.2. Identifier Scope and propagation	5
2.3. Forwarding behavior	6
3. Overlay Encapsulation	6
3.1. Encapsulation specification	7
4. Acknowledgements	8
5. IANA Considerations	8
6. Security Considerations	8
7. References	8
7.1. Normative References	8
7.2. Informative References	9
Authors' Addresses	9
Intellectual Property and Copyright Statements	11

1. Introduction

Virtual network overlays are being designed and deployed in various types of networks, including data center networks. These network overlays serve several purposes including flexible network virtualization, increased scale, multi-tenancy, and mobility. Such overlay networks may be used to provide both Layer-2 and Layer-3 network services to hosts at the network edge. New encapsulations are being defined and standardized to support these virtual networks. These encapsulations are primarily based on IP, such as VxLAN and NvGRE.

BGP based Layer-3 VPNs, as specified in RFC 4364, provide an industry proven and well-defined solution for supporting Layer-3 virtual network services. RFC 4364 mechanisms use MPLS labels to provide the network virtualization capability in the data plane. This document specifies a simple mechanism to use the new IP-based virtual network overlay encapsulations, while continuing to leverage the BGP based Layer-3 VPN control plane techniques and extensions.

1.1. Requirements Language

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 RFC 2119 [RFC2119].

2. Virtual Network Identifier

In RFC 4364 L3VPNs, a 20-bit MPLS label that is assigned to a VPN route determines the forwarding behavior in the data plane for traffic following that route. These labels also serve to distinguish the packets of one VPN from another.

On the other hand, the various IP overlay encapsulations support a virtual network identifier as part of their encapsulation format. A virtual network identifier is a value that at a minimum can identify a specific virtual network in the data plane. It is typically a 24-bit value which can support upto 16 million individual network segments.

There are two useful requirements regarding the scope of these virtual network identifiers.

- o Network-wide scoped virtual network identifiers

Depending on the provisioning mechanism used within a network domain such as a data center, the virtual network identifier may have a

network scope, where the same value is used to identify the specific Layer-3 virtual network across all network edge devices where this virtual network is instantiated. This network scope is useful in environments such as within the data center where networks can be automatically provisioned by central orchestration systems. Having a uniform virtual network identifier per VPN is a simple approach, while also easing network operations (i.e. troubleshooting). It also means simplifies requirements on network edge devices, both physical and virtual devices. A critical requirement for this type of approach is to have a very large amount of network identifier values given the network-wide scope.

- o Locally assigned virtual network identifiers

In an alternative approach supported as per RFC 4364, the identifier has local significance to the network edge device that advertises the route. In this case, the virtual network scale impact is determined on a per node basis, versus a network basis.

When it is locally scoped, and uses the same existing semantics of a MPLS VPN label, the same forwarding behaviors as specified in RFC 4364 can be employed. It thus allows a seamless stitching together of a VPN that spans both an IP based network overlay and a MPLS VPN. This situation can occur for instance at the data center edge where the overlay network feeds into an MPLS VPN. In this case, the identifier may be dynamically allocated by the advertising device.

It is important to support both cases, and in doing so, ensure that the scope of the identifier be clear and the values not conflict with each other.

It should be noted that deployment scenarios for these virtual network overlays are not constrained to the examples used above to categorize the options. For example, a virtual network overlay may extend across multiple data centers.

- o Global unicast table

The overlay encapsulation can also be used to support forwarding for routes in the global or default routing table. A virtual network identifier value can be allocated for the purpose as per the above options.

2.1. Virtual Network Identifier Specification

The above requirements can be achieved in a simple manner by splitting the virtual network ID number space.

- o Values upto 1 million (or less than 20 bits) are treated exactly as MPLS labels and have significance local to the advertiser.

For future expansion, this draft stipulates that the 16 numerical values in the end of the label range, i.e. values 0xffff0 to 0xfffff, be reserved for future use. These special labels could be used to indicate the presence of other types of IP payloads.

- o Values greater than 1 million (greater than 20 bits) are treated as per their original definition.
- o A virtual network identifier value of zero is used by default to indicate the global or routing table.

It should be noted that within an administrative domain, the entire range can be used such that the values have network-wide significance. This is inline with the use of statically assigned labels today.

2.2. Identifier Scope and propagation

The virtual network identifier may be indicated by attaching to the route a new attribute. However, it is also possible to use the MPLS label field in the BGP VPN NLRI to specify this value. The benefit of doing the latter is the reuse of existing NLRI and label processing as is, especially keeping in mind the semantics to be supported. The specification of the identifier value in the label field is described further below.

The use of the virtual network identifier is coupled with the encapsulation used for sending traffic.

The encapsulation used may be MPLS. In this case, the identifier value should be less than 0xffff0, and will be set in the MPLS label field exactly as defined in RFC 3107. There is no change to current RFC 4364 behavior in this case.

When the encapsulation is one of the overlay encapsulation types as listed below, the virtual network identifier will be set in the 3-byte label field described in RFC 3107 as a 24-bit value, irrespective of the actual value being specified.

The value itself may fall into two ranges.

1. Less than 0xffff0 - In this case, the identifier has local significance to the network device that advertised the route.
2. Greater than 0xfffff - In this case, the identifier will have a

significance as per the original definitions, typically within a network domain that is under a common provisioning system.

From a routing perspective, if an intermediate network device changes the BGP next-hop to self before propagating the route, it will assign a new virtual network identifier and advertise it. If not, the virtual network identifier attached by the originator of the route will be carried as is.

When an intermediate network device assigns a virtual network identifier, the assigned value may be a new locally assigned value or it could still be the same network scoped value, if the route is being propagated within the domain.

2.3. Forwarding behavior

o Locally assigned virtual network identifiers

When the virtual network identifier is locally assigned, forwarding based on the identifier follows the semantics of an MPLS label. This label can serve as either an aggregate label or a per-prefix label. This allows a seamless transition out of the overlay network at an MPLS VPN edge, for example, via support of Inter-AS option B.

o Network-scoped virtual network identifiers

With the network-scoped virtual network identifier, any egress device treats the identifier as an aggregate label to lookup the appropriate forwarding table.

In both cases, the forwarding behavior at an ingress edge device, physical or virtual, does not change.

3. Overlay Encapsulation

As mentioned above, different overlay encapsulations may be used to provide an overlay virtual network.

The overlay may use proposed encapsulations such as:

1. VxLAN
2. NVGRE

Based on the encapsulation type being used, the virtual network identifier is appropriately encoded.

When VxLAN encapsulation is used, the virtual network identifier is carried as the 24-bit segment-ID in the VxLAN header.

When NvGRE encapsulation is used, the virtual network identifier is carried as the 24-bit tenant network ID in the NvGRE header.

The fact that a virtual network identifier is carried in the label field in the BGP NLRI is determined by virtue of the accompanying encapsulation attribute, that indicates an overlay encapsulation should be used.

For a given overlay edge device, the same encapsulation may be used for all routes or for selected routes.

3.1. Encapsulation specification

The overlay encapsulation attribute may be carried with each route, or it may be indirectly inferred from the route to the BGP next-hop.

The Tunnel Encapsulation Extended community defined in RFC 5512 can be used to convey this information. [remote-next-hop] specifies an alternative mechanism to carry this information along with each route. The address specified as the remote next-hop identifies the end-point or destination of the encapsulated packets that use the dependent routes.

A single encapsulation may be used on a given device. In this case, the encapsulation may be specified for a given next-hop and inherited by all routes sent with that next-hop (RFC 5512).

When VxLAN and NvGRE encapsulations are used, the header by definition contains an Ethernet MAC address within the overlay header. When these encapsulations are used for Layer-3 as specified in this document, the MAC addresses are not relevant. A single well-known MAC address may be specified for the purpose of deterministically driving a Layer-3 lookup based on the inner IP or IPv6 address.

However, an overlay egress edge device may choose to specify a MAC address as part of the encapsulation header in its route advertisement. In this case, any ingress edge device sending traffic as per this route must use the above specified MAC address as the destination MAC address in the header. The egress device may use this address to drive the Layer-3 table lookup or for other purposes.

When an intermediate device changes the BGP next-hop to self before propagating a received route, it will also need to advertise a new overlay encapsulation attribute with the local address as the

endpoint. While doing so, it may use an overlay encapsulation type that is different from the received route.

4. Acknowledgements

The authors would like to acknowledge and thank Dave Smith, Maria Napierala, Ashok Ganesan and Luyuan Fang for their input and feedback.

5. IANA Considerations

The virtual network identifier values 0xffff0 to 0xfffff should be allocated by IANA as applications for carrying payloads different than regular IP/VPN packets emerge in future.

6. Security Considerations

This draft does not add any additional security implications to the BGP/L3VPN control plane. All existing authentication and security mechanisms for BGP apply here.

The security considerations pertaining to the various IP overlay encapsulations referenced here are described in the respective overlay encapsulation specifications.

7. References

7.1. Normative References

[I-D.mahalingam-dutt-dcops-vxlan]

Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-00 (work in progress), August 2011.

[I-D.sridharan-virtualization-nvgre]

Sridhavan, M., Duda, K., Ganga, I., Greenberg, A., Lin, G., Pearson, M., Thaler, P., Tumuluri, C., and Y. Wang, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-00 (work in progress), September 2011.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

[min_ref] authSurName, authInitials., "Minimal Reference", 2006.

7.2. Informative References

- [I-D.narten-iana-considerations-rfc2434bis]
Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs",
draft-narten-iana-considerations-rfc2434bis-09 (work in progress), March 2008.
- [I-D.vandeveldede-idr-remote-next-hop]
Velde, G., Patel, K., Raszuk, R., and R. Bush, "BGP Remote-Next-Hop", draft-vandeveldede-idr-remote-next-hop-01 (work in progress), July 2012.
- [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, July 2003.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

Authors' Addresses

Dhananjaya Rao
Cisco
San Jose,
USA

Email: dhr Rao@cisco.com

John Mullooly
Cisco
New Jersey,
USA

Email: jmullool@cisco.com

Rex Fernando
Cisco
San Jose,
USA

Email: rex@cisco.com

Internet-Draft

BGP Layer-3 virtual network overlay

October 2012

Rao, et al.

Expires April 19, 2013

[Page 11]

Network Working Group
Internet Draft
Intended status: Informational
Expires: April 15, 2013

Maria Napierala
AT&T
Luyuan Fang
Cisco Systems

October 15, 2012

Requirements for Extending BGP/MPLS VPNs to End-Systems
draft-fang-l3vpn-end-system-requirements-00.txt

Abstract

Service Providers commonly use BGP/MPLS VPNs [RFC 4364] as the control plane for wide-area virtual networks. This technology has proven to scale to a large number of VPNs and attachment points, and it is well suited to provide VPN service to end-systems. Virtualized environment imposes additional requirements to MPLS/BGP VPN technology when applied to end-system networking, which are defined in this document.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.
Napierala, Fang Expire April 2012 [Page 1]

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	3
1.1. Terminology	3
2. Application of MPLS/BGP VPNs to End-Systems	3
3. Connectivity Requirements	4
4. Multi-Tenancy Requirements	5
5. Decoupling of Virtualized Networking from Physical Infrastructure	5
6. Decoupling of Layer 3 Virtualization from Layer 2 Topology	6
7. Encapsulation of Virtual Payloads	6
8. Optimal Forwarding of Traffic	7
9. Inter-operability with Existing MPLS/BGP VPNs	8
10. IP Mobility	9
11. BGP Requirements in a Virtualized Environment	10
11.1. BGP Convergence and Routing Consistency	10
11.2. Optimizing Route Distribution	11
12. Security Considerations	11
13. IANA Considerations	11
14. Normative References	11
15. Informative References	11
16. Authors' Addresses	11
17. Acknowledgements	12

Requirements Language

Although this document is not a protocol specification, 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 RFC 2119 [RFC 2119].

1. Introduction

Networks are increasingly being consolidated and outsourced in an effort, both, to improve the deployment time of services as well as reduce operational costs. This coincides with an increasing demand for compute, storage, and network resources from applications.

In order to scale compute, storage, and network service functions, physical resources are being abstracted from their logical representation. This is referred as server, storage, and network virtualization. Virtualization can be implemented in various layers of computer systems or networks. The virtualized loads are executed over a common physical infrastructure. Compute nodes running guest operating systems are often executed as Virtual Machines (or VMs).

This document defines requirements for a network virtualization solution that provides IP connectivity to virtual resources on end-systems. The requirements address the virtual resources, defined as Virtual Machines, applications, and appliances that require only IP connectivity. Non-IP communication is addressed by other solutions and is not in scope of this document.

1.1. Terminology

AS	Autonomous Systems
End-System	A device where Guest OS and Host OS/Hypervisor reside
IaaS	Infrastructure as a Service
RT	Route Target
ToR	Top-of-Rack switch
VM	Virtual Machine
Hypervisor	Virtual Machine Manager
SDN	Software Defined Network
VPN	Virtual Private Network

2. Application of MPLS/BGP VPNs to End-Systems

MPLS/BGP VPN technology [RFC 4364] have proven to be able to scale to a large number of VPNs (tens of thousands) and customer routes (millions) while providing for aggregated management capability. In traditional WAN deployments of BGP IP VPNs a Customer Edge (CE) is a physical device connected to a Provider Edge (PE). In addition, the forwarding function and control function of a Provider Edge (PE) device co-exist within a single physical router.

MPLS/BGP VPN technology should to able to evolve and adapt to new virtualized environments by extending VPN service to end-systems.

When end-system attaches to MPLS/BGP VPN, CE becomes a Virtual Machine or an application residing on the end-system itself. As in traditional MPLS/BGP VPN deployments, it is undesirable for the end-system VPN forwarding knowledge to extend to the transport network infrastructure. Hence, optimally, with regard to forwarding the end-system should become both the CE and the PE simultaneously. Moreover, it is a current practice to implement PE forwarding and control functions in different processors of the same device and to use internal (proprietary) communication between those processors. Typically, the PE control functionality is implemented in one (or very few) components of a device and the PE forwarding functionality is implemented in multiple components of the same device (a.k.a., "line cards"). In end-system environment, a single end-system, effectively, corresponds to a line card in a traditional PE router. For scalable and cost effective deployment of end-system MPLS/BGP VPNs PE forwarding function should be decoupled from PE control function such that the former can be implemented on multiple standalone devices. This separation of functionality will allow for implementing the end-system PE forwarding on multiple end-system devices, for example, in operating systems of application servers or network appliances. The PE control plane function can itself be virtualized and run as an application in end-system.

3. Connectivity Requirements

A network virtualization solution should be able to provide IPv4 and IPv6 unicast connectivity between hosts in the same and different subnets without any assumptions regarding the underlying media layer.

Furthermore, the multicast transmission, i.e., allowing IP applications to send packets to a group of IPv4 or IPv6 addresses should be supported. The multicast service should also support a delivery of traffic to all endpoints of a given VPN even if those endpoints have not sent any control messages indicating the need to receive that traffic. In other words, the multicast service should be capable of delivering the IP broadcast traffic in a virtual topology. A solution for supporting VPN multicast and VPN broadcast must not require that the underlying transport network supports IP multicast transmission service.

In some deployments, Virtual Machines or applications are configured to belong to an IP subnet. A network virtualization solution should support grouping of virtual resources into IP subnets regardless of whether the underlying implementation uses a multi-access network or not.

4. Multi-Tenancy Requirements

One of the main goals of network virtualization is to provide traffic and routing isolation between different virtual components that share a common physical infrastructure. A collection of virtual resources might provide external or internal services. For example, such collection may serve an external "customer" or internal "tenant" to whom a Service Provider provides service(s). We will refer to collection of virtual resources dedicated to a process or application as a VPN, using the terminology of IP VPNs.

Any network virtualization solution has to assure the network isolation (in data plane and control plane) among tenants or applications sharing the same data center physical resources. Typically VPNs that belong to different external tenants do not communicate with each other directly but they should be allowed to access shared services or shared network resources. It is also common for tenants to require multiple distinct VPNs. In that scenario traffic might need to cross VPN boundaries, subject to access controls and/or routing policies.

A tenant should be able to create multiple VPNs. A network virtualization solution should allow a VM or application end-point to directly access multiple VPNs without a need to traverse a gateway. It is often the case that SP infrastructure services are provided to multiple tenants, for example voice-over-IP gateway services or video-conferencing services for branch offices. A network virtualization solution should support both, isolated VPNs and overlapping VPNs (often referred to as "extranets"), as well as both, any-to-any and hub-and-spoke topologies.

5. Decoupling of Virtualized Networking from Physical Infrastructure

One of the main goals in designing a large scale transport network is to minimize the cost and complexity of its "fabric". It is often done by delegating the virtual resource communication processing to the network edge. Networks use various VPN technologies to isolate disjoint groups of virtual resources. Some use VLANs as a VPN technology, others use layer 3 based solutions, often with proprietary control planes. Service Providers are interested in interoperability and in openly documented protocols rather than in proprietary solutions.

The transport network infrastructure should not maintain any information that pertains to the virtual resources in end-systems. Decoupling of virtualized networking from the physical infrastructure has the following advantages: 1) provides better

scalability; 2) simplifies the design and operation; 3) reduces network cost. It has been proven (in Internet and in large BGP IP VPN deployments) that moving complexity to network edge while keeping network core simple has very good scaling properties.

There should be a total separation between the virtualized segments (i.e., interfaces associated with virtual resources) and the physical network (i.e., physical interfaces associated with network infrastructure). This separation should include the separation of the virtual network IP address space from the physical network IP address space. The physical infrastructure addresses should be routable in the underlying transport network, while the virtual network addresses should be routable only in the virtual network. Not only should the virtual network data plane be fully decoupled from the physical network, but its control plane should be decoupled as well.

6. Decoupling of Layer 3 Virtualization from Layer 2 Topology

The layer 3 approach to network virtualization dictates that the virtualized communication should be routed, not bridged. The layer 3 virtualization solution should be decoupled from the layer 2 topology. Thus, there should be no dependency on VLANs and layer 2 broadcast.

In solutions that depend on layer 2 broadcast domains, host-to-host communication is established based on flooding and data plane MAC learning. Layer 2 MAC information has to be maintained on every switch where a given VLAN is present. Even if some solutions are able to minimize data plane MAC learning and/or unicast flooding, they still rely on MAC learning at the network edge and on maintaining the MAC addresses on every (edge) switch where the layer 2 VPN is present.

The MAC addresses known to guest OS in end-system are not relevant to IP services and introduce unnecessary overhead. Hence, the MAC addresses associated with virtual resources should not be used in the virtual layer 3 networks. Rather, only what is significant to IP communication, namely the IP addresses of the virtual machines and application endpoints should be maintained by the virtual networks.

7. Encapsulation of Virtual Payloads

In a layer 3 end-system virtual network, IP packets should reach the first-hop router in one IP-hop, regardless of whether the first-hop router is an end-system itself (i.e., a hypervisor/Host

OS) or it is an external (to end-system) device. The first-hop router should always perform an IP lookup on every packet it receives from a virtual machine or an application. The first-hop router should encapsulate the packets and route them towards the destination end-system.

In order to scale the transport networks, the virtual network payloads must be encapsulated with headers that are routable (or switchable) in the physical network infrastructure. The IP addresses of the virtual resources are not to be advertized within the physical infrastructure address space.

The encapsulation (and decapsulation) function should be implemented on a device as close to virtualized resources as possible. Since the hypervisors in the end-systems are the devices at the network edge they are the most optimal location for the encap/decap functionality. A device implementing the encap/decap functionality acts as the first-hop router in the virtual topology.

The network virtualization solution should also support deployments where it is not possible or not desirable to implement the virtual payload encapsulation in the hypervisor/Host OS. In such deployments encap/decap functionality may be implemented in an external device. The external device implementing encap/decap functionality should be as close as possible to the end-system itself. The same network virtualization solution should support deployments with both, internal (in a hypervisor) and external (outside of a hypervisor) encap/decap devices.

Whenever the virtual forwarding functionality is implemented in an external device, the virtual service itself must be delivered to an end-system such that switching elements connecting the end-system to the encap/decap device are not aware of the virtual topology.

MPLS/VPN technology based on [RFC 4364] specifies that different encapsulation methods could be for connecting PE routers, namely Label Switched Paths (LSPs), IP tunneling, and GRE tunneling. If LSPs are used in the transport network they could be signaled with LDP, in which case host (/32) routes to all PE routers must be propagated throughout the network, or with RSVP-TE, in which case a full mesh of RSVP-TE tunnels is required. If the transport network is only IP-capable then MPLS in IP or MPLS in GRE [RFC4023] encapsulation could be used. Other transport layers such 802.1ah might also need to be supported.

8. Optimal Forwarding of Traffic

The network virtualization solutions that optimize for the maximum utilization of compute and storage resources require that those resources may be located anywhere in the network. The physical and logical spreading of appliances and workloads implies a very significant increase in the infrastructure bandwidth consumption. Hence, it is important that the virtualized networking solutions are efficient in terms of traffic forwarding and assure that packets traverse the transport network only once.

It must be also possible to send the traffic directly from one end-system to another end-system without traversing through a midpoint router.

9. Inter-operability with Existing MPLS/BGP VPNs

Service Providers want to tie their server-based offerings to their MPLS/BGP VPN services. MPLS/BGP VPNs provide secure and latency-optimized WAN connectivity to the virtualized resources in SP's data center. MPLS/BGP VPN customers may require simultaneous access to resources in both SP and their own data centers. The service provider-based VPN access can provide additional value compared with public internet access, such as security, QoS, OAM, multicast service, VoIP service, video conferencing, wireless connectivity. Service Providers want to "spin up" the L3VPN access to data center VPNs as dynamically as the spin up of compute and other virtualized resources.

The network virtualization solution should be fully inter-operable with MPLS/BGP VPNs, including Inter-AS MPLS/BGP VPN Options A, B, or C [RFC 4364]. MPLS/BGP VPN technology is widely supported on routers and other appliances. BGP/MPLS VPN-capable network devices should be able to participate directly in a virtual network that spans end-systems. The network devices should be able to participate in isolated collections of end-systems, i.e., in isolated VPNs, as well as in overlapping VPNs (called "extranets" in BGP/MPLS VPN terminology).

When connecting an end-system VPN with other services/networks, it should not be necessary to advertize the specific host routes but rather the aggregated routing information. A BGP/MPLS VPN-capable router or appliance can be used to aggregate VPN's IP routing information and advertize the aggregated prefixes. The aggregated prefixes should be advertized with the router/appliance IP address as BGP next-hop and with locally assigned aggregate 20-bit label. The aggregate label should trigger a destination IP lookup in its corresponding VRF on all the packets entering the virtual network.

The inter-connection of end-system VPNs with traditional VPNs requires an integrated control plane and unified orchestration of network and end-system resources.

10. IP Mobility

Another reason for a network virtualization is the need to support IP mobility. IP mobility consists in IP addresses used for communication within or between applications being anywhere across the network. Using a virtual topology, i.e., abstracting the externally visible network address from the underlying infrastructure address is an effective way to solve IP mobility problem.

IP mobility consists in a device physically moving (e.g., a roaming wireless device) or a workload being transferred from one physical server/appliance to another. IP mobility requires preserving device's active network connections (e.g., TCP and higher-level sessions). Such mobility is also referred to as "live" migration with respect to a Virtual Machine. IP mobility is highly desirable for many reasons such as efficient and flexible resource sharing, data center migration, disaster recovery, server redundancy, or service bursting.

To accommodate live mobility of a virtual machine (or a device), it is desirable to assign to it a permanent IP address that remains with the VM/device after it moves. When dealing with IP-only applications it is not only sufficient but optimal to forward the traffic based on layer 3 rather than on layer 2 information. The MAC addresses of devices or applications should be irrelevant to IP services and introduce unnecessary overhead and complications when devices or VMs move (i.e., when a VM moves between physical servers, the MAC learning tables in the switches must be updated; also, it is possible that VM's MAC address might need to change in its new location). In IP-based network virtualization solution a device or a workload move should be handled by an IP route advertisement.

IP mobility has to be transparent to applications and any external entity interacting with the applications. This implies that the network connectivity restoration time is critical. The transport sessions can typically survive over several seconds of disruption, however, applications may have sub-second latency requirement for their correct operation.

To minimize the disruption to established communication during workload or device mobility, the control plane of a network virtualization solution should be able to differentiate between the

activation of a workload in a new location from advertising its route to the network. This will enable the remote end-points to update their routing tables prior to workload's migration as well as allowing the traffic to be tunneled via the workload's old location.

11. BGP Requirements in a Virtualized Environment

11.1. BGP Convergence and Routing Consistency

BGP was designed to carry very large amount of routing information but it is not a very fast converging protocol. In addition, the routing protocols, including BGP, have traditionally favored convergence (i.e., responsiveness to route change due to failure or policy change) over routing consistency. Routing consistency means that a router forwards a packet strictly along the path adopted by the upstream routers. When responsiveness is favored, a router applies a received update immediately to its forwarding table before propagating the update to other routers, including those that potentially depend upon the outcome of the update. The route change responsiveness comes at the cost of routing blackholes and loops.

Routing consistency in virtualized environments is important because multiple workloads can be simultaneously moved between different physical servers due to maintenance activities, for example. If packets sent by the applications that are being moved are dropped (because they do not follow a live path), the active network connections will be dropped. To minimize the disruption to the established communications during VM migration or device mobility, the live path continuity is required.

11.1.1. BGP IP Mobility Requirements

In IP mobility, the network connectivity restoration time is critical. In fact, Service Provider networks already use routing and forwarding plane techniques that support fast failure restoration by pre-installing a backup path to a given destination. These techniques allow to forward traffic almost continuously using an indirect forwarding path or a tunnel to a given destination, and hence, are referred to as "local repair". The traffic path is restored locally at the destination's old location while the network converges to a backup path. Eventually, the network converges to an optimal path and bypasses the local repair. BGP assists in the local repair techniques by advertizing multiple and not only the best path to a given destination.

11.2. Optimizing Route Distribution

When virtual networks are triggered based on the IP communication, the Route Target Constraint extension [RFC 4684] of BGP should be used to optimize the route distribution for sparse virtual network events. This technique ensures that only those VPN forwarders that have local participants in a particular data plane event receive its routing information. This also decreases the total load on the upstream BGP speakers.

12. Security Considerations

The document presents the requirements for end-systems MPLS/BGP VPNs. The security considerations for specific solutions will be documented in the relevant documents.

13. IANA Considerations

This document contains no new IANA considerations.

14. Normative References

[RFC 4363] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC 4023] Worster, T., Rekhter, Y. and E. Rosen, "Encapsulating in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.

[RFC 4684] 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, November 2006.

15. Informative References

[RFC 2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

16. Authors' Addresses

Maria Napierala
AT&T
200 Laurel Avenue
Middletown, NJ 07748
Email: mnapierala@att.com

Luyuan Fang
Cisco Systems
111 Wood Avenue South
Iselin, NJ 08830, USA
Email: lufang@cisco.com

17. Acknowledgements

The authors would like to thank Pedro Marques for his comments and input.

INTERNET-DRAFT
Intended Status: Informational
Expires: April 22, 2013

Luyuan Fang
David Ward
Rex Fernando
Cisco
Maria Napierala
AT&T
Nabil Bitar
Verizon
Dhananjaya Rao
Cisco

October 22, 2012

BGP L3VPN Virtual PE Framework
draft-fang-l3vpn-virtual-pe-framework-01

Abstract

This document describes a framework for BGP/MPLS L3VPN with virtual PE solutions. It provides functional description of the control plane and data plane of the virtual PE solutions. It also describes interactions among the vPE solutions and other network elements. The virtual PE solutions support further control plane and forwarding plane separation when compared with traditional L3VPN PE solutions. It allows the L3VPN functions to be extended to application end devices for large scale and efficient IP application support.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 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	4
1.1	Terminology	4
1.2	Motivation	5
1.3	Scope of the document	6
2.	Virtual PE Architecture and Reference Model	6
2.1	Virtual PE	7
2.2	Architecture reference model	7
3.	Control Plane	10
3.1	vPE Control Plane	10
3.1	Route server of vPE	11
3.3	Use of router reflector	11
3.4	Use of RT constraint	12
4.	Forwarding Plane	12
4.1	Virtual Interface	12
4.2	VPN forwarder	12
4.3	Encapsulation	12
4.4	Optimal forwarding	13
5.	Addressing	14
5.1	IPv4 and IPv6 support	14
5.2	Address space separation	14
6.0	Inter-connection considerations	14
7.	Security Considerations	15
8.	IANA Considerations	15
9.	References	15
9.1	Normative References	15

9.2 Informative References 16

Authors' Addresses 16

1 Introduction

Network virtualization is to provide multiple individual network services through shared common network resources. Network virtualization is not a new concept. For example, BGP/MPLS layer 3 Virtual Private Networks (L3VPNs) [RFC4364] have been widely deployed to provide network based virtual private network services. It provides routing isolation and forwarding separation for individual VPNs, allow IP address overlapping among different VPNs while forwarding traffic over common network infrastructure.

Network virtualization enables the support of multiple isolated individual networks over a common network infrastructure. Network virtualization is not a new concept. For example, BGP/MPLS IP Virtual Private Network (IP VPNs) [RFC4364] have been widely deployed to provide network based, service provider provisioned IP VPNs for multiple customers with overlapping IP address spaces over a common service provider IP/MPLS network. BGP/MPLS IPVPNs provide routing isolation among customers and allow address overlapping among different VPNs by having per-customer Virtual Routing and Forwarding Instance (VRF) at a service provided Edge (PE), while forwarding customer traffic over a common IP/MPLS network infrastructure.

With the advent of compute capabilities and the proliferation of virtualization in end devices for systems and applications, PE functionality virtualization on such end device is becoming feasible, and in some cases attractive for scale and efficiency. Scale and efficiency are crucial factors in the cloud computing environment supporting various applications and services, and in traditional service provider space.

The virtual Provider Edge (vPE) solution described in this document is to extend the functionality of BGP/MPLS L3VPN to the application end device.

1.1 Terminology

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 RFC 2119 [RFC2119].

Term	Definition
-----	-----
3GPP	3rd Generation Partnership Project (3GPP)
AS	Autonomous Systems
ASBR	Autonomous Systems Border Router

BGP	Border Gateway Protocol
ED	End device: where Guest OS, Host OS/Hypervisor, applications, VMs, and virtual router may reside
Forwarder	L3VPN forwarding function
GRE	Generic Routing Encapsulation
IaaS	Infrastructure as a Service
IRS	Interface to Routing System
LTE	Long Term Evolution
MP-BGP	Multi-Protocol Border Gateway Protocol
PCEF	Policy Charging and Enforcement Function
P	Provider backbone router
RR	Route Reflector
RT	Route Target
RTC	RT Constraint
ToR	Top-of-Rack switch
VM	Virtual Machine
Hypervisor	Virtual Machine Manager
VM	Virtual Machine
SDN	Software Defined Network
VI	Virtual Interface
vCE	virtual Customer Router
vPC	virtual Private Cloud
vPE	virtual Provider Edge
VPN	Virtual Private Network
vRR	virtual Route Reflector
WAN	Wide Area Network

Virtual PE is a PE resides in an end device (e.g., a server) along with client/application VMs.

Through out this document, the term virtual PE (vPE) is used to denote BGP/MPLS L3VPN virtual Provider Edge router.

1.2 Motivation

The recent rapid adoption of Cloud Services by enterprises and the phenomenal growth of mobile IP applications accelerate the needs to extend the L3VPN capability to the end devices. For example, Enterprise customers requested Service Providers to extend and integrate their L3VPN services available in the WAN into the new Cloud services; large enterprise have existing L3VPN deployment are extending them into their data centers; mobile providers are adopting L3VPN into their 3GPP Mobile infrastructure are looking to extend the L3VPNs to their end device of their call processing center.

The virtual PE solution described in this document is aimed to meet the following key requirement [I-D.fang-l3vpn-end-system-req].

- 1) Support end device multi-tenancy, per tenant routing isolation and traffic separation.
- 2) Support large scale L3VPNs in service network, upto tens of thousands of end devices and Millions of VMs in the single service network, e.g., a data center.
- 3) Support end-to-end L3VPN connectivity, e.g. L3VPN can start from a service network end device, connect to a corresponding L3VPN in the WAN, and terminate in another service network end device.
- 4) Decoupling control plane and forwarding for network virtualization and abstraction.

L3VPN is the proven technologies which is capable of providing routing and forwarding separation, and it is proven with large scale deployment (e.g. supporting 7-8 million L3VPN routes in single Service Provider networks today).

By extending L3VPN solution to the end device with vPE solution, application end-to-end (VM to VM, applications to the end user) L3VPN connectivity can be achieved, and well as the true network virtualization and abstraction.

The architecture and protocols defined in BGP/MPLS IP VPN [RFC4364] is the foundation for virtual PE extension. Certain protocol extensions or integration may be needed to support the virtual PE solutions.

1.3 Scope of the document

It is assumed that the readers are familiar with BGP/MPLS IP VPN [RFC4364] terms and technologies, the base technology and its operation are not reviewed in details in this document.

The following network elements are discussed in this document: the concept of BGP L3VPN vPE; the interaction of vPE with other network elements, including BGP L3VPN physical PE, physical or virtual BGP Route Reflectors (RR, vRR), and Autonomous System Border Router (ASBR), Service Network gateway routers, external controllers, provisioning/orchestration systems, and the vPE inter-connections with other non L3VPN networks.

The definitions of protocols extensions are out of the scope of this document.

2. Virtual PE Architecture and Reference Model

2.1 Virtual PE

As defined in [RFC4364], a L3VPN is created by applying policies to form a subset of sites among all sites connected the backbone network. It is collection of "sites". A site can be considered as a set of IP systems maintain IP inter-connectivity without connecting through the backbone. The typical use of L3VPM has been to inter-connect different sites of an Enterprise networks through Service Provider's L3VPNs in the WAN.

A virtual PE (vPE) is a PE instance which resides in one or more physical devices, it is commonly placed in a network service (e.g. a Data Center) end device (e.g., a Server) where the client/application VMs are hosted. The control and forwarding components of the vPE are decoupled, they may reside in the same physical device or in different physical devices.

In the case that a vPE is in a Data Center server along with client/application VMs, one can view the vPE to VM relationship as a typical PE-CE relationship. Unlike a regular physical PE, vPE allows L3VPN control plane and forwarding function residing on different physical devices. The full MP-BGP control plane may reside on the end device, or may be external to the end device, e.g., in a BGP L3VPN boarder router (ASBR)/DC gateway router, a Route Reflector (RR), or an external controller.

Virtual PE solution allows the placement of L3VPN termination point right inside the end device (e.g., a server). In this case, the vPE to CE (VM) connection is internal to the device. If both control and forwarding elements are placed on the end device, L3VPN routing and forwarding starts from the end device, the eliminate the needs for additional process in the next hop (e.g., layer2 and layer 3 integration). This approach helps to simplify the operation and improve the routing and forwarding efficiency in large scale deployment.

Another important benefit is that vPE solution allows full control and forwarding decoupling for scale and achieving true network virtualization to allow network abstraction, flexible and dynamic policy control, quick service turn up time and VM mobility support.

2.2 Architecture reference model

Figure 1 illustrate the topology that vPE is reside in the end device where the applications are hosted.

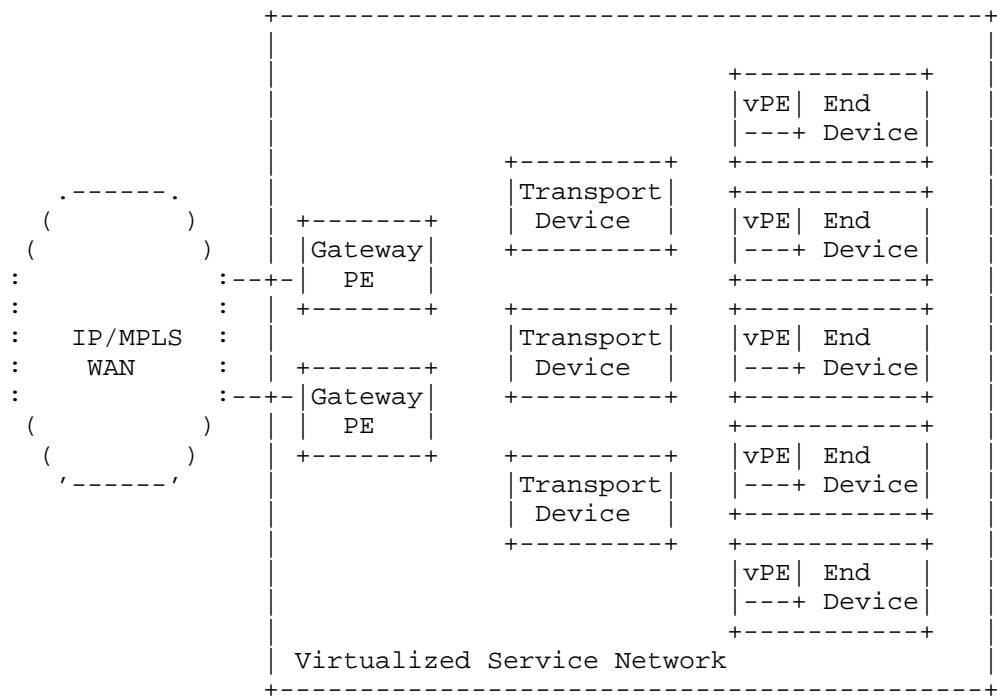


Figure 1. Virtualized Service network with vPE

The Virtualized Service Network in Figure 1 consists of WAN gateway PE devices, transport devices, and end devices. In some networks, it is feasible the VPN Gateways may be implemented as vPEs as well.

Examples of service network may be a network that supports cloud computing services, mobile call centers, and SP or enterprise data centers.

Note that the transport devices in the service network in the diagram do not participate L3VPNs, they function similar as P routers in MPLS back bone, they do not maintain the L3VPN states, and are not L3VPN aware.

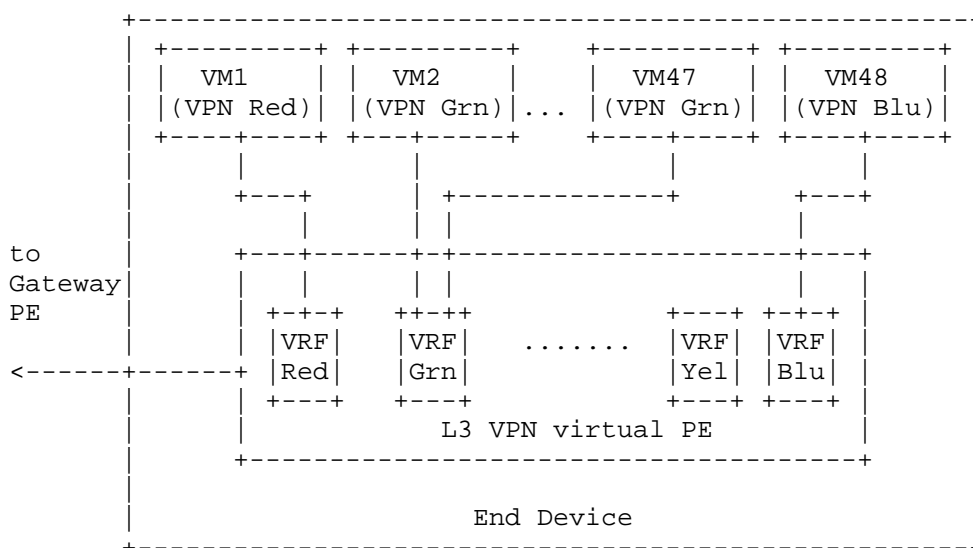


Figure 2. VM in end device to VRF in vPE mapping

An end device shown in Figure 2 is a virtualized server or system which hosts multiple VMs, the virtual PE resides in the end device. The vPE supports multiple VRFs, VRF Red, VRF Grn, VRF Yel, VRF Blu, etc. Each client or application VM is associated to a particular VRF as a member of the particular VPN. For example, VM1 is associated to VRF Red, VM2 and VM47 are associated to RFC Grn, etc. Routing isolation applies between VPNs for multi-tenancy support. For example, VM1 and VM2 can not communicate with each other in a simple intranet L3VPN topology as shown in the configuration.

The vPE connectivity relationship between vPE and the application VM is similar to the PE to CE relationship in a regular BGP L3VPNs.

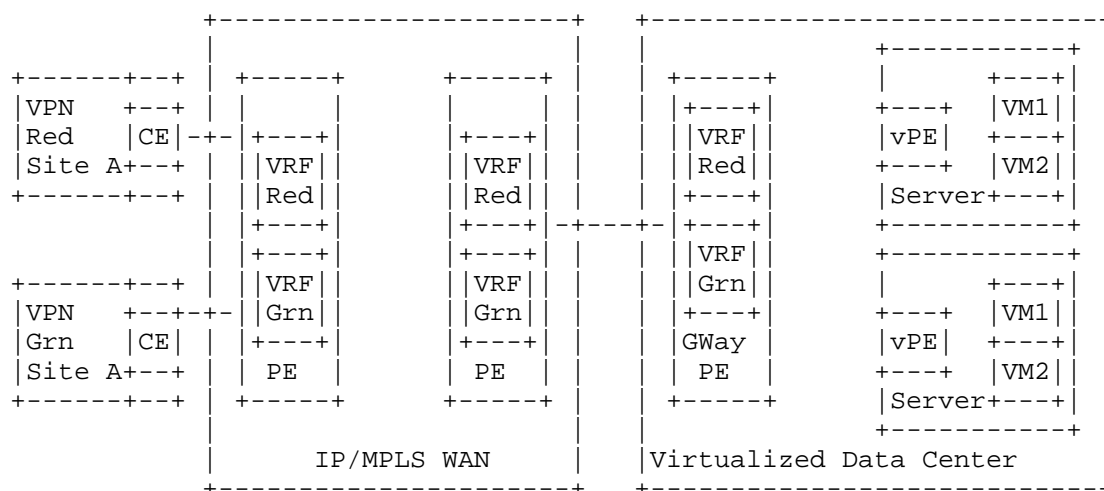


Figure 3. Connecting Enterprise CE to DC VM over WAN

The example of connection from an Enterprise site to application VMs through vPE on the end device of a SP provisioned virtualized data center.

There are multiple options for VPN control plane signaling between the Gateway PE to vPE on the server within the data center. It can use MP-BGP as in regular L3VPN, or use other extensible IP messaging protocols defined in IETF, or use controller direct signaling as a SDN approach.

The inter-connection from DC Gateway PE to MPLS WAN may use one of the Inter-AS options if they are in different ASes. Option B may be more practical for the reasons it is more scalable than Option A, and more restricted than Option C. Consider route aggregation with Option B if both sides have large number of routes.

The connection between backbone VPN to VPN CE on the left hand side is regular L3VPN connection, e-BGP, or static, or other protocols can be used.

3. Control Plane

3.1 vPE Control Plane

The vPE control plane can be distributed or centralized.

1) Distributed control plane

vPE participates in underlay routing through IGP protocols: ISIS or OSPF.

vPE participates in overlay L3VPN control protocol: MP-BGP [RFC4364].

While MP-BGP is the de facto preferred choice between vPE and gateway-PE, using extensible signaling messaging protocols can be alternative, such technologies have been proposed for this segment of signaling [I-D.ietf-l3vpn-end-system].

2. Centralized routing controller

This is a SDN approach. In the virtual PE implementation, not only the service network infrastructure and the VPN overlay networks are decoupled, but also the vPE control plane and data plane are physically decoupled. The control plane directing the data flow may reside elsewhere, such a centralized controller. This requires standard interface to routing system (IRS). The Interface to Routing System (IRS) is work in progress in IETF [I-D.ward-irs-framework], [I-D.rfernando-irs-fw-req].

3.1 Route server of vPE

A virtual PE consist the control plane element and the forwarding plane element. Since the proposed solution decoupled the two element, they may or may not reside on the same physical device.

The Route Server of L3VPN vPE is a software application that implements the BGP/MPLS L3VPN PE control plane functionality.

In the case other control/signaling/messaging protocol are used, the route server is also the server of the particular protocol(s), it interacts with VPN forwarder.

3.3 Use of router reflector

Modern service networks can be very large in scale. For example, the number of VPNs routes in a very large data centers can pass the scale of those in SP backbone VPN networks. There are may be tens of thousands of end devices in a single service network.

Use of Router Reflector (RR) is necessary in large scale L3VPN networks to avoid full iBGP mesh among all vPEs and PEs. The L3 VPN routes can be partitioned to a set of RRs, the partition techniques are detailed in [RFC4364].

When RR is residing in a physical device, e.g., a server, which is

partitioned to support multi-functions and client/applications VMs, the RR becomes virtualized RR (vRR). Since RR's performs control plane only, a physical or virtualized server with large scale of computing power and memory can be a good candidate as host of vRRs. The vRR can also reside be in Gateway PE, or in an end device. Redundant RR design is even more important in when using vRR.

3.4 Use of RT constraint

The Route Target Constraint (RT Constraint, RTC) [RFC4684] is a powerful tool for VPN selective L3VPN route distribution. With RT Constraint, only the BGP receiver (e.g, PE/vPE/RR/vRR/ASBRs, etc.) with the particular L3VPNs will receive the route update for the corresponding VPNs. It is critical to use RT constraint to support large scale L3VPN development.

4. Forwarding Plane

4.1 Virtual Interface

Virtual Interface (VI) is an interface in an end device which is used for connecting the vPE to the application VMs in the end device. The latter cab be treated as CEs in the regular L3VPN's view.

4.2 VPN forwarder

VPN Forwarder is the forwarding component of a vPE.

The VPN forwarder location options:

- 1) within the end device where the virtual interface and application VMs are.
- 2) in an external device which the end device connect to, for example, a Top of the Rack (ToR) in a data center.

Multiple factors should be considered for the location of the VPN forwarder, including device capability, overall solution economics, QoS/firewall/NAT placement, optimal forwarding, latency and performance, operation impact, etc. There are design trade offs, it is worth the effort to study the traffic pattern and forwarding looking trend in your own unique service network as part of the exercise.

4.3 Encapsulation

There are two existing standardized encapsulation/forwarding options for BGP/MPLS L3VPN.

1. MPLS Encapsulation, [RFC3032].

2. Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE), [RFC4023].

The most common BGP/MPLS L3VPNs deployment in SP networks are using MPLS forwarding. This requires MPLS, e.g., Label Switched Protocol (LDP) [RFC5036] to be deployed in the network. It is proven to scale, and it comes with various security mechanisms to protect network against attacks.

However, the service network environment, such as a data center, is different than Service Provider VPN networks or large enterprise backbones. MPLS deployment may or may not be feasible. Two major challenges for MPLS deployment in this new environment: 1) the capabilities of the end devices and the transport/forwarding devices; 2) the workforce skill set.

Encapsulating MPLS in IP or GRE tunnel [RFC4023] may often be more practical in most data center, and computing environment. Note that when IP encapsulations are used, the associated security considerations must be analyzed carefully.

In addition, there are new encapsulation proposals for service network/Data center currently as work in progress in IETF, including several UDP based encapsulations proposals and some TCP based proposal. These overlay encapsulations can be suitable alternatives for a vPE, considering the availability and leverage of support in virtual and physical devices.

4.4 Optimal forwarding

As reported by many large cloud service operators, the traffic pattern in their data centers were dominated by East-West across subnet traffic (between the end device hosting different applications in different subnets) than North-South traffic (going in and out the DC to the WAN) or switched traffic within subnets. This is a primary reason that many large scale new design has moved away from traditional L2 design to L3.

When forwarding the traffic within the same VPN, the vPE should be able to provide direct communication among the VMs/application senders/receivers without the need of going through gateway devices. If it is on the same end device, the traffic should not need to leave the same device. If it is on different end device, optimal routing should be applied.

When multiple VPNs need to be accessed to accomplish the task the

user requested (this is common too), the end device virtual interfaces should be able to directly access multiple VPNs via use of extranet VPN techniques without the need of Gateway facilitation. Use BGP L3VPN policy control mechanisms to support this function.

5. Addressing

5.1 IPv4 and IPv6 support

Both IPv4 and IPv6 should be supported in the virtual PE solution.

This may present challenging to older devices, but may not be issues to newer forwarding devices and servers. A server is replaced much more frequently than a network router/switch in the infrastructure network, newer equipment should be capable of IPv6 support.

5.2 Address space separation

The addresses used for L3VPNs in the service network should be in separate address blocks than the ones used the underlay infrastructure of the service network. This practice is to protect the service network infrastructure being attacked if the attacker gain access of the tenant VPNs.

Similarity, the addresses used for the service network, e.g., a cloud service center of a SP, should be separated from the WAN backbone addresses space, for security reasons.

6.0 Inter-connection considerations

There are also deployment scenarios that L3VPN may not be supported in every segment of the networks to provide end-to-end L3VPN connectivity, a L3VPN vPE may be reachable only via an intermediate inter-connecting network, interconnection may be needed in these cases.

When multiple technologies are employed in the overall solution, a clear demarcation should be preserved at the inter-connecting points. The problems encountered in one domain should not impact the other domains.

From L3VPN point of view: A L3VPN vPE that implements [RFC4364] is a component of L3VPN network only. A L3VPN VRF on physical PE or vPE contains IP routes only, including routes learnt over the locally attached network.

As described earlier in this document, the L3VPN vPE should ideally be located as close to the "customer" edge devices. For cases, where

this is not possible, simple existing "L3VPN CE connectivity" mechanisms should be used, such as static, or direct VM attachments such as described in the vCE option below.

Consider the following scenarios when BGP MPLS VPN technology is considered as whole or partial deployment:

Scenario 1: All VPN sites (CEs/VMs) support IP connectivity. The best suited BGP solution is to use L3 VPNs [RFC4364] for all sites with PE and/or vPE solutions. This is a straightforward case.

Scenario 2: Legacy layer 2 connectivity must be supported in certain sites/CEs/VMs, and the rest sites/CEs/VMs need only 3 connectivity.

One can consider to use combined vPE and vCE solution to solved the problem. Use L3VPN for all sites with IP connectivity, and use a physical or virtual CE (vCE, may reside on the end device) to aggregate the L2 sites which, for example, are in a single container in a data center. The CE/vCE can be considered as inter-connecting point, where the L2 network are terminated and the corresponding routes for connectivity of the L2 network are inserted into L3VPN VRF. The L2 aspect is transparent to the L3VPN in this case.

Reducing operation complicity and maintaining the robustness of the solution are the primary reasons for the recommendations.

7. Security Considerations

vPE solution presented a virtualized L3VPN PE model. There are potential implications to L3VPN control plane, forwarding plane, and management plane. Security considerations are currently under study, will be included in the future revisions.

8. IANA Considerations

None.

9. References

9.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed.,
"Encapsulating MPLS in IP or Generic Routing Encapsulation
(GRE)", RFC 4023, March 2005.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A
Border Gateway Protocol 4 (BGP-4)", RFC 4271, January
2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private
Networks (VPNs)", RFC 4364, February 2006.
- [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, November 2006.
- [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed.,
"LDP Specification", RFC 5036, October 2007.
- [I-D.ietf-l3vpn-end-system] Marques, P., Fang, L., Pan, P., Shukla,
A., Napierala, M., "BGP-signaled end-system IP/VPNs",
draft-ietf-l3vpn-end-system-00, October 2012.

9.2 Informative References

- [I-D.fang-l3vpn-end-system-req] Napierala, M., and Fang, L.,
"Requirements for Extending BGP/MPLS VPNs to End-Systems",
draft-fang-l3vpn-end-system-requirements-00, Oct. 2012.
- [I-D.ward-irs-framework] Atlas, A., Nadeau, T., Ward, D., "Interface
to the Routing System Framework", draft-ward-irs-
framework-00, July 2012.
- [I-D.rfernando-irs-fw-req] Fernando, R., Medved, J., Ward, D., Atlas,
A., Rijsman, B., "IRS Framework Requirements", draft-
rfernando-irs-framework-requirement-00, Oct. 2012.

Authors' Addresses

Luyuan Fang
Cisco
111 Wood Ave. South
Iselin, NJ 08830

Email: lufang@cisco.com

David Ward
Cisco
170 W Tasman Dr
San Jose, CA 95134
Email: wardd@cisco.com

Rex Fernando
Cisco
170 W Tasman Dr
San Jose, CA
Email: rex@cisco.com

Maria Napierala
AT&T
200 Laurel Avenue
Middletown, NJ 07748
Email: mnapierala@att.com

Nabil Bitar
Verizon
40 Sylvan Road
Waltham, MA 02145
Email: nabil.bitar@verizon.com

Dhananjaya Rao
Cisco
170 W Tasman Dr
San Jose, CA
Email: dhrao@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 11, 2013

M. Pathak
Affirmed Networks
K. Patel
A. Sreekantiah
Cisco Systems
July 10, 2012

Inter-AS Option D for BGP/MPLS IP VPN
draft-mapathak-interas-option-d-00.txt

Abstract

This document describes a new option known as an Inter-AS option D to the 'Multi-AS Backbones' section of [RFC4364]. This option combines VPN VRFs at the Autonomous System Border Router (ASBR) as described in 'Option A' with the distribution of labeled VPN-IP routes as described in 'Option B'. In addition, this option allows for a data plane consisting of two methods of traffic forwarding between attached ASBR pairs.

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 <http://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 11, 2013.

Copyright Notice

Copyright (c) 2012 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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	4
1.1. Requirements Language	5
2. Scope	5
3. Inter-AS Option D Reference Model	5
4. Private Interface Operation without Carrier's Carrier (CSC)	7
5. Private Interface Forwarding with CSC	8
6. Shared Interface Forwarding	11
7. Route Advertisement to External BGP Peers	12
7.1. Route Advertisement - Private interface forwarding	12
7.2. Route Advertisement - Shared interface forwarding	13
7.3. Route Advertisement to Internal BGP Peers	14
8. Option D Operation Requirements	14
8.1. Inter-AS IP VPN Route Distribution	14
8.2. Private Interface Forwarding Route Distribution	14
8.3. Shared interface forwarding Route Distribution	14
9. Inter-AS Quality of Service for Option D	15
10. Security Considerations	15
11. Acknowledgements	15
12. References	16
12.1. Normative References	16
12.2. Informative References	16
Authors' Addresses	16

1. Introduction

MPLS VPN providers often need to inter-connect different ASes to provide VPN services to customers. This requires the setting up of Inter-AS connections at ASBRs. The inter-AS connections may or may not be between different providers. The mechanisms to set up inter-as connections are described in [RFC4364]. Of particular interest for this draft are the ones documented in section 10 of [RFC4364].

For the option described in section 10, part (a) of [RFC4364], commonly referred to as Option A, peering ASBRs are connected by multiple sub-interfaces, with at least one interface for each VPN that spans the two ASes. Each ASBR acts as a PE, and thinks that the other ASBR is a CE. The ASBRs associate each sub-interface with a VRF and a BGP session is established per sub-interface to signal IP (unlabeled) prefixes. As a result, traffic within the VPN VRFs is IP. The advantage of this option is that the VPNs are isolated from each other and since the traffic is IP, QoS mechanisms that operate on IP traffic can be applied to achieve customer SLAs. The drawback of this option is that there needs to be one BGP session per sub-interface (and at least one sub-interface per VPN), which can be a potential scalability concern if there are a large number of VRFs.

For the option described in section 10, part (b) of [RFC4364], commonly referred to as Option B, peering ASBRs are connected by one or more sub-interfaces that are enabled to receive MPLS traffic. An MP-BGP session is used to distribute the labeled VPN prefixes between the ASBRs. Therefore, the traffic that flows between them is labeled. The advantage of this option is that it's more scalable, as there is no need to have one sub-interface and BGP session per VPN. The drawback of this option is that QoS mechanisms that can only be applied to IP traffic cannot be used as the traffic is MPLS. There is also no isolation between the VRFs.

The solution described in this draft aims to address the scalability concerns of Option A by using a single BGP session to signal VPN prefixes. In this solution, the forwarding connections between the ASBRs are maintained on a per-VRF basis, while the control plane information is exchanged using a single MP-BGP session.

If the solution is used between any attached ASBR pairs belonging to separate Autonomous Systems (AS), then VRF based route filtering policies via RTs is achieved directly between ASBR pairs, independent of PE based RT filtering within an AS.

1.1. Requirements Language

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 RFC 2119 [RFC2119].

2. Scope

The Inter-AS VPN option described in this draft is applicable to both, the IPv4 VPN services described in [RFC4364] and the IPv6 VPN services defined in [VPN-IPv6]. It is NOT applicable to MVPN IPv4 and MVPN IPv6 services defined in [RFC6513]. Support of existing 'Multi-AS' options, along with the new techniques are beyond the scope of this document.

3. Inter-AS Option D Reference Model

Figure 1 shows an arbitrary Multi-AS VPN interconnectivity scenario between Customer Edge routers. CE1 and CE3, interconnected by Service Providers SP1 and SP2, belong to the same VPN, say Red. CE2 and CE4 belong to a different VPN, say Green. This example shows 3 interfaces ('red', 'white' and 'green') between ASBR1 (belonging to SP1) and ASBR2 (belonging to SP2).

Interface 'red' is a VRF attachment circuit associated to VRF1 (on ASBR1 and ASBR2) for VPN Red and is used to transport labeled or native IP VPN traffic between VRF pairs. Similarly, interface 'green' is a VRF attachment circuit associated to VRF2 (on ASBR1 and ASBR2) for VPN Green and is used to transport labeled or native IP VPN traffic between VRF pairs. Interface 'white' is not associated with any VRF instances i.e. this interface is 'global' in nature (in the context of the connected ASes) and carries as a minimum all ASBR exported VPN-IP routing updates.

We shall use the term "private interface forwarding" to describe the model where packets for the "Red" VPN are forwarded on the "red" interface, while packets belonging to "Green" VPN are forwarded on the "green" interface. There are no BGP sessions running on the "red" and "green" interfaces; rather the 'white' interface carries all ASBR VPN-IP routing updates exported from VRF pairs. We shall use the term "shared interface forwarding" to describe the model where the "white" interface will be used to forward all the traffic between the ASBRs. For shared interface forwarding outside of a VRF context, interfaces 'red' and 'green' are not required. In addition to carrying all ASBR VPN-IP routing updates, interface 'white' transports labeled IP VPN traffic or native IP traffic. IP VPN

packets entering or leaving the ASBR via this interface may be forwarded using normal MPLS mechanisms (e.g. through use of the LFIB) or through a lookup within a VRF that has been identified via MPLS label values.

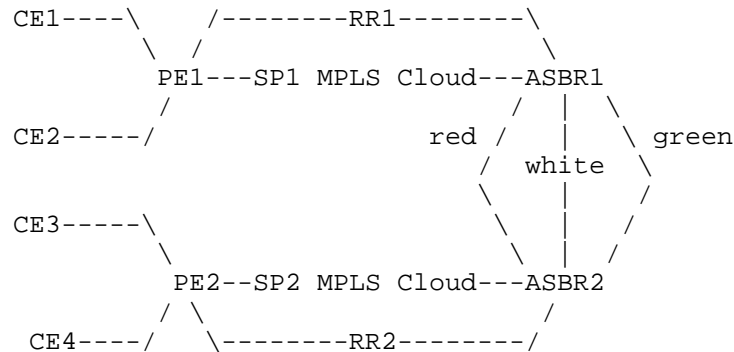


Figure 1

In the diagram above:

1. CE1 and CE3 belong to VPN Red.
2. CE2 and CE4 belong to VPN Green.
3. PE1 uses RDs RD-red1 and RD-green1 for VPN Red (VRF Red) and VPN Green (VRF Green) respectively.
4. PE2 uses RDs RD-red2 and RD-green2 for VPN Red (VRF Red) and VPN Green (VRF Green) respectively.
5. ASBR1 has VRFs Red and Green provisioned with RD-red3 and RD-green3 respectively.
6. ASBR2 has VRFs Red and Green provisioned with RD-red4 and RD-green4 respectively.
7. There are 3 interfaces between ASBR1 and ASBR2.
8. On each ASBR, one interface is associated with VRF Red and one with VRF Green. These are the interfaces marked "red" and "green" respectively.
9. There is a third interface over which there is an MP-BGP session

between the ASBRs. This is the interface marked "white".

10. VPN route importing is achieved by configuring the appropriate RTs.

11. The PE and ASBR routers in each AS peer with a route-reflector in that AS.

The following sections describe in detail the different modes of operation for Option D.

4. Private Interface Operation without Carrier's Carrier (CSC)

This section describes how route distribution and packet forwarding takes place when using the private interface forwarding option without the use of CSC, ie. the traffic sent between the private interfaces is unencapsulated.

Route Distribution:

[The following description is for VPN Red, but Route Distribution for VPN Green is exactly analogous to this]

1. CE1 advertises a prefix N to PE1.
2. PE1 advertises a VPN prefix RD-red1:N to RR1, which in turn advertises it to ASBR1 via iBGP.
3. ASBR1 imports the prefix into VPN Red and creates a prefix RD-red3:N.
4. ASBR1 advertises the imported prefix RD-red3:N to ASBR2. It sets itself as the next-hop for this prefix and also allocates a local label that is signaled with the prefix.
5. By default, ASBR1 does not advertise the source prefix RD-red1:N to ASBR2. This advertisement is suppressed as the prefix is being imported into an Option D VRF.
6. ASBR2 receives the prefix RD-red3:N and imports it into VPN Red as RD-red4:N.
7. While installing the prefix into the VRF Red RIB table, ASBR2 sets the nexthop of RD-red4:N to ASBR1's interface address in VRF Red. The routing context for this entry is also set to that of VRF Red.
8. While installing the MPLS forwarding entry for RD-red4:N, by

default, the label that was advertised by ASBR1 for the prefix is not installed in the Forwarding Information Base. This enables the traffic between the ASBRs to be IP.

9. ASBR2 advertises the imported prefix RD-Red4:N to RR2, which in turn advertises it to PE2. It sets itself as the next-hop for this prefix and also allocates a local label that is signaled as part of the VPN-IP NLRI.

10. By default, ASBR2 does not advertise the source prefix RD5:N to PE2. This advertisement is suppressed.

11. PE2 imports the RD-red4:N into VRF Red as RD-red2:N.

Packet Forwarding

The packet forwarding would work just as it would in an Option A scenario:

1. CE3 sends a packet destined for N to PE2.
2. PE2 encapsulates the packet with the VPN label allocated by ASBR2 and the IGP label (if any) needed to tunnel the packet to ASBR2.
3. The packet arrives on ASBR2 with the VPN Label, ASBR2 pops the VPN Label and sends the packet as IP to ASBR1 on the "red" interface.
4. The IP packet arrives at ASBR1 on the "red" interface. ASBR1 then encapsulates the packet with the VPN Label allocated by PE1 and the IGP label needed to tunnel the packet to PE1.
5. The packet arrives on PE1 with the VPN label; PE1 disposes the VPN label and forwards the IP packet to CE1.

5. Private Interface Forwarding with CSC

Let's assume that VPN Red is used to provide VPN service to its customer carrier who in turn provides a VPN service to a customer. This implies that VPN RED is used to provide an LSP between the PE (PE3 and PE4) loopbacks of the baby carrier in the following topology:

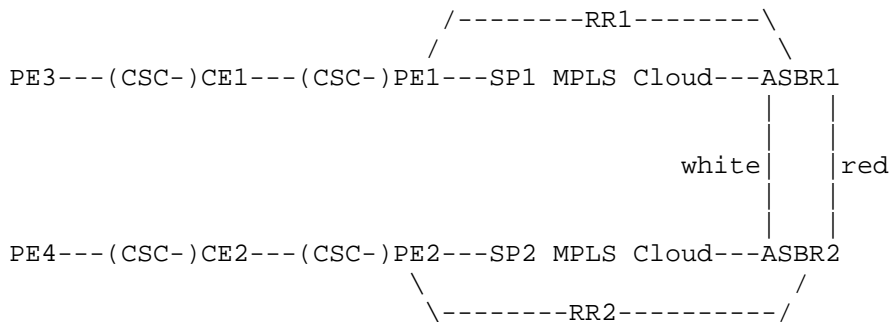


Figure 2

Thus, let's assume that in the diagram above:

1. CSC-PE1 uses RD RD-red1 for VPN Red (VRF Red).
2. CSC-PE2 uses RD RD-red2 for VPN Red (VRF Red).
3. ASBR1 has VRF Red provisioned with RD-red3.
4. ASBR2 has VRF Red provisioned with RD-red4.
5. There are 2 interfaces between ASBR1 and ASBR2.
6. On each ASBR, one interface is associated with VRF Red. This is the interface marked "red" in the Figure 2.
7. There is a second interface over which there is an MP-BGP session between the ASBRs. This interface is in the global context and is marked "white" in the figure.

Route Distribution:

1. CSC-CE1 advertises PE3s loopback N to PE1.
2. CSC-PE1 advertises a VPN prefix RD-red1:N to RR1, which advertises it to ASBR1 via MP-iBGP.
3. ASBR1 imports the prefix into VPN Red and creates a prefix RD-red3:N.
4. ASBR1 advertises the imported prefix RD-red3:N to ASBR2. It sets itself as the next-hop for this prefix and also allocates a local label that is signaled as part of the VPN-IP NLRI.
5. By default, ASBR1 does not advertise the source prefix RD-red1:N

to ASBR2. This advertisement is suppressed as the prefix is being imported into an Option D VRF.

6. ASBR2 receives the prefix RD-red3:N and imports it into VPN Red as RD-red4:N.

7. While installing the prefix into the VRF Red RIB table, ASBR2 sets the nexthop of RD-red4:N to ASBR1's interface address in VRF Red. The nexthop routing context is also set to that of VRF Red.

8. While installing the MPLS forwarding entry for RD-red4:N, the outgoing label is installed in forwarding. This enables the traffic between the ASBRs to be MPLS.

9. ASBR2 advertises the imported prefix RD-red4:N to RR2, which advertises it to CSC-PE2. It sets itself as the next-hop for this prefix and also allocates a local label that is signaled as part of the VPN-IP NLRI.

10. By default, ASBR2 does not advertise the source prefix RD-red4:N to PE2. This advertisement is suppressed.

11. PE2 imports the RD-red4:N into VRF Red as RD-red2:N.

Packet Forwarding:

1. PE4 sends a MPLS packet destined for N to CSC-CE2.

2. CSC-CE2 swaps the MPLS label and sends a packet destined for N to CSC-PE2.

3. CSC-PE2 encapsulates the packet with the VPN label allocated by ASBR2 and the IGP label needed (if any) to tunnel the packet to ASBR2.

4. The packet arrives on ASBR2 with the VPN Label, ASBR2 swaps the received VPN label with the outgoing label received from ASBR1 and sends the MPLS packet on to the VRF Red interface.

5. The MPLS packet arrives at ASBR1 on the VRF red interface, ASBR1 then swaps the received MPLS label with a label stack consisting of the VPN Label allocated by PE1 and the IGP label needed to tunnel the packet to CSC-PE1.

6. The packet arrives on CSC-PE1 with the VPN label; PE1 disposes the VPN label and forwards the MPLS packet to CSC-CE1.

7. CSC-CE1 in turn swaps the label and forwards the labeled packet

to PE3.

6. Shared Interface Forwarding

This section describes how route distribution and packet forwarding takes place when using the shared interface forwarding option. The topology is the same as in Figure 1.

Route Distribution (VPN Red):

1. CE1 advertises a prefix N to PE1.
2. PE1 advertises a VPN prefix RD-red1:N to RR1, which advertises it to ASBR1 via iBGP.
3. ASBR1 imports the prefix into VPN Red and creates a prefix RD-red3:N
4. ASBR1 advertises the imported prefix RD-red3:N to ASBR2. It sets itself as the next-hop for this prefix and also allocates a local label that is signaled with the prefix.
5. By default, ASBR1 does not advertise the source prefix RD-red1:N to ASBR2. This advertisement is suppressed as the prefix is being imported into an Option D VRF.
6. ASBR2 receives the prefix RD-red3:N and imports it into VPN Red as RD-red4:N
7. While installing the prefix into the VRF Red RIB table, ASBR2 retains the nexthop of RD-red4:N as received in the BGP update from ASBR1. This is the address of ASBR1's shared interface address in the global table. The nexthop routing context is also left unchanged and corresponds to that of the global table.
8. While installing the MPLS forwarding entry for RD-red4:N, the outgoing label is installed in forwarding. This enables the traffic between the ASBRs to be MPLS.
9. ASBR2 advertises the imported prefix RD-red4:N to RR2, which advertises it to PE2. It sets itself as the next-hop for this prefix and also allocates a local label that is signaled as part of the VPN-IP NLRI.
10. By default, ASBR2 does not advertise the source prefix RD-red4:N to PE2. This advertisement is suppressed.

11. PE2 imports the RD-red4:N into VRF Red as RD-red2:N.

Packet Forwarding:

The packet forwarding would work just as it would in an Option B scenario:

1. CE3 sends a packet destined for N to PE2.
2. PE2 encapsulates the packet with the VPN label allocated by ASBR2 and the IGP label needed to tunnel the packet to ASBR2.
3. The packet arrives on ASBR2 with the VPN Label. ASBR2 swaps the received VPN label with the outgoing label received from ASBR1 and sends the MPLS packet on to the global shared link interface.
4. The MPLS packet arrives at ASBR1 on the global shared link interface. ASBR1 then swaps the received MPLS label with a label stack consisting of the VPN Label allocated by PE1 and the IGP label needed to tunnel the packet to PE1.
5. The packet arrives on PE1 with the VPN label; PE1 disposes the VPN label and forwards the IP packet to CE1.

7. Route Advertisement to External BGP Peers

ASBR1 (refer Figure 1) does route advertisement and VPN route processing using the standard BGP-VPN rules. It processes the VRF Red RT extended community attributes and learns the label bindings associated with routes for VPN Red. VPN-IP routes are imported into VRF Red's Routing Information Base (RIB) where their RT and RD attributes, assigned by PE1 are removed.

ASBR1 VPN-IP routes are not advertised to RR1 as the original routes applicable to VPN Red sourced by PE1 were received from an internal BGP peer. Any installed routes that are not imported into VRF1 RIB MAY be advertised to external BGP peers using the existing [RFC4364] Multi-AS "Option B" techniques. Dependant on which packet forwarding method is used, routes are then exported from VRFs and advertised from ASBR1 to ASBR2 as described in the following sections.

7.1. Route Advertisement - Private interface forwarding

VPN-IP prefixes are advertised from ASBR1 to ASBR2 via a BGP session that is in the global routing table context. This implies that the advertised next-hop address is also reachable via the global routing table context. In order to force traffic to be forwarded via an

interface 'red' that is in a VRF routing table context, VRF forwarding entries need to be installed using a next-hop address that is in VRF Red's (which resides on ASBR2) routing context. The address of the next-hop could be the same as the global table address of the remote ASBR (in this case ASBR1), although this would require that the same IP address be used across all VRF attachment circuits linking ASBR pairs.

Alternatively, if a Service Provider needs to number the VRF interfaces differently from the global table VPN session, a configuration method SHOULD be available so as to map the corresponding global table VPNv4 neighbor address to an IP address reachable in the given VRF.

ASBR1 exports routes associated to VPN Red from VRF Red's RIB to BGP where RD and RT attributes, plus label bindings are attached to these routes. These labeled VPN-IP routes are advertised across interface 'red' to ASBR2 via BGP, with a label value set to implicit-null and the 'S' bit set. The routes next-hop addresses is set either to ASBR1 (usually interface 'red') or an address reachable via interface 'red'. ASBR2 imports the VRF Red's exported routes into VRF Red's RIB where the routes RT and RD attributes are removed. The next-hop of the imported routes is either set via a policy or left unchanged to an address in VRF Red's routing context. ASBR2 exports routes from VRF Red's RIB to BGP and attaches RT and RD attributes, as configured at VRF Red plus label bindings. Labeled VPN-IP routes are now advertised to PE2 via RR2 and so on. ASBR2 sets itself as the nexthop for these routes and allocates a local label. As an optimization to conserve label space, ASBR2 MAY allocate a per-VRF aggregate label as the local label while advertising the routes to iBGP peers.

7.2. Route Advertisement - Shared interface forwarding

ASBR1 exports routes associated to VPN Red from VRF Red's RIB to BGP where RD and RT attributes, plus label bindings are attached to these routes. These labeled VPN-IP routes are advertised across interface 'white' to ASBR2 via BGP, with a next-hop set to ASBR1. ASBR2 imports the VRF Red exported routes into (its local) VRF Red RIB where the routes RT and RD attributes are removed. The imported routes next-hop is set to ASBR1, available via interface 'white'. ASBR2 exports routes from VRF Red's RIB to BGP and attaches RT and RD attributes, as configured at VRF Red plus label bindings. Labeled VPN-IP routes are now advertised to PE2 via RR2 and so on.

7.3. Route Advertisement to Internal BGP Peers

All the received VPN-IP routes from an adjacent ASBR are imported into local VRFs on the receiving ASBR. The receiving ASBR needs to advertise these routes to its local IBGP sessions. The next-hop for these routes SHOULD be set to itself when the local ASBR advertises these routes to its IBGP sessions.

8. Option D Operation Requirements

8.1. Inter-AS IP VPN Route Distribution

Routes received from internal or external peers that are imported into ASBR VRFs SHOULD NOT be readvertised to any BGP neighbors. Routes that are not imported into VRFs but are installed in the ASBR's global routing table MAY be readvertised using existing Option 'B' techniques as described in the Multi-AS section of [RFC4364]. The ASBR MUST be equipped with RT based filtering mechanisms that explicitly permit all or a subset of such RT values to be readvertised to its neighbors.

VPN-IP routes that are converted by the ASBR MUST NOT be readvertised to the source peer of the route. It SHOULD be possible to remove/manipulate individual RT values when advertising routes on a per neighbor basis. This is useful where the SP wants to separate RT values advertised to EBGP peers from RT values advertised to IBGP peers.

8.2. Private Interface Forwarding Route Distribution

For private interface forwarding, labeled VPN-IP routes advertised from ASBR to ASBR MUST have their next-hop set to an address within a VRF RIB. This address will usually be the VRF attachment circuit interface.

If the Service Provider needs to number the VRF interfaces differently from the global table VPNv4 neighbor, a configuration method SHOULD be available so as to map the corresponding global table VPNv4 neighbor address to an IP address reachable in the given VRF. This route mapping policy SHOULD be configurable on both outbound and inbound peers.

8.3. Shared interface forwarding Route Distribution

For shared interface forwarding outside of a VRF context, the next-hop must be a 'global' (non VRF RIB) address on an ASBR. This address will usually be the interface linking ASBR pairs.

9. Inter-AS Quality of Service for Option D

It SHOULD be possible for the ASBR as a DS boundary node [DS-ARCH] operating traffic classification and conditioning functions to match on ingress and egress a combination of application (TCP, UDP port, RTP session, data pattern etc), IP Source Address, IP Destination Address or DS field per packet, per VRF or per VRF attachment circuit (in the case of private interface forwarding).

Once matched, the packets Layer-2 header (if applicable), IP DSCP and MPLS EXP bits or combinations of the above should be capable of being re-marked, and optionally shaped per traffic stream, depending on the DS domain's Traffic Conditioning Agreement (TCA). This will assist where different DS domains have different TCA rules.

For Private interface forwarding, the ASBR should be capable of forwarding explicit null labeled MPLS packets across VRF attachment circuits. This SHOULD assist with a pipe mode [DIFF-TUNNEL] operation of traffic conditioning behavior at the ASBR. MPLS based forwarding between attached ASBRs inherently should have these mechanisms built in.

10. Security Considerations

This document does not alter the underlying security properties of BGP based VPNs. In particular, the the private interface forwarding using a new Multi-AS option defined in this document has same security implications as Multi-AS option 'a' of [RFC4364]. The global interface forwarding using a new Multi-AS option defined in this document is outside the scope of this document.

This document does not alter the underlying security properties of BGP based VPNs for the shared interface forwarding using the new Multi-AS option. The security implications for this mechanism are same as Multi-AS option 'b' of [RFC4364].

11. Acknowledgements

The authors wish to acknowledge the contributions of the authors of the original Option D draft: Marko Kulmala, Ville Hallivuori, Jyrki Soini, Jim Guichard, Robert Hanzl and Martin Halstead. The authors would like to thank Eric Rosen for his comments.

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

12.2. Informative References

- [RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.

Authors' Addresses

Manu Pathak
Affirmed Networks
35 Nagog Park
Acton, MA 01720
USA

Email: manu_pathak@affirmednetworks.com

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

Arjun Sreekantiah
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: asreekan@cisco.com

INTERNET-DRAFT
Intended Status: Informational
Expires: April 25, 2013

R. Fernando
D. Rao
L. Fang
Cisco
October 22, 2012

Virtual Service Topologies in BGP VPNs
draft-rfernando-virt-topo-bgp-vpn-01

Abstract

This document presents techniques that build on MPLS/VPN control plane mechanisms to construct virtual service topologies in data centers. These virtual service topologies interconnect network zones and help to constrain the flow of traffic that go between zones so that interesting services can be applied to them.

The techniques suggested are required to create a rich overlay network that mimics topology and routing functions of physical networks. Steps to create a virtual service topology and the ability to constrain routing and traffic to flow in this topology are outlined in this document.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 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	3
1.1	Terminology	3
2	Intra-Zone Routing and Traffic Forwarding	3
3	Inter-Zone Routing and Traffic Forwarding	4
4	Proposed Inter-Zone Model	5
4.1	Constructing the Virtual Service Topology	5
4.2	Inter-zone Routing and Service Chaining	7
5	Routing Considerations	8
5.1	Multiple service topologies	8
5.2	Multipath	8
5.3	Supporting redundancy	9
5.4	Route Aggregation	9
6	Security Considerations	10
7	IANA Considerations	10
8	Acknowledgements	10
9	References	10
9.1	Normative References	10
	Authors' Addresses	10

1 Introduction

Network topologies and routing in the enterprise, data center and campus networks reflect the needs of the organization in terms of performance, scale, security and availability. For scale and security reasons, networks are composed of multiple small domains or zones each serving one or more logical functions of the organization.

Hosts within a zone can freely communicate with one another but traffic between hosts in different zones is subjected to additional services that help in scaling and securing the end applications. Traditional networks achieve this using a combination of physical topology constraints and routing.

Porting a traditional network with all its functions and infrastructure elements to a virtualized data center requires network overlay mechanisms that provide the ability to create virtual network topologies that mimic physical networks and the ability to constrain the flow of routing and traffic over these virtual network topologies.

Furthermore, data centers might need multiple virtual topologies per tenant to handle different types of application traffic. Each tenant might dictate a different topology of connectedness between their zones and applications and might need the ability to apply network policies and services for inter-zone traffic in manner specific to their organizational objectives. Therefore, the mechanisms devised should be flexible to accommodate the custom needs of a tenant and their applications at the same time robust enough to satisfy the scale, performance and HA needs that they demand from the virtual network infrastructure.

Towards this end, this document introduces the concept of virtual service topologies and extends MPLS/VPN control plane mechanisms to constrain routing and traffic flow over virtual service topologies.

1.1 Terminology

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 RFC 2119 [RFC2119].

2 Intra-Zone Routing and Traffic Forwarding

This section provides a brief overview of how L3VPNs [RFC4364] can be used in data center networks to create a single zone to host customer

applications. The subsequent sections in the document builds on this base model to create richer topologies by interconnecting these zones and enforcing services for inter-zone traffic.

In a DC, servers host virtual machines where end applications reside. A collection of VMs that can communicate freely form a zone.

The notions of L3VPN when applied to the virtual data center works in the following manner.

The VM that runs the applications that is the CE. A CE/VM belongs to a zone. As in traditional L3VPN, the PE is the first hop router from the CE/VM and the PE-CE link is single hop from an L3 perspective. Any of the available physical, logical or tunneling technologies can be used to create this "direct" link between the CE/VM and its attached PE(s).

The PE helps create the zone that the CE belongs to by placing the CE-PE link in a VRF corresponding to that zone. Intra-zone connectivity is achieved by designating an RT per zone (zone-RT) that is applied on all PE VRFs that terminate the CE/VMs that belong to the zone.

It is further assumed that the CE/VM's are associated with network policies that get activated on an attached PE when a CE/VM becomes alive. These policies dictate how networking should be set up for the CE/VM including the properties of the CE-PE link, the IP address of the CE/VM, the zone(s) that it belongs to, QoS policies etc. There are many ways to achieve this step, a description of which is outside the scope of this proposal.

When the CE/VM gets activated the attached PE starts exporting its IP address with the corresponding zone-RT. This creates a full mesh connectivity between the newly active VM and the rest of the VMs in the zone.

Note that the IP address mask of the CE/VM need not necessarily be a /32. This is the case when the CE/VM's in a zone belong to a single IP subnet. The PE, in this case, would use proxy-arp to resolve ARP's for remote destinations in the IP subnet and use L3VPN style forwarding to carry traffic between the VMs.

3 Inter-Zone Routing and Traffic Forwarding

A simple form of inter-zone traffic forwarding can be achieved using extranets or hub-and-spoke L3VPN configurations. However, the ability

to enforce constrained traffic flow through a set of services is non-existent in extranets and is limited in hub-and-spoke setups.

Note that the inter-zone services cannot always be assumed to reside and inlined on a PE. There is a need to virtualize the services themselves so that they can be implemented on commodity hardware and scaled out 'elastically' when traffic demands increase. This creates a situation where services for traffic between zones may not be applied only at the source-zone PE or the destination-zone PE. Mechanisms are required that make it easy to direct inter-zone traffic through the appropriate set of service nodes that might be remote and virtualized.

A service node for the purposes of this proposal is a physical or virtual service appliance that inspects and/or impacts the flow of inter-zone traffic. Firewalls, load-balancers, deep packet inspectors are examples of service nodes. Service nodes are CE's attached to a service-PE.

A service-PE is a normal L3VPN PE that recognizes and directs the appropriate traffic flows to its attached service nodes through VPN label lookup. Service nodes may be integrated or attached to service-PE's.

A sequence of service-PE's and the corresponding service nodes create a service chain for inter-zone traffic. The service chain is unidirectional and creates a one way traffic flow between source zone and destination zone. The service PE closest to the source zone is the source service-PE and the service PE closest to the destination zone is called the destination service-PE.

4 Proposed Inter-Zone Model

The proposed model has two steps to it.

4.1 Constructing the Virtual Service Topology

The first step involves creating the virtual service topology that ties two or more zones through one or more service nodes.

This is done by originating a service topology route that creates the route resolution state for the zone prefixes in a set of service-PEs. The service topology route is originated in the destination service-PE. It then propagates through the series of service-PE's from the destination service-PE to the source service-PE.

A modification is proposed to the service-PE behavior to allow the automatic and constrained propagation of service topology routes through the service-PE's that form the service chain. A service-PE in a given service chain is provisioned to accept the service topology route and re-originate it such that the upstream service-PE imports it and so on. The sequential import and export of the service topology route along the service chain is controlled by RTs provisioned appropriately at each service-PE.

To create the service chain and give it a unique identity, each service-PE is provisioned with three service RT's for every service chain that it belongs to: {service-import-RT, service-export-RT, service-topology-RT}.

A service-import-RT acts exactly as a regular import RT importing any route that carries that RT into the service-VRF. Additionally, any route that was imported using the service-import-RT MUST be automatically re-originated with the corresponding service-export-RT.

The next-hop of the re-originated route points to the service node attached to the service-PE. The VPN label carried in the re-originated route directs all traffic received by the service-PE to the service node.

The service-export-RT of a downstream service-PE MUST be equal to the service-import-RT of the immediate upstream service-PE. The service topology route MUST be originated in the destination service-PE carrying its service-export-RT. The flow of the service topology route creates both the service chain as well as the route resolution state for the zone prefixes.

Finally, the presence of the service topology route in a service-PE triggers the addition of the service-topology-RT to the regular import RT's of the service-VRF. Every service chain has a single unique service-topology-RT that's provisioned in all participating service-PE's.

The three service RT's (import, export and topology) should not be reused for other purposes within the network. The service RT's that establish the chain and give it its identity can be pre-provisioned or activated due to the appearance of a attached virtual service node. The provisioning system is assumed to have the intelligence to create loop-free virtual service topologies.

There should be one service topology route per virtual service topology. There can be multiple virtual service topologies and hence service topology routes for a given VPN.

Virtual service topologies are constructed unidirectionally. Between the same pair of zones, traffic in opposite directions will be supported by two service topologies and hence two service topology routes. These two service topologies might or might not be symmetrical, i.e. they might or might not traverse the same service-PE's/service-nodes in opposite directions.

As noted above, a service topology route can be advertised with a per-next-hop label that directs incoming traffic to the attached service node. Alternatively, an aggregate label may be used for the service route and an IP route lookup done at the service-PE to send traffic to the service node.

Note that a new service node could be inserted seamlessly by just configuring the three service RT's in the attached service-PE. This technique could be used to elastically scale out the service nodes with traffic demand.

The distribution of the service topology route itself can be controlled by RT constraints [RFC4684] to only the interesting service-PE's.

Finally, note that the service topology route is independent of the zone prefixes which are the actual addresses of the VMs present in the various zones. The zone prefixes use the service topology route to resolve their next-hop.

4.2 Inter-zone Routing and Service Chaining

Routes representing hosts or VMs from a zone are called zone prefixes. A zone prefix will have its regular zone RTs attached when it is originated. This will be used by PEs in the same zone to import these prefixes to enable direct communication between VM's of the same zone.

In addition to the intra-zone RT's, zone prefixes are also tagged with the set of service-topology-RT's that they belong to at the point of origination.

Since they are tagged with the service-topology-RT, zone prefixes get imported into the appropriate service-VRF's of particular service-PE's that form the service chain associated to that topology RT. Note that the topology RT was added to the relevant service-VRF's import RT list during the virtual topology construction phase.

Once the zone prefixes are imported into the service-PE, their next-hops are resolved as follows.

- o If the importing service-PE is the destination service-PE, it uses the next-hop that came with the zone prefix for route resolution. It also uses the VPN label that came with the prefix.

- o If the importing service-PE is not the destination service-PE, it rewrites the received next-hop of the zone prefix with the service topology route.

In an MPLS VPN, the zone prefixes come with VPN labels. The labels also must be ignored when in the intermediate service-PEs. Instead, the zone prefix gets resolved via the service topology route and uses the associated service route's VPN label.

This way the zone prefixes in the intermediate service-PE hops recurse over the service topology route forcing the traffic destined to them flow through the virtual service topology.

Traffic for the zone prefix goes through the service hops created by the service topology route. At each service hop, the service-PE directs the traffic to the service node. Once the service node is done processing the traffic, it then sends it back to the service-PE which forwards the traffic to the next service-PE and so on.

A significant benefit of this next-hop indirection is to avoid redundant advertisement of zone prefixes from the service-PE's. Also, when the virtual service topology is changed (due to addition or removal of service-PEs), there should be no change to the zone prefix's import/export RT configuration.

Note that this proposal introduces a change in the behavior of the service-PE's but does not require protocol changes to BGP.

5 Routing Considerations

5.1 Multiple service topologies

A service-PE can support multiple distinct service topologies for a VPN.

5.2 Multipath

One could use all tools available in BGP to constrain the propagation and resolution state created by the service topology route. A service topology route can have multiple equal cost paths, for inter-zone traffic to get load-balanced over.

5.3 Supporting redundancy

For stateful services an active-standby mechanism could be used at the service level. In this case, the inter-zone traffic should prefer the active service node over the standby service node. At a routing level, this is achieved by setting up two paths for the same service topology route - one path goes through the active service node and the other through the standby service node. The active service path can then be made to win over the standby service path by appropriately setting the BGP path attributes of the service topology route such that the active path succeeds in path selection. This forces all inter-zone traffic through the active service node.

5.4 Route Aggregation

Instead of the actual zone prefixes being imported and used at various points along the chain, the zone prefixes may be aggregated at the destination service-PE and the aggregate zone prefix used in the service chain between zones. In such a case, it is the aggregate zone prefix that carries the service-topology-RT and gets imported in the service-PE's that comprise the service chain.

6 Security Considerations

This proposal does not change the security model of MPLS/VPN BGP.

7 IANA Considerations

This proposal does not have any IANA implications.

8 Acknowledgements

The authors would like to thank the following individuals for their review and feedback on the proposal: Paul Quinn, David Ward, Ashok Ganesan, Peter Bosch.

9 References

9.1 Normative References

[RFC4364] Rosen, E., "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC4364.

[RFC4684] Marques, P., "Constrained Route Distribution for Border Gateway Protocol/Multiprotocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)

Authors' Addresses

Dhananjaya Rao
Cisco
170 W Tasman Dr
San Jose, CA

Email: dhrao@cisco.com

Rex Fernando
Cisco
170 W Tasman Dr
San Jose, CA

Email: rex@cisco.com

Luyuan Fang
Cisco
170 W Tasman Dr
San Jose, CA

Email: lufang@cisco.com

Network working group
Internet Draft
Category: Informational

X. Xu
S. Hares
Huawei Technologies
Y. Fan
China Telecom
C. Jacquenet
France Telecom

Expires: April 2013

October 15, 2012

Virtual Subnet: A L3VPN-based Subnet Extension Solution

draft-xu-virtual-subnet-09

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on April 15, 2012.

Copyright Notice

Copyright (c) 2009 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.

Abstract

This document describes a Layer3 Virtual Private Network (L3VPN)-based subnet extension solution referred to as Virtual Subnet, which mainly reuses existing Border Gateway Protocol (BGP)/Multi-Protocol Label Switch (MPLS) IP Virtual Private Network (VPN)[RFC4364] and Address Resolution Protocol(ARP)/Neighbor Discovery (ND) proxy [RFC925][RFC1027][RFC4389] technologies. Virtual Subnet provides a scalable approach for interconnecting cloud data centers.

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 RFC-2119 [RFC2119].

Table of Contents

1. Introduction	3
2. Terminology	5
3. Solution Description.....	5
3.1. Unicast	5
3.1.1. Intra-subnet Unicast	5
3.1.2. Inter-subnet Unicast	6
3.2. Multicast	9
3.3. CE Host Discovery	9
3.4. ARP/ND Proxy	9
3.5. CE Host Mobility	10
3.6. Forwarding Table Scalability	11
3.6.1. MAC Table Reduction on Data Center Switches	11
3.6.2. PE Router FIB Reduction	11
3.6.3. PE Router RIB Reduction	13
3.7. ARP/ND Cache Table Scalability on Default Gateways	14
3.8. ARP/ND and Unknown Uncast Flood Avoidance	14
3.9. Active-active Multi-homing	15
3.10. Path Optimization	15
4. Security Considerations	15
5. IANA Considerations	16
6. Acknowledgements	16
7. References	16
7.1. Normative References	16
7.2. Informative References	16
Authors' Addresses	17

1. Introduction

For business continuity purposes, Virtual Machine (VM) migration across data centers is commonly used in those situations such as data center maintenance, data center migration, data center consolidation, data center expansion, and data center disaster avoidance. It's generally admitted that IP renumbering of servers (i.e., VMs) after the migration is usually complex and costly at the risk of extending the business downtime during the process of migration. To allow the migration of a VM from one data center to another without IP renumbering, the subnet on which the VM resides needs to be extended across these data centers.

In Infrastructure-as-a-Service (IaaS) cloud data center environments, to achieve subnet extension across multiple data centers in a scalable way, the following requirements SHOULD be considered for any data center interconnect solution:

1) VPN Instance Space Scalability

In a modern cloud data center environment, thousands or even tens of thousands of tenants could be hosted over a shared network infrastructure. For security and performance isolation purposes, these tenants need to be isolated from one another. Hence, the data center interconnect solution SHOULD be capable of providing a large enough Virtual Private Network (VPN) instance space for tenant isolation.

2) Forwarding Table Scalability

With the development of server virtualization technologies, a single cloud data center containing millions of VMs is not uncommon. This number already implies a big challenge for data center switches, especially for core/aggregation switches, from the perspective of forwarding table scalability. Provided that multiple data centers of such scale were interconnected at layer2, this challenge would be even worse. Hence an ideal data center interconnect solution SHOULD prevent the forwarding table size of data center switches from growing by folds as the number of data centers to be interconnected increases. Furthermore, if any kind of L2VPN or L3VPN technologies is used for interconnecting data centers, the scale of forwarding tables on PE routers SHOULD be taken into consideration as well.

3) ARP/ND Cache Table Scalability on Default Gateways

[NARTEN-ARMD] notes that the ARP/ND cache tables maintained by data center default gateways in cloud data centers can raise both scalability and security issues. Therefore, an ideal data center interconnect solution SHOULD prevent the ARP/Neighbor cache table size from growing by multiples as the number of data centers to be connected increases.

4) ARP/ND and Unknown Unicast Flood Suppression or Avoidance

It's well-known that the flooding of Address Resolution Protocol (ARP)/Neighbor Discovery (ND) broadcast/multicast and unknown unicast traffic within a large Layer2 network are likely to affect performances of networks and hosts. As multiple data centers each containing millions of VMs are interconnected together across the Wide Area Network (WAN) at layer2, the impact of flooding as mentioned above will become even worse. As such, it becomes increasingly desirable for data center operators to suppress or even avoid the flooding of ARP/ND broadcast/multicast and unknown unicast traffic across data centers.

5) Active-active Multi-homing

In order to utilize the bandwidth of all available paths between the data center and the transport network in addition to providing resilient connectivity between them, active-active multi-homing is increasingly advocated by data center operators as a replacement of the traditional active-standby multi-homing approach.

6) Path Optimization

A subnet usually indicates a location in the network. However, when a subnet has been extended across multiple geographically dispersed data center locations, the location semantics of such subnet is not retained any longer. As a result, the traffic from a cloud user (i.e., a VPN user) which is destined for a given server located at one data center location of such extended subnet may arrive at another data center location firstly according to the subnet route, and then be forwarded to the location where the service is actually located. This suboptimal routing would obviously result in the unnecessary consumption of the bandwidth resources which are intended for data center interconnection. Furthermore, in the case where the traditional VPLS technology [RFC4761, RFC4762] is used for data center interconnect and default gateways of different data center locations are configured within the same virtual router redundancy group, the returning traffic from that server to the

cloud user may be forwarded at layer2 to a default gateway located at one of the remote data center premises, rather than the one placed at the local data center location. This suboptimal routing would also unnecessarily consume the bandwidth resources which are intended for data center interconnect.

This document describes a L3VPN-based subnet extension solution referred to as Virtual Subnet (VS), which can meet all of the requirements of cloud data center interconnect as described above. Since VS mainly reuses existing technologies including BGP/MPLS IP VPN [RFC4364] and ARP/ND proxy [RFC925][RFC1027][RFC4389], it allows service providers who are offering IaaS cloud services to the public to interconnect their geographically dispersed data centers in a much more scalable way, and more importantly, data center interconnection design can rely upon their existing MPLS/BGP IP VPN infrastructures therefore taking benefit from years of experience in the delivery and the operation of MPLS/BGP IP VPN services.

Please note that VS is targeted at scenarios where the traffic across data centers is routable IP traffic. In such scenario, data center operators who are implementing data center interconnect could benefit from the advantages that such host route-based subnet extension solution uniquely provides, such as MAC table reduction on data center switches, ARP/ND cache table reduction on data center default gateways, path optimization for inter-subnet traffic, and so on.

2. Terminology

This memo makes use of the terms defined in [RFC4364], [RFC2338] [MVPN] and [VA-AUTO].

3. Solution Description

3.1. Unicast

3.1.1. Intra-subnet Unicast

As shown in Figure 1, two CE hosts (i.e., Hosts A and B) which are configured within the same subnet (i.e., 1.1.1.0/24) are located in two different data centers (i.e., DC West and DC East) respectively. PE routers (i.e., PE-1 and PE-2) which are used for interconnecting the above two data centers create host routes for their local CE hosts respectively and then redistribute these routes into BGP. Meanwhile, ARP proxy is enabled on the VRF attachment circuits of these PE routers.

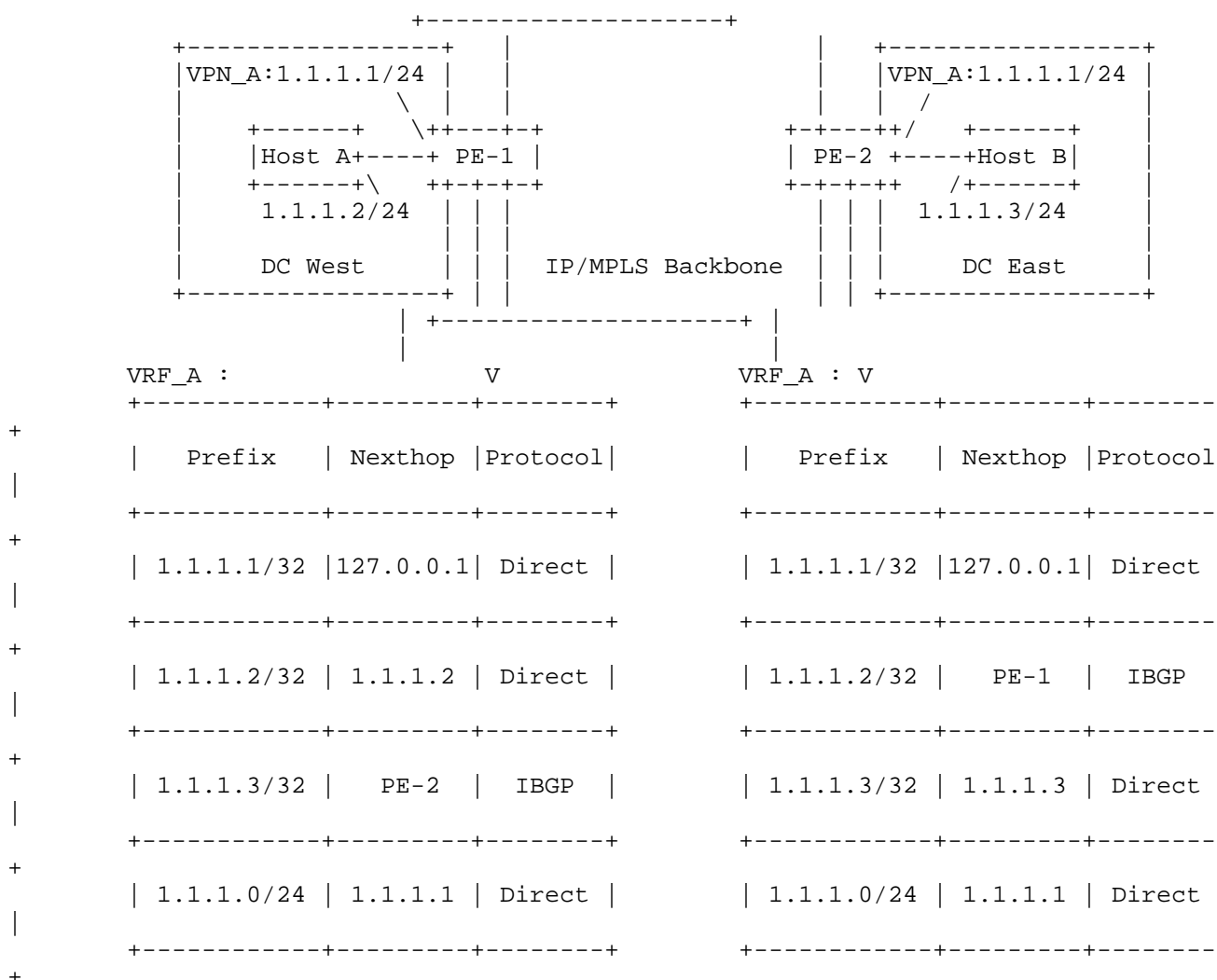


Figure 1: Intra-subnet Unicast Example

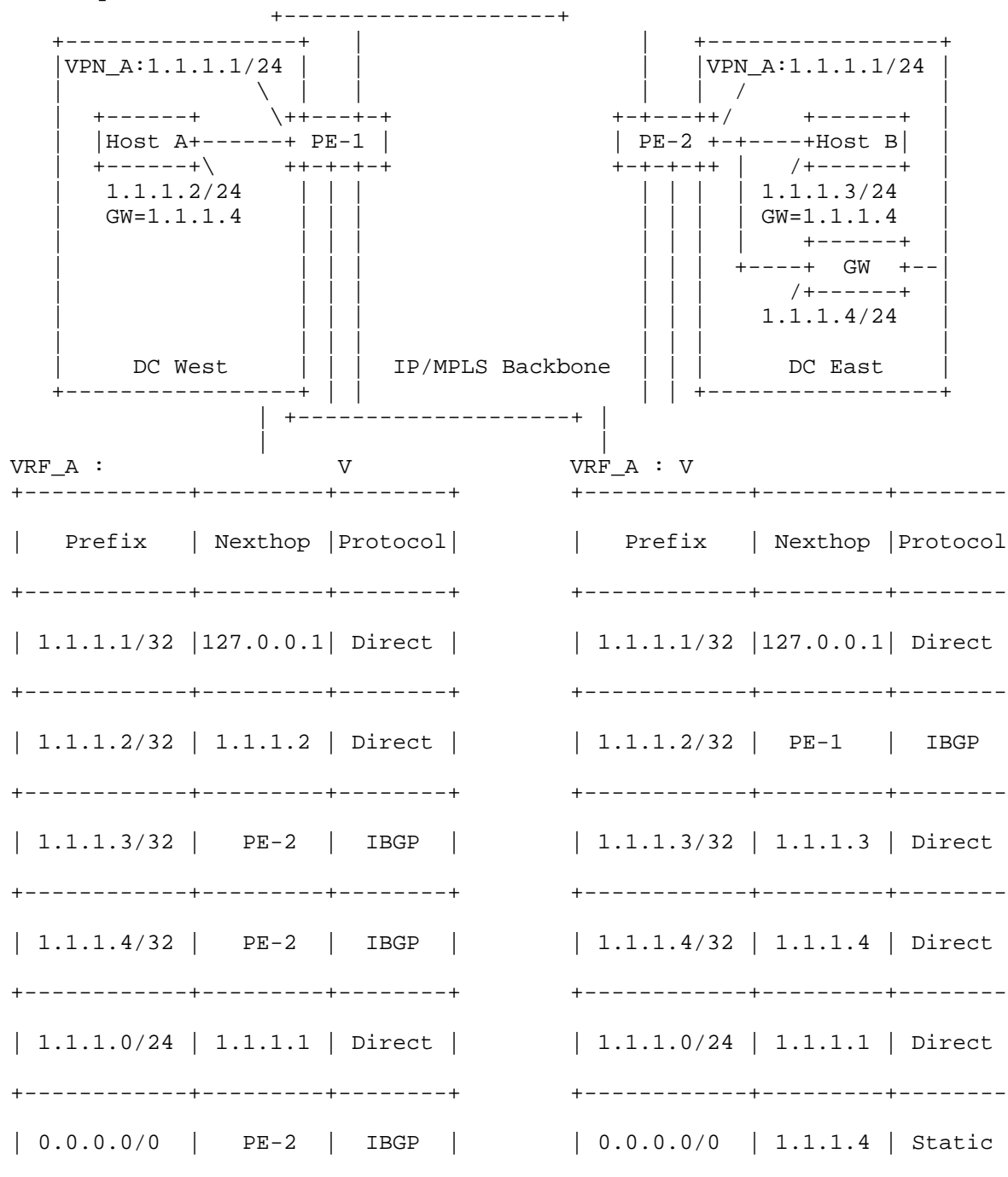
Now assume host A sends an ARP request for host B before communicating with host B. Upon receiving the ARP request, the ARP proxy embedded in PE-1 returns its own MAC address as a response. Host A then sends IP packets for host B to PE-1. Strictly according to the normal L3VPN forwarding procedure, PE-1 tunnels such packets towards PE-2 which in turn forwards them to host B. Thus, hosts A and B can communicate with each other as if they were located within the same subnet or Local Area Network (LAN). In fact, such subnet is a virtual subnet which is emulated by using host routes, rather than a real subnet.

3.1.2. Inter-subnet Unicast

As shown in Figure 2, only one data center (i.e., DC East) is deployed with a default gateway (i.e., GW). PE-2 which is connected to GW would either be configured with or learn from GW a default route with its next-hop being pointed to GW, and this route is distributed to other PE routers (i.e., PE-1) as per normal [RFC4364] operation. Assume host A sends an ARP request for its default gateway (i.e., 1.1.1.4) prior to communicating with a destination

host outside of its subnet (i.e., outside of 1.1.1.0/24). Upon receiving this ARP request, the ARP proxy embedded in PE-1 returns

its own MAC address as a response. Host A then sends a packet towards Host B to PE-1. PE-1 forwards such packet towards PE-2 according to the default route learnt from PE-2, which in turn forwards that packet to GW according to the default route as well. In contrast, if host B sends an ARP request for its default gateway (i.e., 1.1.1.4) prior to communicating with a destination host outside of its subnet, it will receive an ARP response from GW. As such, the packet destined for the destination host will be forwarded directly to GW. Note that since the outgoing interface of the best-match route for the target host (i.e., 1.1.1.4) is the same as the one over which the ARP packet arrived, PE-2 would not respond to this ARP request.



+-----+-----+-----+ +-----+-----+-----+

Figure 2: Inter-subnet Unicast Example (1)

As shown in Figure 3, in this case where each data center is deployed with a default gateway, CE hosts will get ARP responses from their local default gateways, rather than from their local PE routers when sending ARP requests for their default gateways.

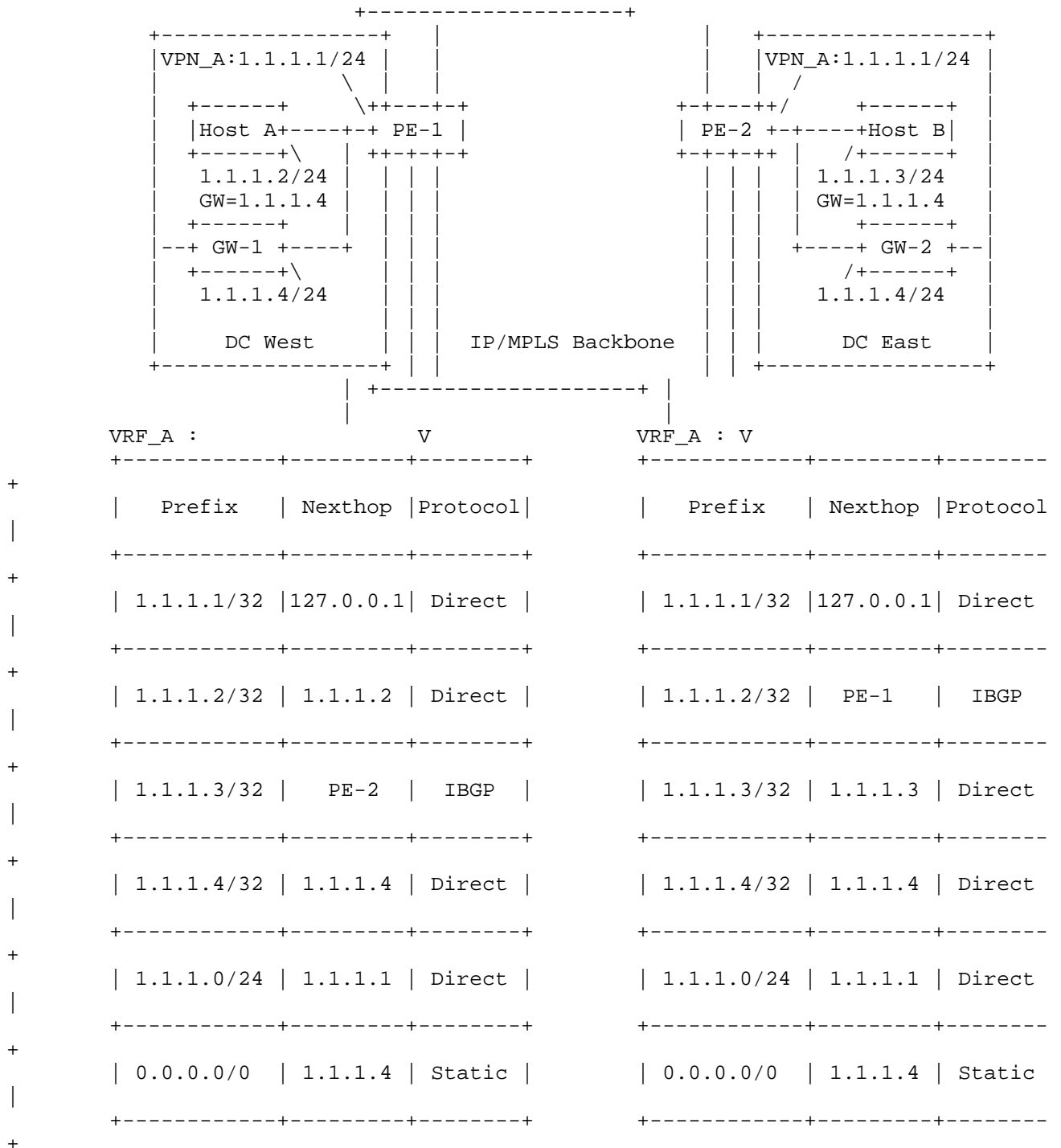
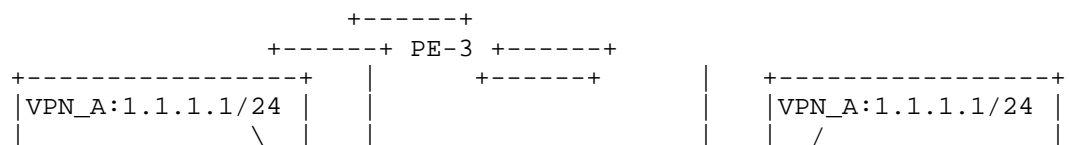
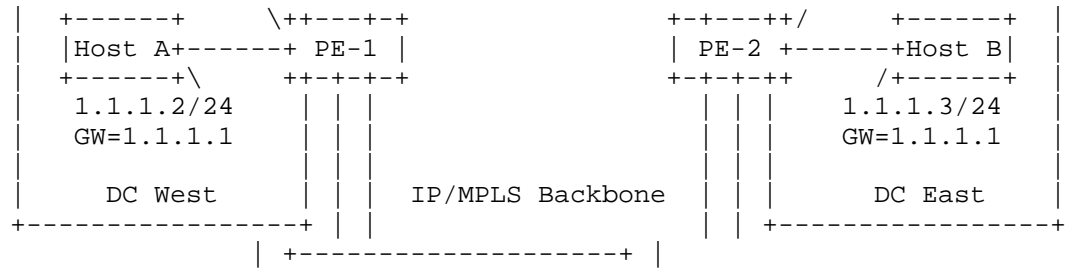


Figure 3: Inter-subnet Unicast Example (2)

Alternatively, as shown in Figure 4, PE routers themselves could be directly configured as the default gateways of their locally connected CE hosts as long as these PE routers have routes for the outside networks.





Internet-Draft	Virtual Subnet	October 2012
VRF_A :	V	VRF_A : V
+-----+-----+-----+		+-----+-----+-----+
Prefix Nexthop Protocol		Prefix Nexthop Protocol
+-----+-----+-----+		+-----+-----+-----+
1.1.1.1/32 127.0.0.1 Direct		1.1.1.1/32 127.0.0.1 Direct
+-----+-----+-----+		+-----+-----+-----+
1.1.1.2/32 1.1.1.2 Direct		1.1.1.2/32 PE-1 IBGP
+-----+-----+-----+		+-----+-----+-----+
1.1.1.3/32 PE-2 IBGP		1.1.1.3/32 1.1.1.3 Direct
+-----+-----+-----+		+-----+-----+-----+
1.1.1.0/24 1.1.1.1 Direct		1.1.1.0/24 1.1.1.1 Direct
+-----+-----+-----+		+-----+-----+-----+
0.0.0.0/0 PE-3 IBGP		0.0.0.0/0 PE-3 IBGP
+-----+-----+-----+		+-----+-----+-----+

Figure 4: Inter-subnet Unicast Example (3)

3.2. Multicast

To support IP multicast between CE hosts of the same virtual subnet, the MVPN technology [MVPN] could be directly reused. For example, PE routers attached to a given VPN join a default provider multicast distribution tree which is dedicated for that VPN. Ingress PE routers, upon receiving multicast packets from their local CE hosts, forward them towards remote PE routers through the corresponding default provider multicast distribution tree.

More details about how to support multicast and broadcast in VS will be explored in a later version of this document.

3.3. CE Host Discovery

PE routers SHOULD be able to discover their local CE hosts and keep the list of these hosts up to date in a timely manner so as to ensure the availability and accuracy of the corresponding host routes originated from them. PE routers could accomplish local CE host discovery by some traditional host discovery mechanisms using ARP or ND protocols. Furthermore, Link Layer Discovery Protocol (LLDP) described in [802.1AB] or VSI Discovery and Configuration Protocol (VDP) described in [802.1Qbg], or even interaction with the data center orchestration system could also be considered as a means to dynamically discover local CE hosts.

More details about the local CE host discovery approach will be explored in a later version of this document or a separate draft.

3.4. ARP/ND Proxy

Acting as ARP or ND proxies, PE routers SHOULD only respond to an ARP request or Neighbor Solicitation (NS) message for the target

Xu, et al.

Expires April 15, 2013

[Page 9]

host for which there is a host route in the associated VRF and the outgoing interface of that route is different from the one over which the ARP request or the NS message arrived. Otherwise, PE routers would not respond.

In the case where it's hard to guarantee each PE router has learnt all of its own local CE hosts entirely, upon receipt of an ARP request or a NS message for an unknown target host for which there is no corresponding host route in the associated VRF yet, ingress PE routers could propagate a BGP UPDATE message containing the IP address of the target host or even that of the requesting host so as to trigger remote PE routers receiving that message to send an ARP request or a NS message for the target host on their own attachment circuits on behalf of the requesting host. As such, the target host which has been silently attached to a given PE router (e.g., there is no any kind of host attachment notification received by the PE router.) could be discovered accordingly. The details of this special BGP update message will be disclosed in a separate draft.

In scenarios where a given VPN site (i.e., a data center) is multi-homed to more than one PE router via an Ethernet switch or an Ethernet network, VRRP [RFC5798] SHOULD be enabled on these PE routers for the sake of the availability of the network connectivity. In this case, only the PE router which is acting as the VRRP Master SHOULD perform the ARP/ND proxy function and respond with the virtual MAC address, instead of its physical MAC address.

3.5. CE Host Mobility

After moving from one VPN site to another, a CE host (e.g., a VM) will send a gratuitous ARP/ND message. Upon receiving that message, the PE router connected to the site where the VM moves to will create a host route for that CE host and then advertise it to remote PE routers.

Upon learning such route, the PE router that previously connected the CE host would immediately check whether that CE host is still connected to it by some means (e.g., ARP/ND PING and/or ICMP PING).

If not, the PE router would accordingly withdraw the corresponding host route which has been advertised before. Meanwhile, the PE router would broadcast a gratuitous ARP/ND message on behalf of that CE host. As such, the ARP/ND entry of that CE host which was cached on any local CE host would be updated accordingly.

3.6. Forwarding Table Scalability

3.6.1. MAC Table Reduction on Data Center Switches

In a VS environment, the MAC learning domain associated with a given virtual subnet which has been extended across multiple data centers is partitioned into segments and each of the segments is confined within a single data center. Therefore data center switches only need to learn local MAC addresses, rather than learning both local and remote MAC addresses as required in the case where the traditional VPLS technology [RFC4761, RFC4762] is used for data center interconnect.

3.6.2. PE Router FIB Reduction

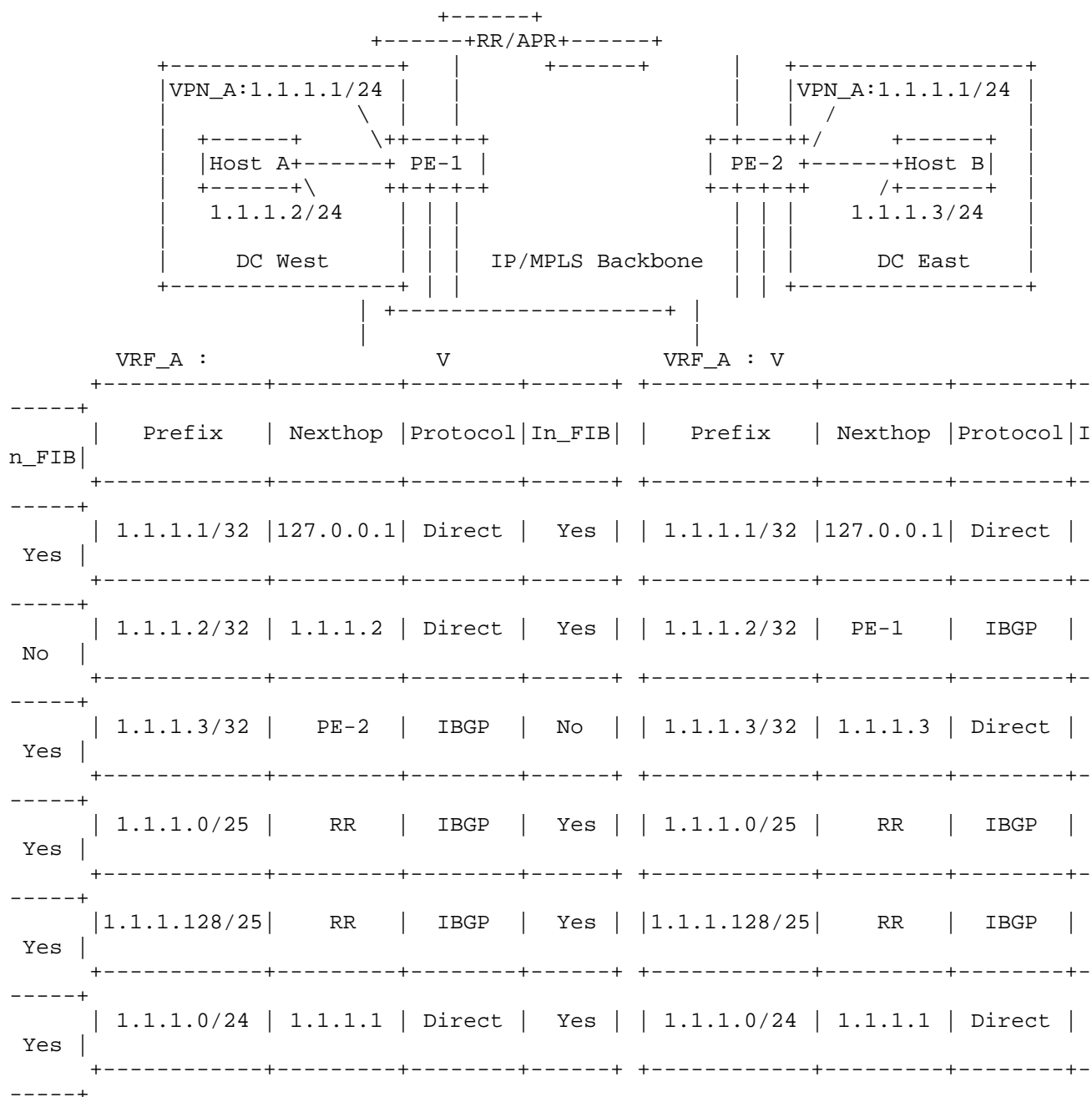


Figure 5: FIB Reduction Example

To reduce the FIB size of PE routers, Virtual Aggregation (VA) [VA-AUTO] technology can be used. Take the VPN instance A shown in Figure 5 as an example, the procedures of FIB reduction are as follows:

- 1) Multiple more specific prefixes (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to the prefix of virtual subnet (i.e., 1.1.1.0/24) are configured as Virtual Prefixes (VPs) and a Route-Reflector (RR) is configured as an Aggregation Point Router (APR) for these VPs. PE routers as RR clients advertise host routes for their own local CE hosts to the RR which in turn, as an APR, installs those host routes into its FIB and then attach the "can-suppress" tag to those host routes before reflecting them to its clients.
- 2) Those host routes which have been attached with the "can suppress" tag would not be installed into FIBs by clients who are VA-aware since they are not APRs for those host routes. In addition, the RR as an APR would advertise the corresponding VP routes to all of its clients, and those of which who are VA-aware in turn would install these VP routes into their FIBs.
- 3) Upon receiving a packet from a local CE host, if no matching host route found, the ingress PE router will forward the packet to the RR according to one of the VP routes learnt from the RR, which in turn forwards the packet to the relevant egress PE router according to the host route learnt from that egress PE router. In a word, the FIB table size of PE routers can be greatly reduced at the cost of path stretch. Note that in the case where the RR is not available for transferring L3VPN traffic between PE routers for some reason (e.g., the RR is implemented on a server, rather than a router), the APR function could actually be performed by a given PE router other than the RR as long as that PE router has installed all host routes belonging to the virtual subnet into its FIB. Thus, the RR only needs to attach a "can-suppress" tag to the host routes learnt from its clients before reflecting them to the other clients. Furthermore, PE routers themselves could directly attach the "can-suppress" tag to those host routes for their local CE hosts before distributing them to remote peers as well.
- 4) Provided a given local CE host sends an ARP request for a remote CE host, the PE router that receives such request will install the host route for that remote CE host into its FIB, in case there is a host route for that CE host in its RIB and has not yet been installed into the FIB. Therefore, the subsequent packets destined for that remote CE host will be forwarded directly to the egress PE router. To save the FIB space, FIB entries corresponding to remote host routes which have been attached with "can-suppress" tags would expire if they have not been used for forwarding packets for a certain period of time.

3.6.3. PE Router RIB Reduction

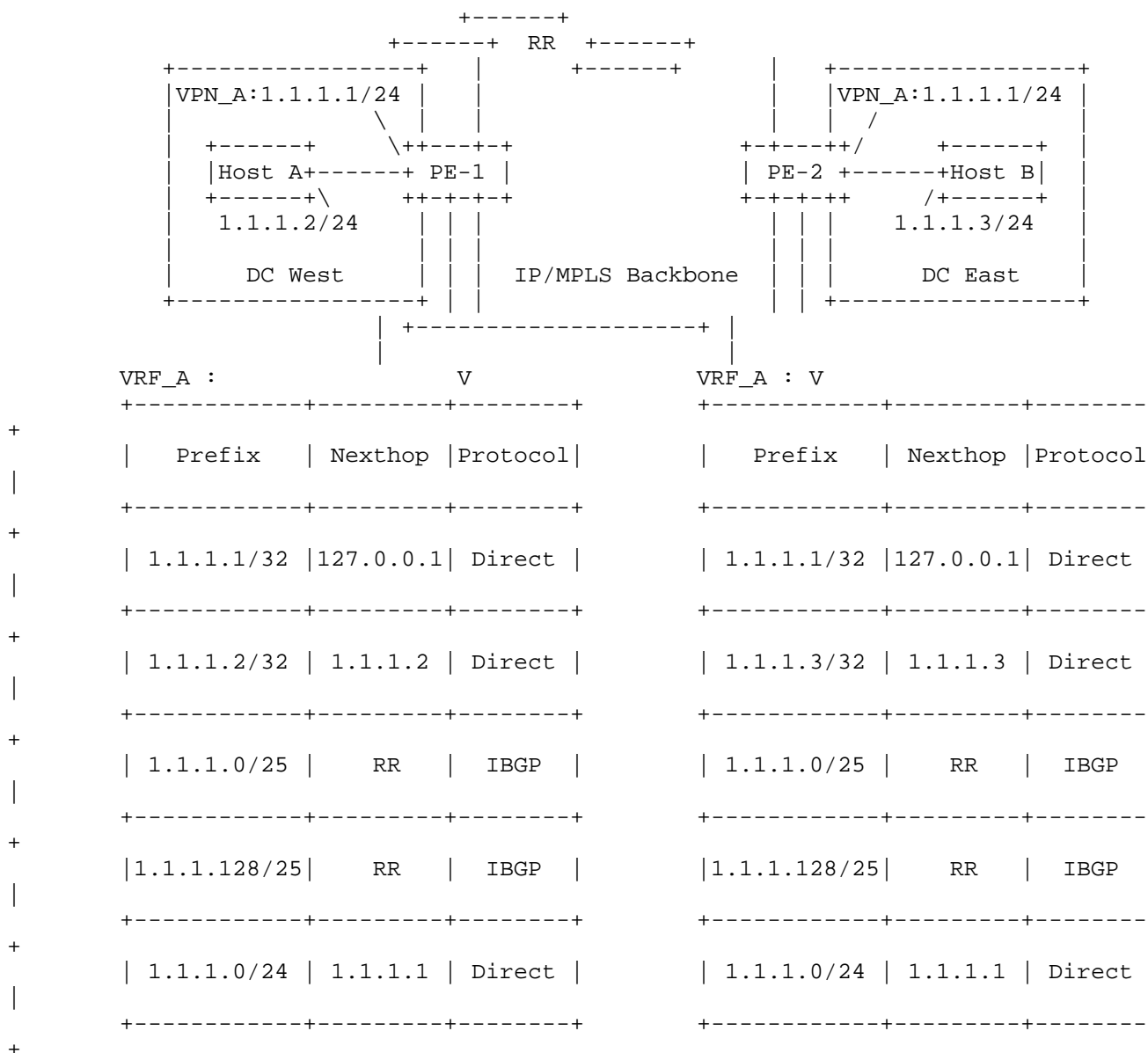


Figure 6: RIB Reduction Example

To reduce the RIB size of PE routers, BGP Outbound Route Filtering (ORF) mechanism is used to realize on-demand route announcement. Take the VPN instance A shown in Figure 6 as an example, the procedures of RIB reduction are as follows:

- 1) PE routers as RR clients advertise host routes for their local CE hosts to a RR which however doesn't reflect these host routes by default unless it receives explicit ORF requests for them from its clients. The RR is configured with routes for more specific subnets (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to the virtual subnet (i.e., 1.1.1.0/24) with next-hop being pointed to Null0 and then advertises these routes to its clients via BGP.
- 2) Upon receiving a packet from a local CE host, if no matching host route found, the ingress PE router will forward the packet to the

RR according to one of the subnet routes learnt from the RR, which in turn forwards the packet to the relevant egress PE router according to the host route learnt from that egress PE router. In a word, the RIB table size of PE routers can be greatly reduced at the cost of path stretch.

- 3) Just as the approach mentioned in section 3.6.2, in the case where the RR is not available for transferring L3VPN traffic between PE routers for some reason, a PE router other than the RR could advertise the more specific subnet routes as long as that PE router has installed all host routes belonging to that virtual subnet into its FIB.
- 4) Provided a given local CE host sends an ARP request for a remote CE host, the ingress PE router that receives such request will request the corresponding host route from its RR by using the ORF mechanism (e.g., a group ORF containing Route-Target (RT) and prefix information) in case there is no host route for that CE host in its RIB yet. Once the host route for the remote CE host is learnt from the RR, the subsequent packets destined for that CE host would be forwarded directly to the egress PE router. Note that the RIB entries of remote host routes could expire if they have not been used for forwarding packets for a certain period of time. Once the expiration time for a given RIB entry is approaching, the PE router would notify its RR not to pass the updates for corresponding host route by using the ORF mechanism.

3.7. ARP/ND Cache Table Scalability on Default Gateways

In case where data center default gateway functions are implemented on PE routers of the VS as shown in Figure 4, since the ARP/ND cache table on each PE router only needs to contain ARP/ND entries of local CE hosts, the ARP/ND cache table size will not grow as the number of data centers to be connected increases.

Alternatively, if dedicated default gateways are directly connected to PE routers of the VS as shown in Figure 3, all remote CE hosts of a given virtual subnet share the same MAC address (i.e., the MAC address of the local PE router) from the point of view of default gateways, because of the use of the ARP/ND proxy function embedded in PE routers. Therefore, ARP/ND entries of those remote CE hosts could be aggregated into one ARP/ND entry (i.e., 1.1.1.0/24-> the MAC address of the PE router in the IPv4 case). Accordingly, default gateways are required to use the longest-matching algorithm for ARP/ND cache lookup instead of the existing exact-matching algorithm. Thus, the ARP/ND cache table size of DC gateways can be reduced greatly as well.

3.8. ARP/ND and Unknown Uncast Flood Avoidance

In VS, the flooding domain associated with a given virtual subnet that has been extended across multiple data centers, has been partitioned into segments and each of the segments is confined

within a single data center. Therefore, the performance impact on networks and servers caused by the flooding of ARP/ND broadcast/multicast and unknown unicast traffic is alleviated.

3.9. Active-active Multi-homing

For PE router redundancy purposes, a VPN site could be connected to more than one PE router. In this case, VRRP SHOULD be enabled on these PE routers and only the PE router which is acting as the VRRP Master SHOULD perform the ARP proxy functionality. However, all PE routers, either as a VRRP master or a VRRP slave, are allowed to advertise host routes for their local CE hosts. Hence, from the perspective of remote PE routers, there will be multiple host routes for a given CE host located within that multi-homed site. In other words, active-active multi-homing is available for the inbound traffic of a given multi-homed site.

3.10. Path Optimization

Take the scenario shown in Figure 4 as an example, to optimize the forwarding path for traffic between cloud users and cloud data centers, PE routers located at cloud data centers (i.e., PE-1 and PE-2), which are also the data center default gateways, propagate host routes for their local CE hosts respectively to remote PE routers which are attached to cloud user sites (i.e., PE-3).

As such, the traffic from cloud user sites to a given server on the virtual subnet which has been extended across data centers would be forwarded directly to the data center location where that server resides, since traffic is now forwarded according to the host route for that server, rather than the subnet route.

Furthermore, for traffic coming from the cloud data center and forwarded to cloud user sites, each PE router acting as a default gateway would forward traffic received from its local CE hosts directly to the remote PE routers (i.e., PE-3) according to the best-match route in the corresponding VRF. As a result, traffic from data centers to enterprise sites is forwarded along the optimal path without consuming the bandwidth resources intended for data center interconnect.

4. Security Considerations

TBD.

5. IANA Considerations

There is no requirement for IANA.

6. Acknowledgements

Thanks to Dino Farinacci, Himanshu Shah, Nabil Bitar, Giles Heron, Ronald Bonica, Monique Morrow for their valuable comments and suggestions on this document.

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

7.2. Informative References

[RFC4364] Rosen. E and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[MVPN] Rosen. E and Aggarwal. R, "Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10.txt, Work in Progress, January 2010.

[VA-AUTO] Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, "Auto-Configuration in Virtual Aggregation", draft-ietf-grow-va-auto-05.txt, Work in Progress, December 2011.

[RFC925] Postel, J., "Multi-LAN Address Resolution", RFC-925, USC Information Sciences Institute, October 1984.

[RFC1027] Smoot Carl-Mitchell, John S. Quarterman, "Using ARP to Implement Transparent Subnet Gateways", RFC 1027, October 1987.

[RFC4389] D. Thaler, M. Talwar, and C. Patel, "Neighbor Discovery Proxies (ND Proxy) ", RFC 4389, April 2006.

[RFC5798] S. Nadas., "Virtual Router Redundancy Protocol", RFC 5798, March 2010.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [802.1AB] IEEE Standard 802.1AB-2009, "Station and Media Access Control Connectivity Discovery", September 17, 2009.
- [802.1Qbg] IEEE Draft Standard P802.1Qbg/D2.0, "Virtual Bridged Local Area Networks -Amendment XX: Edge Virtual Bridging", Work in Progress, December 1, 2011.
- [NARTEN-ARMD] Narten, T., Karir, M., and I. Foo, "Problem Statement for ARMD", draft-ietf-armd-problem-statement-01.txt, Work in Progress, February 2012.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
Beijing, China.
Phone: +86 10 60610041
Email: xuxiaohu@huawei.com

Susan Hares
Huawei Technologies (FutureWei group)
2330 Central Expressway
Santa Clara, CA 95050
Phone: +1-734-604-0332
Email: Susan.Hares@huawei.com
shares@ndzh.com

Yongbing Fan
Guangzhou Institute, China Telecom
Guangzhou, China.
Phone: +86 20 38639121
Email: fanyb@gsta.com

Christian Jacquenet
France Telecom
Rennes
France
Email: christian.jacquenet@orange.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 02, 2013

R. Bush
Internet Initiative Japan
K. Patel
P. Mehta
A. Sreekantiah
Cisco Systems
L. Jalil
Verizon
October 2012

Authenticating L3VPN Origination Signaling
draft-ymbk-l3vpn-origination-02

Abstract

A BGP-signaled Layer-3 VPN's prefix bindings sent over BGP are subject to unintentional errors, both by the legitimate originator and by non-legitimate origins. This is of special concern if the VPN traverses untrusted networks. This document describes how the sender of the Prefix/VPN binding may sign it so that recipient of the binding may authenticate it.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in RFC 2119 [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning.

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 <http://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 02, 2013.

Copyright Notice

Copyright (c) 2012 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.

This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

Table of Contents

1. Introduction	2
2. NLRI Deaggregation	3
3. L3VPN Origination BGP Path Attribute (L3OPA)	3
4. Validation of Routes Having an L3OPA	4
5. L3VPN Deployment Scenarios	5
5.1. End CE to CE Authentication	5
5.2. Provider/ASBR Based Validation/Authentication	5
5.3. PE-PE Based Validation	6
6. Notes	6
7. Security Considerations	7
8. IANA Considerations	7
9. Acknowledgements	7
10. References	7
10.1. Normative References	7
10.2. Informative References	8
Authors' Addresses	8

1. Introduction

RFC 4364 [RFC4364] Section 7.4 describes how a Customer Edge (CE) router uses eBGP to announce to a Provider Edge (PE) router the address prefix(es) the customer provides to an L3VPN. It is possible that the originator of such an announcement could unintentionally announce prefixes they do not own.

```
Cust(West)-CE--PE-Provider(West)--TransitA-~
~-TransitB--Provider(East)-PE--CE-Cust(East)
```

This document describes how the PE receiving the CE's originating announcement, West, may sign the announcement so that the PE proximal to the destination CE, East, may authenticate the NLRI see RFC 4364 [RFC4364] Section 4.3.1. Alternatively, the originating CE router may sign the announcement so that the destination CE router may authenticate the NLRI.

It is assumed that the providers already have the key creation, storage, and distribution infrastructure needed. Keys might be configured on the routers, or in some shared PKI, or, for example, the Resource Public Key Infrastructure (RPKI) could be used, see RFC 6480 [RFC6480].

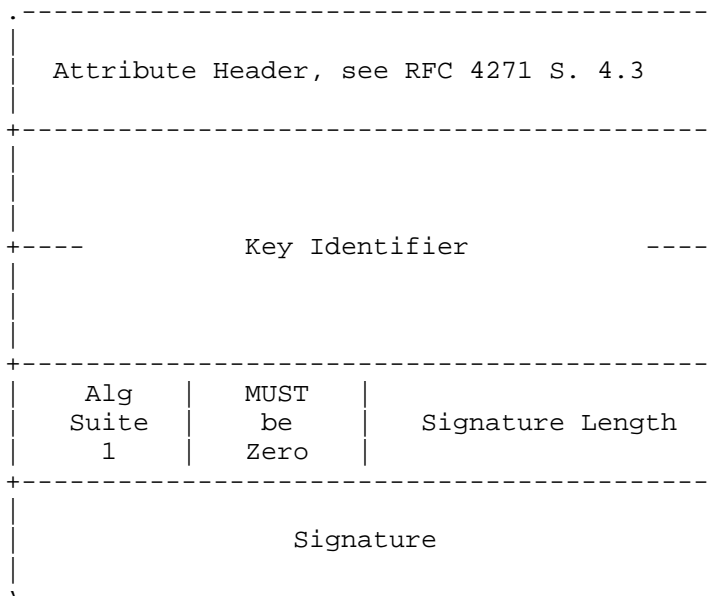
A new BGP PATH Attribute, called L3VPN Origination BGP PATH Attribute (L3OPA), is created to contain the necessary keying information and signature.

2. NLRI Deaggregation

Normally, a BGP Update may contain multiple NLRI which all share the identical set of attributes. As L3OPA signalling signs over the NLRI, and NLRI can become separated as they transit the network, separation would break the signature. Therefore, a BGP announcement using L3OPA signalling MUST contain one and only one NLRI.

3. L3VPN Origination BGP Path Attribute (L3OPA)

The L3OPA is a BGP optional transitive Path Attribute RFC 4271 [RFC4271]. BGP Path Attributes are Type/Length/Value tuples.



The Attribute Type is two octets, the first of which, Attribute Flags, MUST have the two high order bits set to signify that attribute is optional and transitive.

The second octet of the Attribute Flags, Attribute Type, MUST be set to 0xXX, as assigned by the IANA, see Section 8, to signal that this is an L3OPA.

The Length field is one or two octets with a value of the number of octets in the entire attribute. If the length of the L3OPA is less than 256 octets, only the first octet of the length field is used. Otherwise, both octets are used to represent the Length.. See RFC 4271 [RFC4271] Section 4.2 for another explanation of this byte saving.

The Key Identifier is an eight octet value identifying the key (pair) used for the Signature. It is used when the keying is not implied by the NLRI, as it would be, for example, if the RPKI was used. It is often the VPN Identifier. If not used to identify the key, it MUST be zero.

The Algorithm Suite is a one-octet identifier specifying the digest algorithm and digital signature algorithm used to produce the Signature. The values reference the IANA registry for Algorithm Identifiers from BGPsec, see [I-D.ietf-sidr-bgpsec-algs].

The Signature Length is two octets and is the number of octets in the Signature field.

The Signature field is a digital signature that covers the NLRI and the Key Identifier.

To compute the Signature, the digest algorithm for the specified Algorithm Suite is applied to the catenation of the NLRI and the Key Identifier. This is then fed to the signature algorithm for the specified algorithm suite and the resulting value is the Signature.

Signature = sign (hash (NLRI || Key Identifier))

4. Validation of Routes Having an L3OPA

A BGP speaker receiving routes with an L3OPA MUST perform the necessary validation if configured to do so.

The digest algorithm for the specified Algorithm Suite is applied to the catenation of the NLRI and the Key Identifier. This is then fed to the signature algorithm for the specified algorithm suite and the resulting value is compared with the Signature.

If the signature value matches the Signature in the attribute, the route MUST be marked as Valid, otherwise it MUST be marked as Invalid.

A route received without an L3OPA SHOULD be marked as having an Unknown validity state.

If L3OPA marking is disabled in the router configuration, routes are

considered to have the Unknown validity state.

Configured local policy on the router may use the validity state markings to implement policy. For example, a route marked as Invalid or Unknown may be dropped or de-preferenced by appropriate use of normal BGP policy mechanisms.

Note that this is similar to announcement marking while allowing the user to control policy as described in RPKI-Based BGP origin validation, see [I-D.ietf-sidr-pfx-validate].

5. L3VPN Deployment Scenarios

The following L3VPN deployment scenarios illustrate use of the scheme. The examples use the language of symmetric keys which have been previously agreed upon between the signer of the route and the validator. Asymmetric keying, a PKI, etc. could also be used. Signing and validation are as described above.

5.1. End CE to CE Authentication

```
CE1 ---- PE1 ----- PE2 -- CE2
                AS1
```

```
CE1 ---- PE1 ----- ASBR1 ----- ASBR2 ----- PE2 ---- CE2
                AS1                      AS2
```

CE1 and CE2 are end CEs in the same VPN. PE1, PE2, ASBR1, ASBR2 are provider PE/ASBRs which are blindly propagating the announcement with the L3OPA as generated by CE1.

As the authorization is between the originating CE1 and the terminating CE2, the keying should not be known by the provider(s). The CEs are configured with the keying information, the originating CE1 creates and signs an L3OPA for each NLRI participating in the VPN.

An update received by CE2 without an L3OPA, or having an Invalid Signature would likely be dropped. Thus the CEs are protected from incorrect prefixes originating from a provider network or unauthorized CEs.

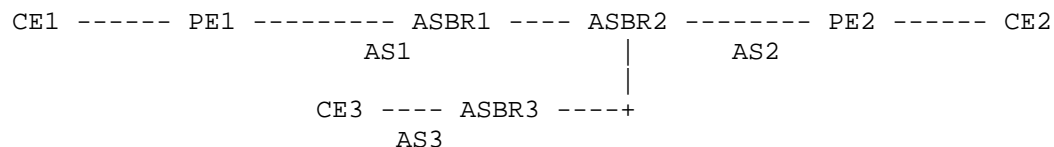
5.2. Provider/ASBR Based Validation/Authentication

```
CE1 ---- PE1 ----- ASBR1 ----- ASBR2 ----- PE2 ---- CE2
                AS1                      AS2
```

In the diagram, CE1 is the originating/signing CE. ASBR2 is the trusted provider with whom CE1 has collaborated. Updates generated by CE1 may be passed transparently through any number of intermediate providers, ASBR1s, which blindly propagate the L3OPA. Validation is performed when the announcement reaches the trusted validating provider, ASBR2.

Keying is agreed between CE1 and the trusted provider ASBR2, likely on per-VPN basis.

5.3. PE-PE Based Validation



Here PEs, possibly across ASes, agree on the keying. The Key Identifier and associated keys would normally be configured on a per VPN basis, with the PE1 signing and PE2 and PE3 validating similarly to the CEs in the previous examples.

CE1 originates an announcement, possibly with multiple NLRI, but without an L3OPA. PE1 de-aggregates the NLRI into separate announcements, signs each with the keying agreed with PE2 and PE3, and propagates them. Arbitrary providers carry the announcements toward PE2 and PE3, where the announcements have their Signatures validated, the L3OPAs removed, and are then propagated to CE2 and CE3.

6. Notes

The keying could either come from the Global RPKI or the customer or carrier running their own PKI. The keying is assumed to be asymmetric, but possibly could be symmetric. The keys can be statically configured (beware scaling and key-roll issues), dynamic, in some public or private infrastructure, etc.

If the RPKI is used, and the public key is taken from the CA certificate which owns the NLRI, the classic problem arises where all the NLRI on that certificate share fate. I.e. if one causes the need for a re-key, then all must re-key. RPKI-based origin validation solves this problem by a level of indirection, the CA certificate is used to sign an End Entity (EE) certificate which signs a Route Origin Authorization (ROA), see RFC 6480 [RFC6480] and RFC 6482 [RFC6482]. As the Key Identifier of an L3VPN signal is larger than the four octets of a ROA, a new RPKI object, for the moment let's call it a VOA, would have to be defined and then it would have to be carried in the RPKI-Router Protocol [I-D.ietf-sidr-rpki-rtr].

If the value of the signing key, as identified by the Key Identifier, is to be rolled, in case of compromise or security policy, the technique in RFC 4808 [RFC4808] should be used.

While it is poor security practice to trust a different entity for your security/authentication/..., should a non-validating router choose to trust a validating router, they could use normal policy and signaling mechanisms, e.g. communities, to signal validation status. This page is too small to enumerate the vulnerabilities this creates.

7. Security Considerations

Signing (NLRI || Key Identifier) with the key of the NLRI-owner or some other pre-agreed key, only says that the contents were produced by the owner of the key (NLRI or other), and that no one in between has changed the (NLRI || Key Identifier). This is not protection against attacks, only configuration errors, aka 'fat fingers'. If we were trying to protect against an attacking PE replaying a signed (NLRI || Key Identifier) it has no business announcing, this design does not help.

If Key Identifier based keying is used, then the Key Identifier, and hence the signing key, MUST be unique to the VPN.

Adding a VOA which binds (NLRI || Key Identifier) still could be replayed from anywhere so really offers nothing. Like RPKI-based origin validation, this only catches fat fingers, not black hats.

8. IANA Considerations

This document requests the IANA create a new entry in the BGP Path Attributes Registry as follows:

Value	Code	Reference
-----	-----	-----
TBD	L3VPN Origination	This Document

9. Acknowledgements

The authors would like to thank Eric Rosen, John Scudder, Russ Housley, and Sandy Murphy.

We note the long expired draft draft-ietf-l3vpn-auth by Ron Bonica, Yakov Rekhter, Eric Rosen, Robert Raszuk, and Dan Tappan.

10. References

10.1. Normative References

- [I-D.ietf-sidr-bgpsec-algs]
Turner, S., "BGP Algorithms, Key Formats, & Signature Formats", Internet-Draft draft-ietf-sidr-bgpsec-algs-03, September 2012.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

- [RFC4271] Rekhter, Y., Li, T. and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4808] Bellovin, S., "Key Change Strategies for TCP-MD5", RFC 4808, March 2007.
- [RFC6480] Lepinski, M. and S. Kent, "An Infrastructure to Support Secure Internet Routing", RFC 6480, February 2012.

10.2. Informative References

- [I-D.ietf-sidr-pfx-validate]
 Mohapatra, P., Scudder, J., Ward, D., Bush, R. and R. Austein, "BGP Prefix Origin Validation", Internet-Draft draft-ietf-sidr-pfx-validate-10, October 2012.
- [I-D.ietf-sidr-rpki-rtr]
 Bush, R. and R. Austein, "The RPKI/Router Protocol", Internet-Draft draft-ietf-sidr-rpki-rtr-26, February 2012.
- [RFC6482] Lepinski, M., Kent, S. and D. Kong, "A Profile for Route Origin Authorizations (ROAs)", RFC 6482, February 2012.

Authors' Addresses

Randy Bush
Internet Initiative Japan
5147 Crystal Springs
Bainbridge Island, Washington 98110
US

Email: randy@psg.com

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

Pranav Mehta
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: pmehta@cisco.com

Arjun Sreekantiah
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: asreekan@cisco.com

Luay Jalil
Verizon
1201 E Arapaho Rd.
Richardson, TX 75081
USA

Email: luay.jalil@verizon.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 18, 2013

L. Zheng
Z. Li
Huawei Technologies
October 15, 2012

Performance Monitoring Analysis for L3VPN
draft-zheng-l3vpn-pm-analysis-00

Abstract

To perform the measurement of packet loss, delay and other metrics on a particular VPN flow, the egress PE need to tell to which specific ingress VRF a packet belongs. But for L3VPN, multipoint-to-point or multipoint-to-multipoint (MP2MP) network model applies, flow identifying is a big challenge. This document summarizes the current performance monitoring mechanisms for MPLS networks, and analyzes the challenge for L3VPN performance monitoring. This document also discuss the key points need to be taken in consideration when designing L3VPN performance monitoring mechanisms.

Requirements Language

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 RFC 2119 [RFC2119].

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 <http://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 18, 2013.

Copyright Notice

Copyright (c) 2012 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	4
2. Overview of current mechanisms for MPLS networks	5
2.1. Packet Loss and Delay Measurement for MPLS Networks	5
2.2. Profile for MPLS-based Transport Networks	5
3. Challenge for L3VPN Performance Monitoring	6
4. Design Consideration	8
4.1. P2P Connection	8
4.2. Control Plane	8
4.3. Data Plane	8
4.4. MPLS OAM	8
4.5. QoS	9
4.6. Configuration	9
5. Security Considerations	10
6. IANA Considerations	11
7. Acknowledgements	12
8. References	13
8.1. Normative References	13
8.2. Informative References	13
Authors' Addresses	14

1. Introduction

Level 3 Virtual Private Network (L3VPN) [RFC4364] service is widely deployed in the production network. It is deployed to provide enterprise interconnection, Voice over IP (VoIP), video, mobile, etc. services. Most of these services are sensitive to the packet loss and delay. The capability to measure and monitor performance metrics for packet loss, delay, as well as related metrics is essential for SLA. The requirement for SLA measurement for MPLS networks has been documented in [RFC4377].

One popular deployment of L3VPN nowadays is in mobile backhaul networks. When deploying MPLS-TP in mobile backhaul network, due to the scalability issue with PW, L3VPN is used either for end-to-end service delivery, or L2VPN and L3VPN hybrid networking. The measurement capability of L3VPN provides operators with greater visibility into the performance characteristics of their networks, and provides diagnostic information in case of performance degradation or failure and helps for fault localization.

To perform the measurement of packet loss, delay and other metrics on a particular VPN flow, the egress PE need to tell to which specific ingress VRF a packet belongs. But for L3VPN, multipoint-to-point or multipoint-to-multipoint (MP2MP) network model applies, flow identifying is a big challenge. This document summarizes the current performance monitoring mechanisms for MPLS networks, and analyzes the challenge for L3VPN performance monitoring. This document also discuss the key points need to be taken in consideration when designing L3VPN performance monitoring mechanisms.

2. Overview of current mechanisms for MPLS networks

2.1. Packet Loss and Delay Measurement for MPLS Networks

[RFC6374] defines procedure and protocol mechanisms to enable the efficient and accurate measurement of packet loss, delay, as well as related metrics in MPLS networks.

The LM protocol can perform two distinct kinds of loss measurement. In inferred mode, it can measure the loss of specially generated test packets (in order to infer the approximate data-plane loss level). In direct mode, it can directly measure data-plane packet loss. Direct measurement provides perfect loss accounting, but may require specialized hardware support and is only applicable to some LSP types. Inferred measurement provides only approximate loss accounting but is generally applicable.

The LM and DM protocols are initiated from a single node, the querier. A query message may be received either by a single node or by multiple nodes; i.e. these protocols provide point-to-point or point-to-multipoint measurement capabilities.

2.2. Profile for MPLS-based Transport Networks

Procedures for the measurement of packet loss, delay, and throughput in MPLS networks are defined in [RFC6374]. [RFC6375] describes a profile, i.e. a simplified subset, of procedures that suffices to meet the specific requirements of MPLS-based transport networks [RFC5921] as defined in [RFC5860]. This profile is presented for the convenience of implementors who are concerned exclusively with the transport network context.

LM session is externally configured and the values of several protocol parameters can be fixed in advance at the endpoints involved in the session, so that inspection or negotiation of these parameters is not required.

3. Challenge for L3VPN Performance Monitoring

To perform the measurement of packet loss, delay and other metrics on a particular VPN flow, the egress PE need to tell to which specific ingress VRF a packet belongs.

The above mentioned existing mechanisms for MPLS networks provide either point-to-point or point-to-multipoint measurement capabilities. For a specific receiver, it could easily identify a specific flow by the label stack information, when Penultimate Hop Pop (PHP) function is disabled .

But in the case of L3VPN, multipoint-to-point or multipoint-to-multipoint (MP2MP) network model applies , it makes the flow identifying a big challenge for packets loss and delay measurement. According to the label allocation mechanisms of L3VPN, a private label itself cannot uniquely identify a specific VPN flow. That is, when the egress PE allocates VPN label for a specific prefix of a VPN, the same label will be advertised to all its peers. Given a VPN flow, the egress PE cannot tell which ingress VRF is from based on the private label it carries. As a result, it's not feasible to perform the loss or delay measurement on this flow.

In L3VPN the LSPs may be merged at any intermediate nodes along the LSP (e.g., Label Distribution Protocol (LDP) [RFC5036] based LSP). The egress PE cannot derive a unique identifier of the source PE from label stack. The tunnel label cannot help for flow identification due to the LSP merge.

In L3VPN, the ingress PE could be identified by the tunnel label when TE LSP applies [RFC3209], but the egress PE cannot tell to which specific VRF a packet belongs when extranet (If the various sites in a VPN are owned by different enterprises) exist on ingress PE. Figure 1 shows an example of extranet. In Figure1, Site A,B,C,D all belong to the same VPN-A, but Site C and Site D does not belong to the same enterprise (Site D also belongs to a VPN-B), so different VRFs are maintained for each site on PE3. PE1 assign the same label L for prefix 10.0.0.1 to PE3 of VPN-A, when it recieve the VPN-A flow from PE3, it can not tell the flow is from either VRFC or VRFD by the label stack.

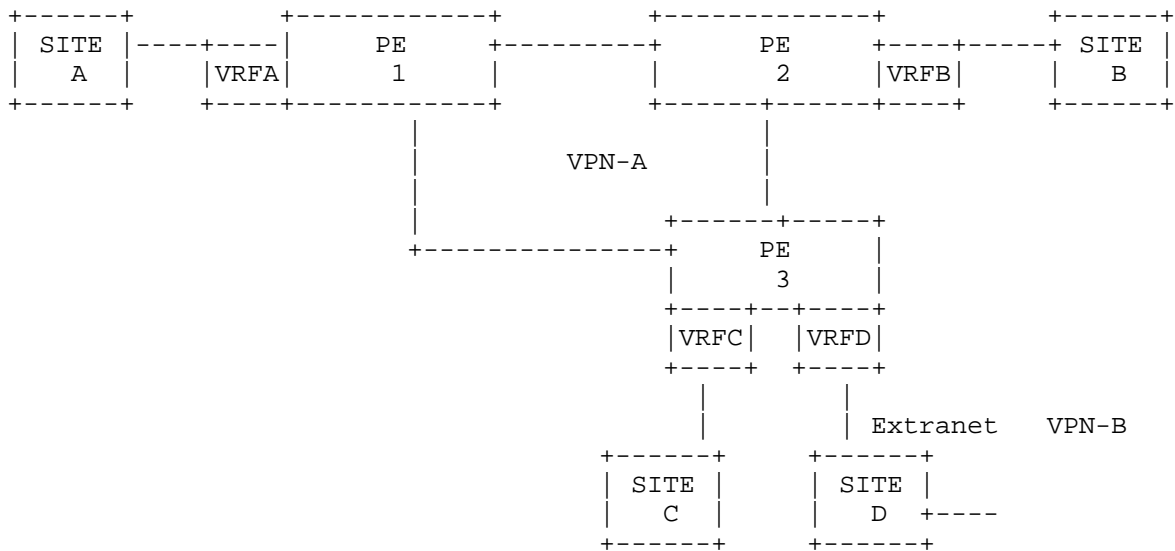


Figure1: Extranet on Ingress PE

The current label allocation mechanism of L3VPN make the flow identification a big challenge for L3VPN performace monitoring, as a result the current performace monitoring mechanisms for MPLS networks cannot be applied to L3VPN networks. Extension or alteration to current label allocation mechanism is needed to solve the problem.

4. Design Consideration

This section discuss the key points need to be taken in consideration when designing L3VPN performance monitoring mechanism.

4.1. P2P Connection

As analyzed above, to perform the packet loss or delay measurement on a specific VPN flow, it is critical for the egress PE to identify the unique ingress VRF, i.e. to establish the Point-to-Point connection between the two VRFs. Current allocation mechanism may need extension or alteration to help build up the Point-to-Point connection. Once the Point-to-Point connection is built up, current measurement mechanisms may be applied to L3VPN .

Conditions like Penultimate Hop Popping (PHP), Equal-Cost Multi-Path (ECMP) load-balancing and BGP multi-path may make it infeasible for receiving PE to identify the ingress PE. These conditions SHOULD be excluded for consideration for mechanism design.

4.2. Control Plane

In L3VPN, BGP is used to distribute a particular route, as well as an MPLS label that is mapped to that route [RFC4364]. The label mapping information for a particular route is piggybacked in the same BGP Update message that is used to distribute the route itself. In order to setup the Point-to-Point connection between ingress and egress VRFs the current label distribution mechanism may be altered. For compatibility, this alteration SHOULD NOT change the current label distribution mechanism dramatically.

4.3. Data Plane

Same as for control plane, for compatibility reason, the data plane should as far as be compatible with the current L3VPN forwarding procedure.

4.4. MPLS OAM

[RFC6374], [RFC6375] defines procedure and protocol mechanisms to enable the measurement of packet loss, delay, as well as related metrics in MPLS networks. These mechanisms SHOULD be reasonably reused in L3VPN networks. The addressing of source and destination of Loss Measurement (LM) and Delay Measurement (DM) messages may needed to be changed to identify the measured VRF.

4.5. QoS

To perform the packet loss or delay measurement in L3VPN network, either proactive or on-demand, SHOULD NOT impact the customer QoS experience.

4.6. Configuration

Measurement entities and functions MUST be configurable either statically or dynamically. It SHOULD be possible to configure and activated/deactivated the measurement capability as part of connectivity establishment, and it SHOULD also be possible to configure and activated/deactivated the capability after connectivity has been established .

5. Security Considerations

This document does not change the security properties of L3VPN.

6. IANA Considerations

This document makes no request to IANA.

7. Acknowledgements

The authors would like to thank XXX for their valuable comments.

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

8.2. Informative References

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4377] Nadeau, T., Morrow, M., Swallow, G., Allan, D., and S. Matsushima, "Operations and Management (OAM) Requirements for Multi-Protocol Label Switched (MPLS) Networks", RFC 4377, February 2006.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.
- [RFC6375] Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-Based Transport Networks", RFC 6375, September 2011.

Authors' Addresses

Lianshu Zheng
Huawei Technologies
China

Email: vero.zheng@huawei.com

Zhenbin Li
Huawei Technologies
China

Email: lizhenbin@huawei.com

