Network Working Group
Internet Draft
Intended Status: Informational
Expires: May 2008

Yiqun Cai
Mike McBride
Chris Hall
Maria Napierala

November 2007

PIM Based MVPN Deployment Recommendations

draft-ycai-mboned-mvpn-pim-deploy-00.txt

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of 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/1id-abstracts.txt.

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

Copyright Notice

Copyright (C) The IETF Trust (2007).

Internet Draft draft-ycai-mboned-mvpn-pim-deploy-00.txt November 2007

Abstract

Multicast VPN, based on pre-standard drafts, has been in operation in production networks for many years. This document describes some of the practices and experiences gained from implementation and deployment of MVPN using PIM with GRE tunnels. It is informational only.

Table of Contents

1 Introduction 3
2 Implementation 3
2.1 RPF 3
2.2 MTU 4
2.3 EIBGP Load Balancing 4
2.4 MTRACE 5
3 Operational Experience 5
3.1 Multicast VPN Design Considerations 5
3.2 PIM Modes For MI-PMSI 6
3.2.1 PIM-SSM for MI-PMSI 6
3.2.2 ASM for MI-PMSI 6
3.3 PIM Modes For S-PMSI 7
3.4 CE to PE PIM Modes 7
3.5 Timer Alignment 8
3.6 Addressing 8
3.7 Filtering 8
3.8 Scalability 9
3.9 QOS 9
4 Security Considerations 10
5 Iana Considerations 10
6 Acknowledgments 10
7 Normative References 10
8 Informative References 11
9 Authors' Addresses 11
10 Full Copyright Statement 11
11 Intellectual Property 12

1. Introduction

Multicast support for L3VPN based on RFC2547 [2547bis] was first presented in San Diego IETF, 2000. It had not been included in the charter of L3VPN (formerly PPVPN) working group until San Diego IETF in 2004 and stayed on as an individual submission. During the time, the draft, known as "rosen-draft", continued to evolve as pre-standards work. Several vendors provided implementations based on this draft. Service providers began deploying this mvpn solution in production networks.

Once the working group officially accepted the challenge to define a solution or solutions to support multicast, several proposals have been suggested. They are now captured in [MVPN] which forms a base for future standards work.

This document provides MVPN deployment experience based solely on the original PIM and GRE based MVPN solution. This solution is now outlined as one of the options in the standards track [MVPN] document.

In this document, we describe some of the lessons learned from implementing and deploying MVPN. We hope it will benefit implementors as well network operators looking to deploy MVPN services. Throughout the document, where the term "MVPN" is used, the reference is to the original MVPN deployment based upon the Rosen-08 GRE tunnels.

2. Implementation

There are two known MVPN implementations: IOS from Cisco and JunOS from Juniper Networks. Contact these vendors for implementation details beyond what is provided in this draft. The following sections describe common mvpn deployment considerations.

2.1. RPF

[MVPN] specifies that the source address of any PIM packets that a PE router generates over the MDT tunnel must be the same as the BGP nexthop for updates originated by the PE router for all multicast traffic sources existing in the site. Otherwise, a PE router will not resolve the RPF neighbour towards the source connected to a remote PE router.

A PE needs to have a particular IP address which it uses in both the IP source address field of the PIM packet and the next hop field of

the BGP updates. If this requirement is overlooked, RPF determination may fail. This has caused interoperability problems in the past and implementors should be careful about it in the future.

2.2. MTU

When GRE encapsulation is used in the core, 24 bytes are added to the IP packets generated in the VPNs. Due to the lack of a path MTU discovery mechanism for multicast, a PE router may have to fragment the incoming packets.

The best practice is to fragment the packets before performing any GRE encapsulation. This spares the egress PE routers from reassembling the fragments, and leaves that for the end-systems. This doesn't work if the "DF" bit is set in the original packet since the packet will be dropped. Its best to ensure that the backbone does not have any links with a 1500 byte MTU.

It is further recommended to read Section 5.1 of [Worster] for a more detailed look at preventing fragmentation and reassembly.

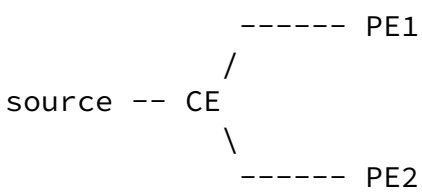
2.3. EIBGP Load Balancing

External and Internal Border Gateway Protocol (eiBGP) load sharing is an enhancement to BGP that enables load sharing over parallel links between CE and PE routers. EIBGP enables service providers to share customer traffic loads over parallel paths within an MPLS core

network.

When EIBGP load balancing is enabled on all PE routers, we have seen that multicast RPF check, inside the VRF, may be affected if the path towards the source resolves via iBGP. The best practice is to ensure that, when both eBGP and iBGP routes are present, multicast RPF selects eBGP paths only.

Example:



With EIBGP configured, PE1 will have two paths towards the source, one directly via the CE using eBGP, and one via PE2 using iBGP.

By default PIM will pick the neighbor with the highest IP address. If

this happens to be PE2, the RPF check will fail as it will use the global table.

A workaround would be either a static mroute on PE1 pointing towards the CE router, or make sure that CE has a higher ip address than PE2. The better solution is additional logic in the RPF code that where EIBGP is used the EBGP link is preferred.

2.4. MTRACE

MTRACE is a tool that allows a network operator to obtain multicast routing information from routers and to explore a path to the source of the traffic or the RP.

Since there is no security mechanism embedded in MTRACE, some service providers express concern when the mtrace packet has to traverse the PE routers in order to obtain the full information.

Vendors have their own mechanisms to remove, or hide, certain fields

in the MTRACE packets in order to satisfy the needs of their customers. We need to define a better mechanism for MTRACE in an MVPN environment.

3. Operational Experience

3.1. Multicast VPN Design Considerations

When deploying a multicast VPN service, providers try to optimize multicast traffic distribution and delays while reducing the amount of state. The following considerations have given MVPN providers direction in their MVPN deployment:

+ Core multicast routing states should typically be kept to a minimum

+ MVPN packet delays should typically be the same as unicast traffic

+ Data should typically be sent only to PEs with interested receivers

3.2. PIM Modes For MI-PMSI

The MI-PMSI ("Default-MDT" of pre-standard drafts) is used to build an overlay network connecting all PE routers attaching to the same MVPN.

Service providers have implemented PIM-SM and PIM-SSM instantiated MI-PMSI in production networks. The majority of MI-PMSI deployments are using PIM-SM using static Anycast RP with MSDP assignment. But a dynamic RP discovery protocol, such as BSR, is also being used.

The decision to deploy either PIM-SM or PIM-SSM is based on the following concerns,

+ the number of multicast routing states

+ the overhead of managing the RP if PIM-SM is used

+ the difference of forwarding delay between shared tree and source trees

3.2.1. PIM-SSM for MI-PMSI

Optimal MVPN forwarding is most easily achievable when there is a single multicast tree per MVPN per PE. Such trees are naturally built with PIM-SSM since it permits the PE to directly join a source tree for an MDT. With PIM-SSM, no Rendezvous Points are required. With SSM, however, all PEs on an MVPN tree need to maintain source state. Each PE, which is participating in MVPN, is a source. Unless VPN customers locate their multicast sources within a constrained set of sites, SSM may become a scalability concern in the service providers network.

3.2.2. ASM for MI-PMSI

One solution to minimize the amount of multicast state in an MVPN environment is to configure PIM-SM or BIDIR PIM to stay on the shared tree. With shared trees, multicast state scalability is no longer a function of the number of PE's but rather of the number of VPNs.

The scale benefit of shared trees comes at the cost of less efficient multicast distribution. MVPN providers use the MI-PMSI to achieve bandwidth optimality. MVPN providers may address the sub-optimality of shared tree forwarding by deploying an RP at the best location for

each VPN. Such an assignment would be based on the VPN source locations, something which may be difficult to maintain.

3.3. PIM Modes For S-PMSI

The S-PMSI ("Data MDT" of pre-standard drafts) has also been widely deployed by service providers. While both PIM-SM and PIM-SSM are used, PIM-SSM is the more widely deployed, and recommended, S-PMSI tree building model. The majority of S-PMSI deployments today are using SSM since the source address is included in the PIM Hello packet sent from the source PE.

MVPN providers deploy the S-PMSI to achieve optimal bandwidth usage, especially when SSM is deployed as well. S-PMSIs are optimized for active sources and receivers and triggered per (S,G) for a subset of (S,G) of a given VPN. Since S-PMSIs are triggered by (S,G) states in a VPN, they could increase the amount of multicast states in an MVPN network.

The decision to switch from MI-PMSI to S-PMSI is always made by the ingress PE based upon the traffic load exceeding a configurable threshold.

3.4. CE to PE PIM Modes

The PIM protocols that are deployed within the customer VPN are independent of the PIM Protocols in use within the Provider core. Customers can choose to deploy PIM-DM, PIM-SM, Bidir, or SSM.

With SM or Bidir, customers may choose to deploy the RP on either a PE or CE router. It is generally recommended to have a CE router serve as the RP. This is done primarily to avoid an increase in customer/provider interaction on matters such as the integration of the PE/RP into the customer chosen RP discovery mechanism and to avoid any additional burden on a busy PE router. While RP deployment is most commonly performed on CE routers, we have seen RPs deployed successfully on PE as well as CE routers.

If a customer desires to have a provider managed RP, they should consider requesting the service provider manage a CE and have it serve as the RP. To avoid managing an RP altogether, SSM should be deployed.

3.5. Timer Alignment

When PIM-SM is used in an SP's core MVPN environment, some interesting observations were made. When BSR, for example, is used in the service provider network, to discover RPs in the provider tunnel, it takes more than 3 minutes to detect the failure of the RP if the default timer is used. During the window, PIM Hellos originated by C-PIM instances will be dropped, which cause PIM adjacencies to be torn down. But since the default PIM Hello timer is 30 seconds, C-PIM instance on a PE router detects an outage much faster than the P-PIM instance on the same PE router.

This is a factor to be considered when choosing the protocol for RP redundancy and fast failover. One option, for fast failover, is to use BSR only for RP discovery and then utilize Anycast-RP for RP redundancy.

3.6. Addressing

It has become general practice to use 239/8 private address space when assigning address space to mvpn's. This helps to prevent vpn traffic from being sent outside the mvpn core. When SSM is used, 239.232/16 addressing is the common practice according to [Meyer], Administratively Scoped IP Multicast. Operators typically deploy an addressing tool to manage their addresses.

This addressing practice can also be used to prevent non-VPN traffic, originating outside the SP boundaries, from entering a VPN.

The reader should also reference section 11.5.4 of the [MVPN] draft entitled "Avoiding Conflict with Internet Multicast".

3.7. Filtering

Filtering at the SP boundaries is needed to prevent VPN security violations. It may be necessary to modify these deployed filters to permit GRE and possibly UDP port 3232. UDP port 3232 is the UDP port used for the S-PMSI Join messages.

3.8. Scalability

PIM retransmission overhead on a given MI-PMSI increases in linear proportion to any increase in the number of PEs that join the MI-PMSI. The overhead also increases in linear proportion with an increase in the number of J/P messages received from the CEs.

There have been no scaling issues with current deployments of MVPN. Current MVPN deployments consist of up to a few hundred sites per MVPN. The number of PE's participating in a MI-PMSI continues to increase as customers extend the multicast group participation to additional VPN sites. There are unicast VPN customers with several thousand sites. These sites are gradually becoming multicast enabled. The number of J/P messages received from CEs will also increase over time.

At some level of scaling of the MI-PMSI, PIM Hello's and J/P messages will become a scaling issue. The scaling point at which these messages become a real operational problem is not clear. Empirical field data shows they do not affect the broad range of MVPN deployments today. MVPN is scalable as specified across a wide range of deployments.

Some analysis is needed to clarify at what operational level PIM messages do become a problem. The L3VPN WG has gathered requirements information in [MORIN]. A benchmarking draft [DRY] has been submitted to the BMWG to provide consistent MVPN test methodology. The PIM WG is evaluating methods to decrease PIM messages when this becomes of operational value. Extensions to PIM such as PIM J/P Acks and TCP based approaches are being evaluated by the working group.

Increasing the Hello timer and increasing the periodic join/prune timer may also help in future MVPN scaling. Doing so, however, may affect join and leave latency in times when control messages are lost. OAM, to verify the health of the data and control paths, would also be affected if the Hello timer were increased or removed altogether.

3.9. QOS

Deployments of MVPN, that have deployed QOS, typicall use the same

QOS mechanisms for the MVPN GRE header that they would use for their other data traffic. VPN customers may want to separate the queuing of multicast data from unicast data. Service Providers are extending their QOS portfolio to support more classes of service to allow for better separation of multicast and unicast traffic. Enhanced QOS

mechanisms support applications with short bursts but which require bounded delay (such as video streaming). Since multicast (UDP) traffic might not be subject to the same drop behavior as TCP traffic, QOS profiles support Weighted Random Early Detection (WRED) treatment.

4. Security Considerations

The use of GRE encapsulations and IP Multicast has certain security implications. As discussed in [Farinacci], security in a network using GRE should be relatively similar to security in a normal IPv4 network. And Section 6 of [Fenner] clearly outlines the various security concerns related to PIM and how to use IPsec to secure the protocol.

5. Iana Considerations

This document does not require any action on the part of IANA.

6. Acknowledgments

We'd like to thank Dino Farinacci, Yuji Kamite, Hitoshi Fukuda and Eric Rosen for their feedback on this draft.

7. Normative References

[2547bis] "BGP/MPLS VPNs", Rosen, Rekhter, et. al., February 2006, RFC 4364

[MVPN] "Multicast in MPLS/BGP IP VPNs", Rosen, Aggarwal, July 2007, draft-ietf-l3vpn-2547bis-mcast-05.txt

8. Informative References

[MORIN] T. Morin, "Requirements for Multicast in L3 Provider-Provisioned VPNs", RFC 4834

[DRY] S. Dry, "Multicast VPN Scalability Benchmarking", draft-sdry-bmwg-mvpnscale-02.txt

[Meyer] D. Meyer, "Administratively Scoped IP Multicast". RFC 2365

[Farinacci] D. Farinacci, "Generic Routing Encapsulation (GRE)". RFC 2784

[Fenner] B. Fenner, "Protocol Independent Multicast - Sparse Mode (PIM-SM)". RFC 4601.

[Worster] T. Worster, "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)" RFC 4023.

9. Authors' Addresses

Yiqun Cai
ycai@cisco.com

Mike McBride
mmcbride@cisco.com

Chris Hall

chall@sprint.net

Maria Napierala
mnapierala@att.com

10. Full Copyright Statement

Copyright (C) The IETF Trust (2007).

This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS

OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

11. Intellectual Property

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.