

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 7, 2016

K. Patel
Cisco Systems
E. Rosen
Juniper Networks
Y. Rekhter
October 5, 2015

**BGP as an MVPN PE-CE Protocol
draft-ietf-bess-mvpn-pe-ce-01**

Abstract

When a Service Provider offers BGP/MPLS IP VPN service to its customers, RFCs 6513 and 6514 describe protocols and procedures that the Service Provider can use in order to carry the customer's IP multicast traffic from one customer site to others. BGP can be used to carry customer multicast routing information from one Provider Edge (PE) router to another, but it is assumed that PIM is running on the interface between a Customer Edge (CE) router and a PE router. This document specifies protocols and procedures that, under certain conditions, allow customer multicast routing information to be carried between PE and CE via BGP. This can eliminate the need to run PIM on the PE-CE interfaces, potentially eliminating the need to run PIM on the PE routers at all.

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 7, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

- [1.](#) Introduction [2](#)
- [2.](#) C-MCAST SAFI NLRI Format [4](#)
 - [2.1.](#) C-MCAST Join and Prune Routes [5](#)
- [3.](#) Exchanging C-MCAST Join Routes [6](#)
 - [3.1.](#) Originating C-MCAST Join Route at the CE router [6](#)
 - [3.2.](#) Receiving a C-MCAST Join Route by the CE Router [8](#)
 - [3.3.](#) Originating a C-MCAST Join Route at the PE Router [8](#)
 - [3.4.](#) Receiving a C-MCAST Join Route by the PE Router [10](#)
- [4.](#) Pruning Sources off the Shared Tree [10](#)
- [5.](#) IANA Considerations [11](#)
- [6.](#) Security Considerations [11](#)
- [7.](#) References [11](#)
 - [7.1.](#) Normative References [11](#)
 - [7.2.](#) Informative References [12](#)
- Authors' Addresses [12](#)

1. Introduction

When a Service Provider offers BGP/MPLS IP VPN service to its customers, [[RFC6513](#)] and [[RFC6514](#)] describe protocols and procedures that the Service Provider can use in order to carry the customer's IP multicast traffic from one customer site to others. BGP can be used to carry customer multicast routing information from one Provider Edge (PE) router to another, but it is assumed that PIM is running on the interface between a Customer Edge (CE) router and a PE router. This document specifies protocols and procedures that under certain conditions, allow customer multicast routing information to be carried between PE and CE via BGP. This can eliminate the need to run PIM on the PE-CE interfaces, potentially eliminating the need to run PIM on the PE routers at all. It is however assumed that PIM is the multicast routing protocol running at the customer sites.

This document defines a new SAFI known as Customer-Multicast (C-MCAST) SAFI. This SAFI is used to carry customer multicast routing information from CE to PE (and vice versa). The use of this SAFI is only defined when the AFI is either IPv4 or IPv6. The

procedures of this document are applicable only if BGP is being used as the PE-PE control protocol for carrying customer multicast routing. It is presupposed that if these procedures are being used on any interface of a given VRF, then PIM is NOT enabled on that interface or on any other VRF interface of that same VRF. It is also assumed that if a CE is using BGP C-MCAST on its interface to one PE, then it is using BGP C-MCAST on its interfaces to all the PEs to which it is connected, and that PIM is not enabled on any of these interfaces.

Throughout this document, we will use the terms "MCAST-VPN route" and "C-MCAST route" to mean routes that have the corresponding SAFI.

This document assumes that a CE and a PE exchange C-MCAST routes over a direct BGP session (i.e., the C-MCAST routes do not pass through a route reflector or other third party on their way from CE to PE, or vice versa).

The NLRI format of a C-MCAST route is modeled on the NLRI format of the MCAST-VPN SAFI, as defined in [[RFC6514](#)]. However, since the C-MCAST routes are always exchanged in the context of a particular VPN, Route Distinguishers (RDs) are not used. Also, C-MCAST routes are never distributed from one PE to another.

Where the procedures of this document require a PE or a CE to determine the upstream neighbor for a particular multicast (*,G) or (S,G) state, the upstream neighbor is determined using the procedures of [[RFC4601](#)], rather than the procedures of [[RFC6513](#)] or [[RFC6514](#)].

That is, the upstream neighbor selected for a particular (*,G) or (S,G) is the same as it would be if PIM were being used between the PE and CE. This includes support for non-congruent multicast topologies. The upstream neighbor address determined through these procedures MUST be the same address that the upstream neighbor puts in the next hop field of BGP Updates that it sends to the "downstream" PE or CE. Note that neither the VRF Route Import Extended Community nor the Source AS Extended Community, defined in [[RFC6514](#)], are used.

When following the procedures of this document, customer IP multicast traffic is sent natively from CE to PE, and vice versa. There is no encapsulation or tunneling of the multicast traffic. Therefore there is no need for C-MCAST routes corresponding to the I-PMSI A-D routes or S-PMSI A-D routes or [[RFC6514](#)]. Similarly, there is no need for the PMSI Tunnel attribute of [[RFC6514](#)].

This document does not provide a mechanism corresponding to the "PIM Assert" mechanism of [[RFC4601](#)]. It is assumed either that the PE-CE

interfaces are point-to-point interfaces, or else that the multicast procedures in the PE and CE can determine, by examination of the layer 2 headers, the node from which a given multicast data packet was received.

Support for "Dense Mode Multicast" (PIM-DM) is outside the scope of this document.

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

2. C-MCAST SAFI NLRI Format

The BGP Multiprotocol Extensions [RFC4760] allow BGP to carry routes from multiple different "AFI/SAFIs". This document defines a new new SAFI known as a C-MCAST SAFI with a value to be assigned by the IANA. This SAFI is used along with the AFI of IPv4 (1) or IPv6 (2).

The C-MCAST NLRI defined below is carried in the BGP UPDATE messages [RFC4271] using the BGP multiprotocol extensions [RFC4760] with a AFI of IPv4 (1) or IPv6 (2) assigned by IANA and a C-MCAST SAFI with a value to be assigned by the IANA.

The Next hop field of MP_REACH_NLRI attribute SHALL be interpreted as an IPv4 adress whenever the length of the Next Hop address is 4 octets, and as an IPv6 address whenever the length of the Next Hop is address is 16 octets.

The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a prefix with a maximum length of 12 octers for IPv4 AFI and 36 octets for IPv6 AFI. The following is the format of the C-MCAST NLRI:

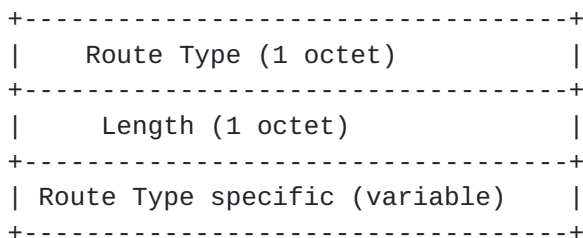


Figure 1

The Route Type field defines encoding of the rest of the C-MCAST NLRI. (Route Type specific C-MCAST NLRI).

The Length field indicates the length in octets of the Route Type specific field of C-MCAST NLRI.

This document defines the following Route Types:

- 1 - Shared Tree Join Route
- 2 - Source Tree Join Route
- 4 - Source Prune A-D Route

The encodings and procedures for these route types are described in the subsequent sections. The NLRI encodings are modeled after those of [RFC6514], in order to facilitate implementation in generation of MCAST-VPN routes.

In order for two BGP speakers to exchange C-MCAST NLRI, they must use BGP Capabilities Advertisement [RFC5492] to ensure that they both are capable of properly processing the C-MCAST NLRI. This is done as specified in [RFC4760], by using a capability code 1 (multiprotocol BGP) with an AFI of IPv4 (1) or IPv6 (2) and a SAFI of C-MCAST with a value to be assigned by IANA.

2.1. C-MCAST Join and Prune Routes

The "route type specific" part of the C-MCAST NLRI is the same for the Shared Tree Join, Source Tree Join, and Source Prune A-D Route Types.

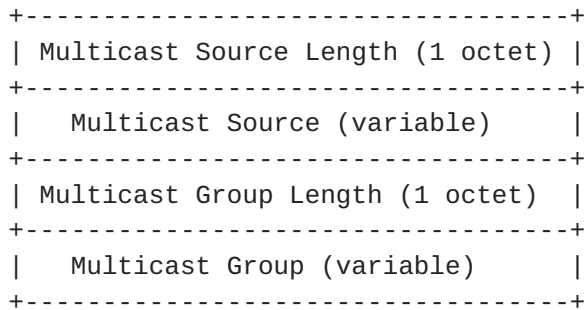


Figure 2

The value of the AFI field in the MP_REACH_NLRI/MP_UNREACH_NLRI that carries the C-MCAST NLRI determines whether the multicast source and multicast group addresses fields are IPv4 addresses (AFI 1) or IPv6 addresses (AFI 2). The length field of IPv4 source and group addresses is 32 bits and the length field of IPv6 addresses is 128 bits.

Use of other values of the Multicast Source Length and Multicast Group Length fields is outside the scope of this document. If other values occur, or if the NLRI length is not as expected, given the AFI

value, the NLRI should be considered to be malformed. An implementation SHOULD treat such an UPDATE as though the NLRI has been withdrawn, SHOULD log an error.

In a Source Tree Join route, the Multicast Source field and the Multicast Group field identify an (S,G) multicast flow, where the IP address of S appears in the Multicast Source field and the IP address of G appears in the Multicast Group field.

In a Shared Tree Join route, the Multicast Source field contains the Rendezvous Point (RP) address that is associated, at the originator of the UPDATE, with the multicast group whose address appears in the Multicast Group field.

In the Source Prune route, the Multicast Source field contains the address of a source that is to be pruned from the shared tree associated with the multicast group whose address appears in the Multicast Group field.

Usage of C-MCAST Join routes is described in [Section 3](#). Usage of C-MCAST Source Prune routes is described in [Section 4](#).

[3](#). Exchanging C-MCAST Join Routes

Multicast Routing Information is exchanged between a CE router and a PE a router using C-MCAST NLRIs. If the procedures of this draft are followed, the CE router and the PE router MUST NOT exchange PIM messages with each other. The C-MCAST routes are originated and propagated as follows.

[3.1](#). Originating C-MCAST Join Route at the CE router

Whenever a) PIM instance on a particular CE router creates a new (S,G) or (*,G) state and b) the selected upstream neighbor address for that state is a PE router to which the CE router is directly connected and with which the CE has a BGP session on which the C-MCAST SAFI is enabled, then the CE router MUST originate a C-MCAST Source Tree Join route or a C-MCAST Shared Tree Join route. The CE router uses the procedures of [\[RFC4601\]](#) to determine the upstream neighbor (also called the "RPF neighbor"). Note that, unlike [\[RFC6513\]](#), the upstream nexthop selection procedures defined in this document are not based on the VRF Route Import Extended community [\[RFC6513\]](#).

The Route Type field of the NLRI is set to "Source Tree Join" if the corresponding state is (S,G), and to "Shared Tree Join" if the corresponding state is (*,G).

The Multicast Source field of the NLRI of a C-MCAST Source Tree Join route MUST be set to the source address of the corresponding (S,G). The Multicast Source field of the NLRI of a C-MCAST Shared Tree Join route MUST be set to the address of the RP that is mapped to the corresponding (*,G) state.

The Multicast Group field of the NLRI of either kind of C-MCAST Join route MUST be set to the multicast group address of the (S,G) or (*,G) state.

The CE router MUST put its own IP address in the Next Hop field of either type of C-MCAST Join route.

The CE router MUST attach a particular RT to the C-MCAST Join route. This RT MUST be either an IP Address Specific RT or an IPv6 Address Specific RT, depending upon whether the address of the upstream neighbor is an IPv4 address or an IPv6 address. In either case, the address of the upstream neighbor is placed in the Global Administrator field of the RT, and the Local Administrator field is set to 0. (Compare to the "C-multicast Import RT" described in [section 7 of \[RFC6514\]](#).) The "address of the upstream neighbor" MUST be the address that the upstream neighbor puts in the Next Hop field of BGP UPDATES that it sends to the CE router.

The CE MUST send the C-MCAST Join route on the BGP session to the PE that it has selected as the upstream neighbor for the (*,G) or (S,G) identified in the NLRI of the route. Although not required for correctness, it is RECOMMENDED that the C-MCAST Join not be sent on other BGP sessions. How this distribution constraint is met is a local matter. Policy based filtering based on the RT is one possible way to meet the constraint. Use of [\[RFC4684\]](#) is another.

The C-MCAST Shared Tree Join is used to join the shared tree of an "Any Source Multicast" (ASM) group. The C-MCAST Source Tree Join is used in both ASM mode and in "Single Source Multicast" (SSM) mode to join a source tree. The C-MCAST Shared Tree Join can be used to join a BIDIR-PIM [\[RFC5015\]](#) multicast group, as long as the interface between the PE and the CE is a point-to-point interface.

If a PIM instance on a particular CE router deletes its (S,G) or a (*,G) entry, and if there is a currently originated corresponding C-MCAST Join route, then that route MUST be withdrawn. The C-MCAST route is withdrawn using the MP_UNREACH_NLRI attribute. If the upstream neighbor of a (S,G) or (*,G) route changes, and if there is a currently originated corresponding C-MCAST Join route, then the RT MUST be changed to identify the new upstream neighbor, and the route MUST be advertised on the BGP session to that neighbor.

3.2. Receiving a C-MCAST Join Route by the CE Router

When a CE router receives a C-MCAST route from its PE router, it checks to see if the route carries an IP Address Specific Route Target Extended Community whose Global Administrator field contains the address that the CE puts in the Next Hop field of the BGP UPDATES it sends to the PE router. If there is no such Route Target, the received route does not impact the CE's multicast state. Otherwise, the received route is used to create or modify the corresponding (S,G) or (*,G) state in the multicast forwarding information base (FIB). The CE-to-PE interface is stored in the outgoing interface list of the corresponding (S,G) or a (*,G) state. If the C-MCAST route is received with the Route Type of Shared Tree Join, then the CE router attaches RP address to the corresponding (*,G) state.

If the CE router is connected to the multiple PE routers, it is possible that it will receive C-MCAST routes with the same NLRI from multiple PE routers. In this case, the CE router adds multiple CE-to-PE interfaces to its outgoing interface list, one for each PE router from which the given route was received. (Note that a particular CE-to-PE interface is added only if the route from the corresponding PE identifies the CE in its IP Address Specific RT.)

When the last C-MCAST route for a given (S,G) or a (*,G) is withdrawn, resulting in a state where BGP C-MCAST SAFI has no route for a given (S,G) or a (*,G) state, the CE router MUST remove all the interfaces learnt via BGP from the outgoing interface list. If this results in an empty outgoing interface list then the CE router using PIM procedures MUST prune itself off the corresponding (S,G) or a (*,G) tree.

3.3. Originating a C-MCAST Join Route at the PE Router

The C-MCAST routes on a PE router are originated in BGP as a result of updates in the (C-S,C-G) or (C-*,C-G) state in the associated VRF of a PE router. These states are created by BGP routes learnt from other PEs via the BGP MCAST-VPN SAFI [[RFC6514](#)], or from CEs via the C-MCAST SAFI. Whenever a) a particular VRF on a particular PE router creates a new (C-S,C-G) or a (C-*,C-G) state, b) the selected upstream neighbor address identifies a CE, and c) the PE has a direct BGP session to the CE, on which C-MCAST SAFI is enabled, then the PE router MUST originate a C-MCAST Join route from the given VRF. The PE router finds the upstream neighbor address based on the procedures of [[RFC6513](#)].

The Route Type field of the NLRI is set to "Source Tree Join" if the corresponding state is (S,G), and to "Shared Tree Join" if the corresponding state is (*,G).

The Multicast Source field of the NLRI of a C-MCAST Source Tree Join route MUST be set to the source address of the corresponding (S,G). The multicast source address field of the NLRI of a C-MCAST Shared Tree Join route MUST be set to the address of the RP that is mapped to the corresponding (*,G) state.

The Multicast Group field of the NLRI of either kind of C-MCAST Join route MUST be set to the multicast group address of the (S,G) or (*,G) state.

The PE router must put its own IP address in the Next Hop field of the C-MCAST Join route.

The PE router MUST attach a particular RT to the C-MCAST Join route. This RT MUST be either an IP Address Specific RT or an IPv6 Address Specific RT, depending upon whether the address of the upstream neighbor is an IPv4 address or an IPv6 address. In either case, the address of the upstream neighbor is placed in the Global Administrator field of the RT, and the Local Administrator field is set to 0. (Compare to the "C-multicast Import RT" described in [section 7 of \[RFC6514\]](#).) The "address of the upstream neighbor" MUST be the address that the upstream neighbor puts in the Next Hop field of BGP UPDATES that it sends to the PE router.

The PE MUST send the C-MCAST Join route on the BGP session to the CE that it has selected as the upstream neighbor for the (*,G) or (S,G) identified in the NLRI of the route. Although not required for correctness, it is RECOMMENDED that the C-MCAST Join not be sent on other BGP sessions. How this distribution constraint is met is a local matter. Policy based filtering based on the RT is one possible way to meet the constraint. Use of [\[RFC4684\]](#) is another. Of course, a C-MCAST route originated by a particular PE is originated in the context of a particular VRF, and MUST NOT be advertised to a particular CE unless the interface between that PE and that CE is associated with that VRF.

The C-MCAST Shared Tree Join is used to join the shared tree of an "Any Source Multicast" (ASM) group. The C-MCAST Source Tree Join is used in both ASM mode and in "Single Source Multicast" (SSM) mode to join a source tree. The C-MCAST Shared Tree Join can be used to join a BIDIR-PIM [\[RFC5015\]](#) multicast group, as long as the interface between the PE and the CE is a point-to-point interface.

If a PE router deletes its (S,G) or a (*,G) entry in the context of a particular VRF, and if there is a currently originated corresponding C-MCAST Join route from that VRF, then that route MUST be withdrawn. The C-MCAST route is withdrawn using the MP_UNREACH_NLRI attribute. If the upstream neighbor of a (S,G) or (*,G) route changes, and if

there is a currently originated corresponding C-MCAST Join route, then the RT MUST be changed to identify the new upstream neighbor.

3.4. Receiving a C-MCAST Join Route by the PE Router

When a PE router receives a C-MCAST Join route from a CE router, it checks to see if the route carries an IP Address Specific Route Target Extended Community whose Global Administrator field contains the address that the PE puts in the Next Hop field of the BGP UPDATES it sends to the CE router. If there is no such Route Target, the received route does not impact the PE's multicast state. Otherwise, the received route is used to create or modify the corresponding (S,G) or (*,G) state in the MVPN-TIB ([RFC6514]). The PE-to-CE interface is stored in the outgoing interface list of the corresponding (S,G) or a (*,G) state. If the C-MCAST route is received with the Route Type of Shared Tree Join, then the CE router attaches RP address to the corresponding (*,G) state.

If the PE router is connected to the multiple CE routers, it is possible that it will receive C-MCAST routes with the same NLRI from multiple CE routers. In this case, the PE router adds multiple PE-to-CE interfaces to its outgoing interface list, one for each CE router from which the given route was received. (Note that a particular PE-to-CE interface is added only if the route from the corresponding CE identifies the PE in its IP Address Specific RT.) When the last C-MCAST route for a given (S,G) or a (*,G) is withdrawn, resulting in a state where BGP C-MCAST SAFI has no route for a given (S,G) or a (*,G) state, the withdrawal of the route has the (S,G) or a (*,G) prune semantics. The corresponding MCAST-VPN route is withdrawn using the MP_UNREACH_NLRI attribute.

4. Pruning Sources off the Shared Tree

Suppose a router, say R1, has originated a C-MCAST Source Tree Join A-D route for (*,G), and has identified another router, say R2, in that route's RT. Under certain conditions, R1 may need to prune a particular source, say S1, off the (*,G) tree. To do so, R1 originates a C-MCAST Prune Source A-D route whose NLRI contains S1 in the multicast source field and G in the multicast group field. This route MUST have the same IP Address Specific RT (identifying R2), and the same Next Hop field, as the corresponding C-MCAST Source Tree Join A-D route. This route MUST be sent on the BGP session to R2. It is RECOMMENDED that it not be sent on other BGP sessions. Note that either R1 is a CE and R2 is a PE, or vice versa. The procedures are the same in either case. When R1 no longer needs to prune S1 from the (*,G) tree, R1 MUST withdraw the (S1,G) Source Prune route. If R1 changes the RT on the (*,G) Shared Tree Join route, it MUST change the RT on all the corresponding (S,G) Source Prune routes. If

R1 withdraws the (*,G) Shared Tree Join route, it MUST also withdraw all the corresponding (S,G) Source Prune routes. When a router withdraws a Source Tree Join route for (*,G), it MUST withdraw all its Source Prune (S,G) routes. A received Source Prune (S,G) route does not impact a router's multicast state unless there is a the router has also received a Shared Tree Join (*,G) route over the same BGP session. However, a router should handle the case where one or more Source Prune routes are received before the corresponding Shared Tree Join route. Further details will be provided in future revisions of this document.

5. IANA Considerations

This document define a new SAFI called "C-MCAST SAFI". IANA is requested to allocate a code point for C-MCAST SAFI.

6. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [[RFC4271](#)] and [[RFC6514](#)].

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", [RFC 4271](#), DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", [RFC 4601](#), DOI 10.17487/RFC4601, August 2006, <<http://www.rfc-editor.org/info/rfc4601>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", [RFC 4760](#), DOI 10.17487/RFC4760, January 2007, <<http://www.rfc-editor.org/info/rfc4760>>.

- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", [RFC 5015](#), DOI 10.17487/RFC5015, October 2007, <<http://www.rfc-editor.org/info/rfc5015>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", [RFC 5492](#), DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.
- [RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", [RFC 6513](#), DOI 10.17487/RFC6513, February 2012, <<http://www.rfc-editor.org/info/rfc6513>>.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", [RFC 6514](#), DOI 10.17487/RFC6514, February 2012, <<http://www.rfc-editor.org/info/rfc6514>>.

7.2. Informative References

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

Authors' Addresses

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

Email: keyupate@cisco.com

Eric C. Rosen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: erosen@juniper.net

Yakov Rekhter

