

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 4, 2012

M. Bhatia
Alcatel-Lucent
D. Zhang
Huawei
July 3, 2011

In-Band Authentication Extension for Protocol Independent Multicast
(PIM)
draft-bhatia-zhang-pim-auth-extension-00

Abstract

Existing security mechanisms for the Protocol Independent Multicast - Sparse Mode (PIM-SM) routing protocol mandates to use IPsec to provide message authenticity and integrity. This draft proposes an embedded authentication mechanism to facilitate data origin authentication and integrity verification for PIM packets in the cases where IPsec is not applied.

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 January 4, 2012.

Copyright Notice

Copyright (c) 2011 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. Proposed Solution	3
3. PIM Security Association	5
4. AEP Packet Processing	6
4.1. Cryptographic Aspects	6
4.2. Outbounding Packet Processing	8
4.3. Inbounding Packet Processing	8
5. Security Considerations	9
5.1. Register Packet Processing	9
5.2. New Packet Type Versus Authentication Trailer	9
5.3. Inter-Session Replay Attack Issue	9
6. Acknowledgements	9
7. References	10
7.1. Normative References	10
7.2. Informative References	10
Authors' Addresses	11

1. Introduction

[RFC5796] describes the methods of using the IP security (IPsec) Encapsulating Security Payload (ESP) [RFC4303] or the Authentication Header (AH) [RFC4302] (which is optional) to protect the authenticity and integrity of the link-local messages of Protocol Independent Multicast - Sparse Mode (PIM-SM) [RFC4601]. [RFC5796] mandates the application of manual key management mechanisms and provide optional support for an automated group key management mechanism. However, the procedures for implementing automated group key management are left undone yet.

It has been clarified in [I-D.bhatia-karp-pim-gap-analysis] that without the support of automated group key management mechanisms, the PIM packets protected by IPsec will be vulnerable to both inter-session and inner-session replay attacks. In addition, the poor scalability of manual keying may cause deployment issues in many typical scenarios. This document proposes a new type of PIM packet, called the Authentication Extension PIM packet (AEP), which is able to facilitate data origin authentication and message integrity verification for PIM packets without the support of IPsec. An AEP actually encapsulates all the essential information of a PIM packet being protected and provides cryptographic methods for the receiver to assess the authenticity and integrity of the packet. In this solution, it is assumed that manual keying is performed while the automatic key management mechanisms are not precluded. Within a packet proposed in this document, a monotonically increasing sequence number is adopted to address the replay attack issues. However, the work of addressing the scalability issues imposed by manual keying is out of scope of this draft.

2. Proposed Solution

Figure 1 illustrates the format of an example packet header.

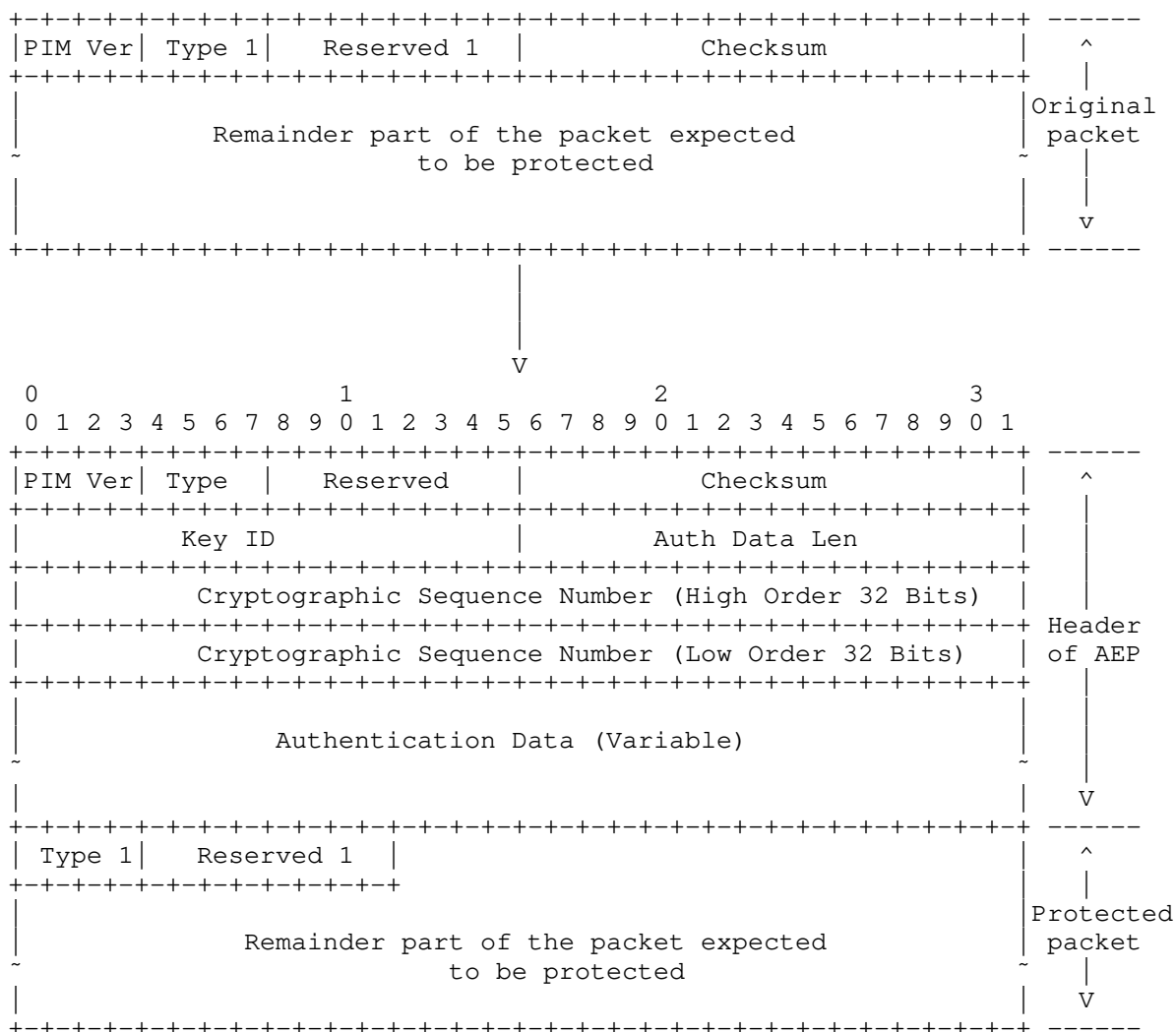


Figure 1. The format of an example AEP

In compliance with [RFC4601], the first four fields of the AEP header is identical to those of the original types of PIM packets. Particularly, the PIM Version number is set to 2. The type number of AEP is 9 in order to distinguish AEP from other types of PIM packets. The Reserved field is set to zero on transmission and ignored upon receipt. The checksum field of the AEP is set to zero, and the checksum calculation and verification are omitted.

Other fields of in the AEP header are described as follows:

Key ID: A 16-bit field that identifies the secret key and the algorithm used to create the authentication data.

Cryptographic Sequence Number: A 64-bit strictly increasing sequence number that is used to guard against replay attacks. The 64-bit sequence number **MUST** be incremented for every AEP packet sent by a PIM router. Upon reception, the sequence number **MUST** be greater than the sequence number in the last AEP packet accepted from the PIM router sending the packet. Otherwise, the AEP packet is considered a replayed packet and dropped. PIM routers implementing this specification **SHOULD** use available mechanisms to preserve the sequence number's strictly increasing property for the deployed life of the PIM router (including cold restarts). Techniques such as sequence number space partitioning and non-volatile storage preservation can be used but are beyond the scope of this specification.

Authentication Data: A field with a variable length. The field carries the digest for the protocol packet and other optional information.

Type 1: This 4-bit field indicate the type of the encapsulated PIM packet.

Reserved 1: This 8-bit field is identical to the Reserved field of the encapsulated PIM packet. Because the Version field and the Checksum field in the header of the encapsulated PIM packet are redundant, they are removed.

3. PIM Security Association

An PIM Security Association (SA) consists of a set of parameters for PIM routers to correctly generate or verify AEP packets. In manual keying, it is the responsibility of network operators to generate and deploy PIM SAs amongst PIM routers appropriately to ensure the routers can exchange PIM signalling messages securely.

The parameters associated with a PIM SA:

- o **Key Identifier (Key ID) :** A 16-bit unsigned integer which is used to uniquely identify an PIM SA within a PIM domain.
- o **Authentication Algorithm:** This parameter is used to indicate the authentication algorithm to be used with the PIM SA. The value of this parameter can be implementer specific. Currently, the

following algorithms SHOULD be supported: HMAC-SHA-1, HMAC-SHA-256, HMAC-SHA-384, and HMAC-SHA-512.

- o Key: The value of this parameter denotes the cryptographic key associated with the key ID. The length of this key is determined by the algorithm specified in the PIM SA.
- o Key Start Accept: The time after which a PIM router will accept a packet if it is created with this PIM SA.
- o Key Start Generate: The time after which a PIM router will begin using this PIM SA for PIM packet generation.
- o Key Stop Generate: The time after which a PIM router will stop using this PIM SA for PIM packet generation.
- o Key Stop Accept: The time after which a PIM router will refuse to accept a packet if it is generated with this PIM SA.

4. AEP Packet Processing

4.1. Cryptographic Aspects

In the algorithm description below, the following nomenclature, which is consistent with [FIPS-198], is used:

H is the specific hashing algorithm (e.g. SHA-256).

K is the Authentication Key for the PIM security association.

Ko is the cryptographic key used with the hash algorithm.

B is the block size of H, measured in octets rather than bits.

Note that B is the internal block size, not the hash size.

For SHA-1 and SHA-256: B == 64

For SHA-384 and SHA-512: B == 128

L is the length of the hash, measured in octets rather than bits.

XOR is the exclusive-or operation.

Opad is the hexadecimal value 0x5c repeated B times.

Ipad is the hexadecimal value 0x36 repeated B times.

Apad is a value which is the same length as the hash output or message digest. If the packet is transported upon IPv6, the first 16 octets contain the IPv6 source address followed by the hexadecimal value 0x878FE1F3 repeated (L-16)/4 times. If the packet is transported upon IPv4, the first 4 octets contain the IPv4 source address followed by the hexadecimal value 0x878FE1F3 repeated (L-4)/4 times.

1. Preparation of the Key

In this application, Ko is always L octets long.

If the Authentication Key (K) is L octets long, then Ko is equal to K. If the Authentication Key (K) is more than L octets long, then Ko is set to H(K). If the Authentication Key (K) is less than L octets long, then Ko is set to the Authentication Key (K) with zeros appended to the end of the Authentication Key (K) such that Ko is L octets long.

2. First Hash

First, the AEP packet's Authentication Data field in the AEP header is filled with the value Apad.

Then, a First-Hash, also known as the inner hash, is computed as follows:

If the original packet is a Register packet

First-Hash = H(Ko XOR Ipad || (AEP Packet-Data Part))

else

First-Hash = H(Ko XOR Ipad || (AEP Packet))

The digest length for SHA-1 is 20 octets; for SHA-256, 32 octets; for SHA-384, 48 octets; and for SHA-512, 64 octets.

3. Second Hash

Then a second hash, also known as the outer hash, is computed as follows:

Second-Hash = H(Ko XOR Opad || First-Hash)

4. Result

The resulting Second-Hash becomes the authentication data that is sent in the AEP header. The length of the authentication data is always identical to the message digest size of the specific hash function H that is being used.

4.2. Outbounding Packet Processing

First of all, a sender needs to find a proper PIM SA and generate a PIM header. The checksum field of the AEP header is set as zero. The length of the Authentication Data field is determined according to the algorithm specified in the SA. The sequence number for this SA is increased, and the new value is inserted into the Sequence Number field. The Authentication Data field is set as Apad. After these, the sender appends the encapsulated PIM packet (without the redundant fields) at the end of the AEP header and generates the authentication data as illustrated in Section 4.1. After inserting the calculated authentication data into the Authentication Data field, the sender delivers the packet.

4.3. Inbounding Packet Processing

A router identifies a received PIM packet is an AEP by examining the Type field in PIM packet header. If the cryptographic sequence number of the packet is less than or equal to the last sequence number received from the PIM router, the AEP packet MUST be dropped. If the Checksum fields in the AEP header and in the PIM header of the encapsulated PIM packet are not zero, the AEP packet MUST be dropped.

According to the key ID in the packet header, the receiver tries to find the associated PIM SA. If no valid PIM SA exists for this packet or the key is not in its valid period, the receiver MUST discard the packet. If the appropriate PIM SA for the received packet is found, the receiver starts performing the authentication algorithm dependent processing, using the algorithm specified in the SA.

In the first step, the receiver derives the cryptographic algorithm from the PIM SA and identify the length of the Authentication Data field. Then the receiver fills the Authentication Data field with Apad. After this, the receiver calculate the authentication data for the AEP as described in Section 4.1. The calculated data is compared with the received authentication data in AEP header. If the two do not match, the packet MUST be discarded. In such a case, an error event SHOULD be logged.

5. Security Considerations

5.1. Register Packet Processing

The solution proposed in this draft only intends to secure PIM signaling packets. The efforts of protecting data packets transported among PIM routers are out of scope. Therefore, for a register packet, only the Type field, B field, and N field are secured while the Multicast data packet part is not protected by the authentication data.

5.2. New Packet Type Versus Authentication Trailer

Both PIM and OSPFv3 rely on IPsec to secure packet transmission, and they meet similar security issues, such as the vulnerability to the replay attacks and lack of support to priority packets. [I-D.ietf-ospf-auth-trailer-ospfv3] proposes an authentication trailer which is appended at the end of an OSPFv3 packet and provides IPsec independent authentication for the packet. This idea can also be adopted into PIM. However, compared with the OSPFv3 packet header, the PIM header lacks a field to point out the length the PIM packet. The length of the PIM packet is actually indicated by the length of the IP payload and can be variable. This raises a issue. If an authentication trailer is attached at the end of a PIM packet, it will be difficult to locate. This issue can be addressed by extending the PIM headers with an Length field.

5.3. Inter-Session Replay Attack Issue

When a router is rebooted , the sequence number will be re-initialized. This will cause a problem. When a PIM router received a hello message with a changed GenID and an re-inialized sequence number, it is difficult for the receiver to distinguish this message from a replay attack. The soltuion proposed in this document is subject to this problem. However, the experience in [I-D.ietf-ospf-security-extension-manual-keying] can be used to address this problem. In the solution proposed in [I-D.ietf-ospf-security-extension-manual-keying], there is a reboot counter maintained in non-violate memory which is increased by 1 after every reboot. The count value is set into the first 32bit of the sequence number. Therefore, even after a restart, the sequence number will still be increased.

6. Acknowledgements

We would like to thank Stig Venaas for his kindly review work and comments 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

- [I-D.bhatia-karp-pim-gap-analysis]
Bhatia, M., "Analysis of Protocol Independent Multicast Sparse Mode (PIM-SM) Security According to KARP Design Guide", draft-bhatia-karp-pim-gap-analysis-00 (work in progress), April 2011.
- [I-D.ietf-ospf-auth-trailer-ospfv3]
Bhatia, M., Manral, V., and A. Lindem, "Supporting Authentication Trailer for OSPFv3", draft-ietf-ospf-auth-trailer-ospfv3-05 (work in progress), May 2011.
- [I-D.ietf-ospf-security-extension-manual-keying]
Bhatia, M., Hartman, S., Zhang, D., and A. Lindem, "Security Extension for OSPFv2 when using Manual Key Management", draft-ietf-ospf-security-extension-manual-keying-00 (work in progress), May 2011.
- [RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, January 2005.
- [RFC4302] Kent, S., "IP Authentication Header", RFC 4302, December 2005.
- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, December 2005.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC5796] Atwood, W., Islam, S., and M. Siami, "Authentication and Confidentiality in Protocol Independent Multicast Sparse Mode (PIM-SM) Link-Local Messages", RFC 5796, March 2010.

Authors' Addresses

Manav Bhatia
Alcatel-Lucent

Email: manav.bhatia@alcatel-lucent.com

Dacheng Zhang
Huawei

Email: zhangdacheng@huawei.com

Network Working Group
Internet-Draft
Intended status: Standard Track
Expires: November 20, 2011

Michael Brig
Aegis BMD Program Office
17211 Avenue D, Suite 160
Dahlgren, VA 22448-5148
Phone: 540-663-1919
Email: michael.brig@mda.mil

Deterministic RP (D-RP) Specification
draft-brigm-deterministicrp-00.txt

Copyright (c) 2011 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.

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."

Comments are solicited and should be addressed to the working group's mailing list and/or the author(s).

Abstract

This document specifies the Deterministic Rendezvous Point (D-RP) mechanism for Protocol Independent Multicast (PIM) Sparse Mode (SM) networks. It intends to provide a simple and robust RP service. The mechanism is deterministic since it elects the highest priority candidate to be the D-RP from those available for each IP address family. If a D-RP fails, the election process begins again using the remaining C-RPs for the IP address family. If no candidates are available, the network will transition to PIM Dense Mode (DM) routing for that IP address family. In the future if C-RPs emerge for the address family, the network will elect a new D-RP and return to operations with PIM SM routing.

1. Introduction

From [2], a PIM SM multicast domain requires at least one Rendezvous Point (RP) and each RP may service one or more multicast groups. Concurrently, each multicast group can be serviced by one and only one RP. This protocol mechanism is intended for high availability, moderately sized, well managed, and tightly controlled multicast domains; therefore, only a single RP will be needed to service all multicast groups of each IP address family. If IPv4 and IPv6 multicast are simultaneously operational in a PIM SM domain running this protocol, one D-RP will service IPv4 multicast while another D-RP will service IPv6 multicast.

This mechanism provides a simple and fault tolerant RP service for IP multicast domains. For this protocol to operate effectively, all routers in the PIM domain

must utilize it and be configured either as candidate-RPs (C-RPs) or non-candidates.

The mechanism will support IPv4 multicast by itself, IPv6 multicast by itself, or IPv4 and IPv6 multicast operating simultaneously on the same infrastructure but distinct from one other. The C-RP sets for IPv4 and IPv6 should, therefore, be distinct and not intersect. In the later case, there would be at most one D-RP for each IP address family at any time.

This mechanism is built upon reference [1] and reference [2] for PIM ver.2. It is specifically not intended to operate with PIM ver.1.

2. Protocol Specification

During the D-RP election process, C-RPs periodically flood the PIM domain with PIM type 8 "Candidate-RP-Advertisement" messages declaring their candidacy for D-RP of the IP address family. The protocol will elect the candidate with the highest priority to D-RP from the available C-RPs and any existing D-RP of the IP address family. After election, the D-RP will periodically flood the network with a new PIM ver.2 type 11 "elected-RP" message for the duration of its operation as D-RP. If the D-RP fails, the election process begins again using the remaining C-RPs for the IP address family. If no candidates are available, the network will transition to PIM Dense Mode (DM) routing for that IP address family. In the future if C-RPs emerge for the address family, the network will elect a new D-RP and return to PIM SM routing.

Alternately, the PIM ver.2 type 8 "Candidate-RP-Advertisement" message defined in [3] could be modified by using a single bit from its reserve field as the "Elected" (E) bit. When E = 0, the Candidate/Elected-RP (C/E-RP) message would be a candidate-RP advertisement for that IP address family. When E = 1, the (C/E-RP) message would be an elected-RP advertisement for that IP address family.

Each IP address family will have 10 integer priority values ranging from 1 to 10 for the network administrator to assign relative importance to C-RPs. It is believed that 10 C-RPs per IP address family represents the largest practical set of C-RPs which a PIM ver.2 network may require. The greater the priority value of the C-RP, the greater its relative importance to the network. These values shall fill the priority fields of Candidate-RP-Advertisement, Elected-RP-Advertisement, and Candidate/Elected-RP-Advertisement messages when transmitted in the multicast domain.

When a Elected-RP-Advertisement message has a Holdtime = 0 or a Candidate/Elected-RP-Advertisement message with E = 1 has a Holdtime = 0, the E-RP Valid Time and Timer shall be considered infinite.

2.1 State Transitions for PIM ver.2 Routers configured as Candidate RPs.

On startup, Candidates enter C-RP state after transmitting a C-RP message, setting the C-RP Xmit Timer, C-RP Valid Timer, and RP Election Timer.

When in Active C-RP state				
Event	RP Election Expires	Rcvd E-RP Message with lower priority than candidate.	C-RP Valid Timer Timeout.	Rcvd C-RP message with higher priority than candidate.
	->	->	->	->

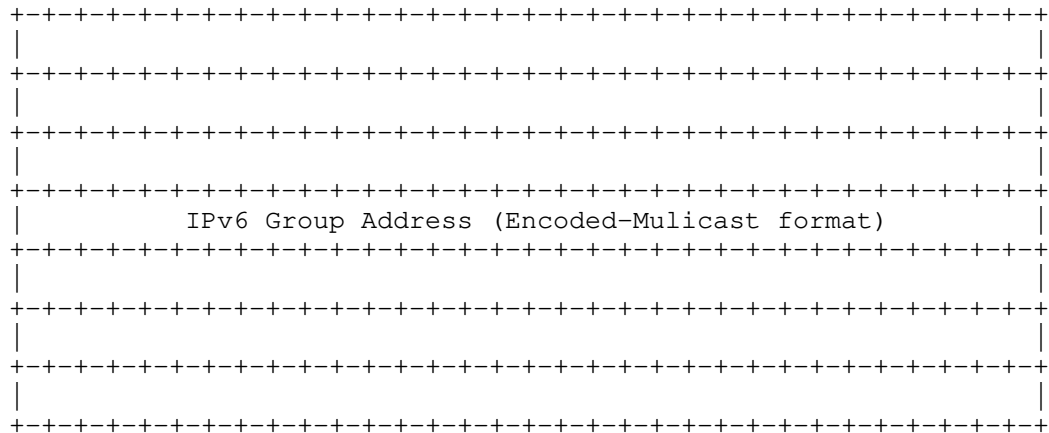
Action	E-RP; Xmit E-RP message, Set E-RP Xmit Timer, Set E-RP valid Timer, set RP Election Timer, set E-RP Valid Timer	E-RP; Xmit E-RP message, Set E-RP Xmit Timer, Set E-RP valid Timer. set RP Election Timer	Standby C-RP; Set RP Alive Timer.	Standby C-RP; Set RP Alive Timer.
--------	--	--	---	---

When in Active C-RP state (continued)		
Event	Rcvd E-RP message with higher priority than candidate.	C-RP Xmit Timeout
Action	-> Standby C-RP; Set RP Alive Timer.	-> Active C-RP; Xmit C-RP message; Set C-RP Xmit Timer.

When in E-RP state			
Event	rcvd C-RP message is higher priority than the priority of the Elected RP	E-RP Valid Timer Timeout	E-RP Xmit Timeout
Action	-> Standby C-RP; Set RP Alive Timer.	-> Active C-RP; Xmit C-RP message, Set C-RP Xmit Timer. Set C-RP Valid Timer.	-> E-RP; Xmit E-RP message; Set E-RP Xmit Timer.

When in Standby C-RP state		
Event	RP Alive Timer expires	rcvd Elected RP message
Action	-> Active C-RP; Xmit C-RP message, Set C-RP Valid Timer, Set C-RP Xmit Timer, set RP Election Timer.	-> Standby C-RP; Set RP Alive Timer.

2.2 State Transition Diagrams for PIM ver.2 Routers configured as non-candidates.



3.2 Elected-RP-Advertisement Message (type = 11) proposed for IPv6.

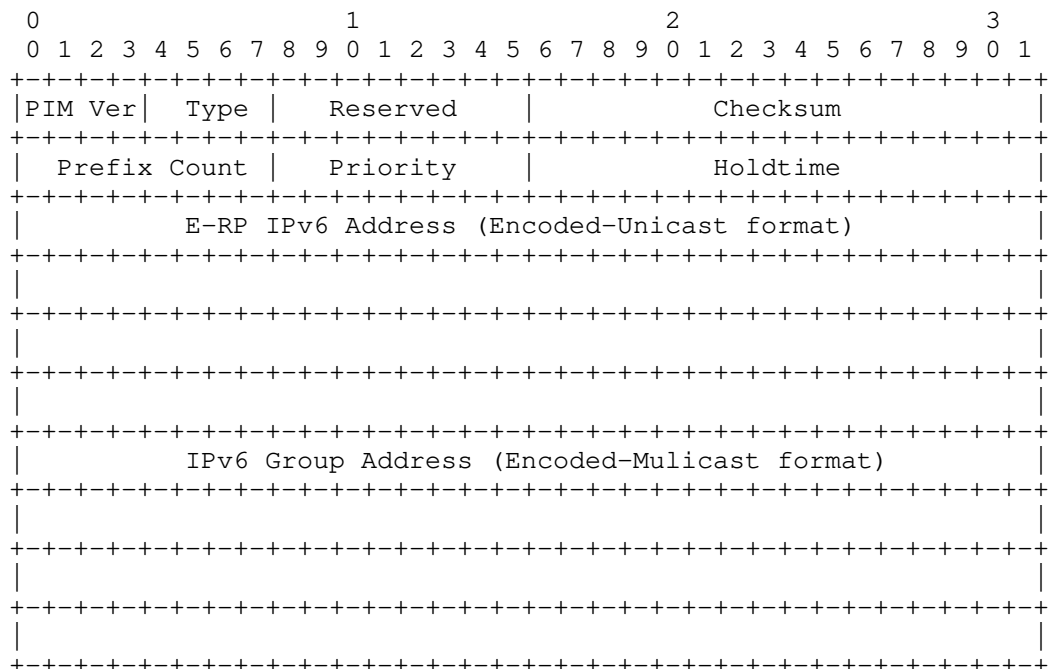
This mechanism could utilize a new PIM ver.2 type 11 "elected-RP-Advertisement" message. It is a means of determining Group to RP mappings. The message would be defined identically to the "candidate-RP-Advertisement" defined in [3] with the exception that the type field would be set to 11. This message would flood the PIM domain periodically to announce the D-RP. Only the D-RP should utilize this message at any time. Both IPv4 multicast and IPv6 multicast are supported by This message but only IPv6 is illustrated for the sake of brevity.

PIM VER = 2

Type = 11

Prefix Count = 1

IPv6 Group Address is the entire IPv6 multicast address range.



3.3 Candidate/Elected-RP-Advertisement Message (type=8) proposed for IPv6.

As an alternative to the PIM type 11 "elected-RP-advertisement" message, this mechanism could utilize a modified PIM ver.2 type 8 "candidate-RP-advertisement" message renamed the "candidate/elected-RP-Advertisement" message. It is a means of determining Group to RP mappings. This would be defined identically to the

"candidate-RP-Advertisement" defined in [3] with the exception of a new one bit "elected" field taken from the reserve bits. This message would flood the PIM domain periodically to announce the D-RP or candidate RPs. Only one PIM router, the D-RP, should utilize this message with the "elected" bit set to 1 while many PIM Routers could utilize this message with the "elected" bit set to 0. Both IPv4 multicast and IPv6 multicast are supported by This message but only IPv6 is illustrated for the sake of brevity.

PIM VER = 2

Type = 8

Prefix Count = 1

E = 0; Candidate-RP Message

E = 1; Elected-RP Message

IPv6 Group Address is the entire IPv6 multicast address range.

0										1										2										3																			
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1																		
PIM Ver										Type										E										Reserved										Checksum									
Prefix Count										Priority										Holdtime																													
C/E-RP IPv6 Address (Encoded-Unicast format)																																																	
IPv6 Group Address (Encoded-Mulicast format)																																																	

5 State Information and Timers

5.1 RP Election Timer - default of 4 seconds.

5.2 C-RP Transmit Timer - default of 1 second.

5.3 C-RP Valid Timer - set to a configured value and transmitted in the Candidate-RP-Advertisement message or the Candidate/Elected-RP-Advertisement message.

5.4 E-RP Transmit Timer - default of 1 second.

5.5 E-RP Valid Timer - set to a configured value and transmitted in the Elected-RP-Advertisement message or the Candidate/Elected-RP-Advertisement message.

5.6 RP Alive Timer - default of 5 seconds.

5.7 Transcient Timer - default of 3 seconds.

6 Security Considerations

Since D-RP is specifically designed to provide a reliable and fault-tolerant RP

service for PIM SM multicast networks, it is vulnerable to the security considerations and mitigations outlined in [5] and [6] while a D-RP is operational. D-RP is vulnerable to routers masquerading as C-RPs and D-RPs with and without high configured priority values. It is vulnerable to denial of service if an attacker could sufficiently flood the IP multicast domain with data and therefore prevent the majority of the PIM routers from receiving timely C-RP and D-RP messages.

When an D-RP cannot be elected, this mechanism falls back to PIM DM operations until a C-RP becomes available, and a new D-RP is elected. While in DM, it is vulnerable to the security considerations and mitigations outlined in [1].

7 Contributors

LCDR Charles Schlise
AEGIS BMD B33
540-663-1763
charles.schlise@mda.mil

Jeff Chaney
AEGIS BMD B33C
540-663-1790
Jeff.chaney@mda.mil

Thomas Tharp
IO Technologies
540-663-1865
thomas.tharp.ctr@mda.mil

8 References

- [1] Adams, A., Nicholas, J., Siadak, W.,
"Protocol Independent Multicast - Dense Mode (PIM-DM):
Protocol Specification (Revised)", RFC 3973, January 2005
- [2] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas,
"Protocol Independent Multicast - Sparse Mode (PIM-SM):
Protocol Specification (Revised)", RFC 4601, August 2006.
- [3] Bhaskar, N., Gall, A., Lingard, J., and S. Venaas,
"Bootstrap Router (BSR) Mechanism for Protocol Independent
Multicast (PIM)", RFC 5059, January 2008.
- [4] Venaas, S.,
"A Registry for PIM Message Types", RFC 6166, April 2011
- [5] Savola, P., Lehtonen, R., Meyer, D.
"Protocol Independent Multicast - Sparse Mode (PIM-SM)
Multicast Routing Security Issues and Enhancements",
RFC 4609, August 2006.
- [6] Atwood, W., Islam, S., Siami, M., "Authentication and
Confidentiality in Protocol Independent Multicast Sparse
Mode (PIM-SM) Link-Local Messages", RFC 4601, March 2010

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 9, 2012

Yiqun Cai
Liming Wei
Heidi Ou
Cisco Systems, Inc.
Vishal Arya
Sunil Jethwani
DIRECTV Inc.
July 8, 2011

Protocol Independent Multicast ECMP Assert
draft-hou-pim-ecmp-01.txt

Abstract

A PIM router uses RPF procedure to select an upstream interface and router to build forwarding state. When there are equal cost multiple paths (ECMP), existing implementations often use hash algorithms to select a path. Such algorithms do not allow the spread of traffic among the ECMPs according to administrative metrics. This usually leads to inefficient or ineffective use of network resources. This document introduces the ECMP Assert, a mechanism to improve the RPF procedure over ECMPs. It allows ECMP path selection to be based on administratively selected metrics, such as data transmission delays, path preferences and routing metrics.

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 9, 2012.

Copyright Notice

Copyright (c) 2011 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. Requirements Notation	3
2. Introduction	3
2.1. Overview	3
2.2. Applicability	4
3. Protocol Specification	5
3.1. ECMP Bundle	5
3.2. Sending ECMP Assert	5
3.3. Receiving ECMP Assert	6
3.4. Transient State	6
3.5. Interoperability	7
3.6. Packet Format	7
3.6.1. PIM ECMP Assert Hello Option	7
3.6.2. PIM ECMP Assert Format	8
4. IANA Considerations	9
5. Security Considerations	9
6. Acknowledgement	9
7. References	9
7.1. Normative Reference	9
7.2. Informative References	10
Authors' Addresses	10

1. Requirements Notation

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. Introduction

A PIM [RFC4601] router uses RPF procedure to select an upstream interface and a PIM neighbor on that interface to build forwarding state. When there are equal cost multiple paths (ECMP) upstream, existing implementations often use hash algorithms to select a path. Such algorithms do not allow the spread of traffic among the ECMP according to administrative metrics. This usually leads to inefficient or ineffective use of network resources. This document introduces the ECMP Assert, a mechanism to improve the RPF procedure over ECMP. It allows ECMP path selection to be based on administratively selected metrics, such as data transmission delays, path preferences and routing metrics, or a combination of metrics.

ECMPs are frequently used in networks to provide redundancy and to increase available bandwidth. A PIM router selects a path in the ECMP based on its own implementation specific choice. The selection is a local decision. One way is to choose the PIM neighbor with the highest IP address, another is to pick the PIM neighbor with the best hash value over the destination and source addresses.

While implementations supporting ECMP have been deployed widely, the existing RPF selection methods have weaknesses. The lack of administratively effective ways to allocate traffic over alternative paths is a major issue. For example, there is no straightforward way to tell two downstream routers to select either the same or different RPF neighbor routers for the same traffic flows.

With the ECMP Assert mechanism introduced here, the upstream routers use a new PIM ECMP Assert message to instruct the downstream routers on how to tie-break among the upstream neighbors. The PIM ECMP Assert message conveys the tie-break information based on metrics selected administratively.

2.1. Overview

The existing PIM Assert mechanism allows the upstream router to detect the existence of multiple forwarders for the same multicast flow onto the same downstream interface. The upstream router sends a PIM Assert message containing a routing metric for the downstream routers to use for tie-breaking among the multiple upstream

forwarders on the same RPF interface.

With ECMP interfaces between the downstream and upstream routers, the PIM ECMP Assert mechanism works in a similar way, but extends the ability to resolve the selection of forwarders among different interfaces in the ECMP.

When a PIM router downstream of the ECMP interfaces creates a new (*,G) or (S,G) entry, it will populate the RPF interface and RPF neighbor information according to the rules specified by [RFC4601]. This router will send its initial joins to that RPF neighbor.

When the RPF neighbor router receives the join message and finds that the receiving interface is one of the ECMP interfaces, it will check if the same flow is already being forwarded out of another ECMP interface. If so, this RPF neighbor router will send a PIM ECMP Assert message onto the interface the join was received on. The PIM ECMP Assert message contains the address of the desired RPF neighbor, an interface ID [INTID], along with other parameters used as tie breakers. In essence, a PIM ECMP Assert message is sent by an upstream router to notify downstream routers to redirect PIM Joins to the new RPF neighbor via a different interface. When the downstream routers receive this message, they should trigger PIM Joins toward the new RPF neighbor specified in the packet.

This new message is named PIM ECMP Assert for the following reasons,

1. It is sent by an upstream router;
2. It is used to influence the RPF selection by downstream routers;
And
3. A tie breaker metric is used.

This new message functions in similar ways to the existing PIM Assert message, with the exception that the existing Assert message is used to select an upstream router within the same multi-access network (such as a LAN) while the new message is used to select both a network and an upstream router.

One advantage of this design is that the control messages are only sent when there is need to "re-balance" the traffic. This reduces the amount of control traffic.

2.2. Applicability

The use of ECMP Assert applies to shared trees or source trees built with procedures described in [RFC4601]. The use of ECMP Assert in "Protocol Independent Multicast - Dense Mode" [RFC3973] or in "Bidirectional Protocol Independent Multicast" [RFC5015] is not

considered.

The enhancement described in this document can be applicable to a number of scenarios. For example, it allows a network operator to use ECMP paths and have the ability to perform load splitting based on bandwidth. To do this, the downstream routers perform RPF selection with bandwidth instead of IP addresses as a tie breaker. The ECMP Assert mechanism assures that all downstream routers select the desired network link and upstream router whenever possible. Another example is for a network operator to impose a transmission delay limit on certain links. The ECMP Assert mechanism provides a mean for an upstream router to instruct a downstream router to choose a different RPF path.

This specification does not dictate the scope of applications of this mechanism.

3. Protocol Specification

3.1. ECMP Bundle

An ECMP bundle is a set of PIM enabled interfaces on a router, where all interfaces belonging to the same bundle share the same routing metric. The ECMP paths reside between the upstream and downstream routers over the ECMP bundle.

There can be one or more ECMP bundles on any router, while one individual interface can only belong to a single bundle.

ECMP bundles are created on a router via configuration.

3.2. Sending ECMP Assert

ECMP Asserts are sent by an upstream router in a rate limited fashion, under the following conditions,

- o It detects a PIM Join on a non-desired outgoing interface; or
- o It detects multicast traffic on a non-desired outgoing interface.

In both cases, an ECMP Assert is sent to the non-desired interface. An outgoing interface is considered "non-desired" when,

- o The upstream router is already forwarding the same flow out of another interface belonging to the same ECMP bundle;
- o The upstream router is not forwarding the flow yet out any interfaces of the ECMP bundle, but there is another interface with more desired attributes.

An upstream router may choose not to send ECMP Asserts if it becomes aware that some of the downstream routers do not support the new message, or unreachable via some links in ECMP bundle.

3.3. Receiving ECMP Assert

When a downstream router receives an ECMP Assert, and detects the desired RPF path from its upstream router's point of view is different from its current one, it should choose to prune from the current path and join to the new path. The exact order of such actions is implementation specific.

If a downstream router receives multiple ECMP Asserts sent by different upstream routers, it SHOULD use the Preference, Metric, or other fields as specified below, as the tie breakers to choose the most preferred RPF interface and neighbor.

If an upstream router receives an ECMP Assert from another upstream router, it SHOULD NOT change its forwarding behavior even if the ECMP Assert makes it a less preferred RPF neighbor on the receiving interface.

3.4. Transient State

During a transient network outage with a single link cut in an ECMP bundle, a downstream router may lose connection to its RPF neighbor and the normal ECMP Assert operation may be interrupted temporarily. In such an event, the following actions are recommended.

The down stream router may re-select a new RPF neighbor. Among all ECMP upstream routers, the one on the same LAN as the previous RPF neighbor is preferred.

If there is no upstream router reachable on the same LAN, the down stream router will select a RPF neighbor on a different LAN. Among all ECMP upstream routers, the one served as RPF neighbor before the link failure is preferred. Such a router can be identified by the Router ID which is part of the Interface ID in the PIM ECMP Assert Hello option.

During normal ECMP Assert operations, when PIM Joins for the same (*,G) or (S,G) are received on a different LAN, an upstream router will send ECMP Assert to prune the non-preferred LAN. Such ECMP Asserts during partial network outage can be suppressed if the upstream router decides that the non-preferred PIM Join is from a router that is not reachable via the preferred LAN. This check can be performed by retrieving the downstream's Router ID, using the source address in the PIM join, and searching neighbors on the

preferred LAN for one with the same router ID.

3.5. Interoperability

If a PIM router supports this draft, it MUST send the new Hello option ECMP-Assert-Supported TLV in its PIM Hello messages. A PIM router sends ECMP Asserts on an interface only when it detects that all neighbors have sent this Hello option. If a PIM router detects that any of its neighbor does not support this Hello option, it MUST not send ECMP Asserts, however, it SHOULD still process any ECMP Asserts received.

3.6. Packet Format

3.6.1. PIM ECMP Assert Hello Option

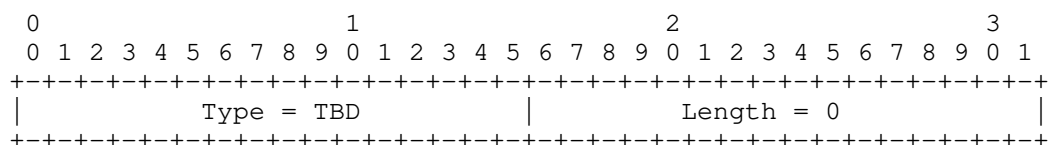


Figure 1: ECMP Assert Hello Option

Type: TBD.
Length: 0

3.6.2. PIM ECMP Assert Format

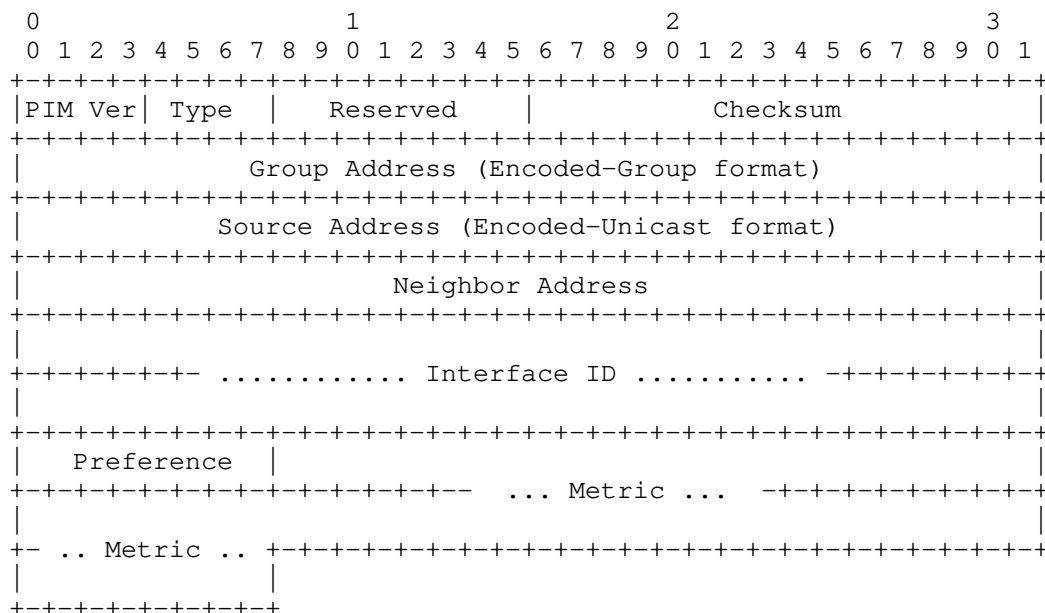


Figure 2: ECMP Assert Message Format

Type: TBD

Neighbor Address (32/128 bits): Address of desired upstream neighbor where the downstream receiver should redirect PIM Joins to. This address MUST be associated with an interface in the same ECMP bundle as the ECMP Assert message's outgoing interface. If the "Interface ID" field (see below) is ignored, this "Neighbor Address" field uniquely identifies a LAN and an upstream router to which a downstream router should redirect its Join messages to, and an ECMP Assert message MUST be discarded if the "Neighbor Address" field in the message does not match cached neighbor address.

Interface ID (64 bits): This field is used in IPv4 when one or more RPF neighbors in the ECMP bundle are unnumbered, or in IPv6 where link local addresses are in use. For other IPv4 usage, this field is zero'ed when sent, and ignored when received. If the "Router ID" part of the "Interface ID" is zero, the field must be ignored. See [INTID] for details of its assignment and usage in PIM Hellos. If the "Interface ID" is not ignored, the receiving router of this message MUST use the "Interface ID", instead of "Neighbor

Address", to identify the new RPF neighbor, and an ECMP Assert message MUST be discarded if the "Interface ID" field in the message does not match cached interface ID.

Preference (8 bits): The first tie breaker when ECMP Asserts from multiple upstream routers are compared against each other. Numerically smaller value is preferred. A reserved (15) value is used to indicate the metric value following the "Preference" field is a timestamp, taken at the moment the sending router started to forward out of this interface.

Metric (64 bits): The second tie breaker if the the "Preference" values are the same. Numerically smaller metric is preferred. This "Metric" can contain path parameters defined by users. When both "Preference" and "Metric" values are the same, "Neighbor Address" or "Interface ID" field is used as the third tie-breaker, depends on which field is used to identify the RPF neighbor, and the bigger value wins.

4. IANA Considerations

A new PIM Type is required to be assigned to the ECMP Assert messages. According to [PIMREG], this document recommends 11 (0xB) as the new "PIM ECMP Assert Type".

5. Security Considerations

Security of the ECMP Assert is only guaranteed by the security of the PIM packet, so the security considerations for PIM Assert packets as described in [RFC4601] apply here. Spoofed ECMP Assert packets may cause the downstream routers to send PIM Joins to an undesired upstream router, and trigger more ECMP Assert messages.

6. Acknowledgement

The authors would like to thank Apoorva Karan for helping with the original idea, Eric Rosen, Isidor Kouvelas, Toerless Eckert and Stig Venaas for their review comments.

7. References

7.1. Normative Reference

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

- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

7.2. Informative References

- [RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, January 2005.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.
- [INTID] Gulrajani, S. and S. Venaas, "An Interface ID Hello Option for PIM", draft-gulrajani-pim-hello-intid-01.txt (work in progress).
- [PIMREG] Venaas, S., "A Registry for PIM Message Types", draft-ietf-pim-registry-04.txt (work in progress).

Authors' Addresses

Yiqun Cai
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: ycai@cisco.com

Liming Wei
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: lwei@cisco.com

Heidi Ou
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: hou@cisco.com

Vishal Arya
DIRECTV Inc.
2230 E Imperial Hwy
El Segundo, CA 90245
USA

Email: varya@directv.com

Sunil Jethwani
DIRECTV Inc.
2230 E Imperial Hwy
El Segundo, CA 90245
USA

Email: sjethwani@directv.com

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 7, 2012

Dino Farinacci
Greg Shepherd
Yiqun Cai
Stig Venaas
cisco Systems
July 6, 2011

Population Count Extensions to PIM
draft-ietf-pim-pop-count-04.txt

Abstract

This specification defines a method for providing multicast distribution-tree accounting data. Simple extensions to the PIM protocol allow a rough approximation of tree-based data in a scalable fashion.

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 7, 2012.

Copyright Notice

Copyright (c) 2011 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. Requirements Notation	3
2. Introduction	4
2.1. Terminology	4
3. New Hello TLV Pop-Count Support	5
4. New Pop-Count Join Attribute Format	6
4.1. Options	9
4.1.1. Link Speed Encoding	10
4.2. Example message layouts	11
5. How to use Pop-Count Encoding	13
6. Implementation Approaches	14
7. Caveats	15
8. IANA Considerations	16
9. Security Considerations	17
10. Acknowledgments	18
11. References	19
11.1. Normative References	19
11.2. Informative References	19
Authors' Addresses	20

1. Requirements Notation

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. Introduction

This draft proposes a mechanism to convey accounting information using the PIM protocol [RFC4601] [RFC5015]. Putting the mechanism in PIM allows efficient distribution and maintenance of such accounting information. Previous mechanisms require data to be correlated from multiple router sources.

This proposal allows a single router to be queried to obtain accounting and statistic information for a multicast distribution tree as a whole or any distribution sub-tree downstream from a queried router. The amount of information is fixed and does not increase as multicast membership, tree diameter, or branching increase.

The sort of accounting data this draft provides, on a per multicast route basis, are:

1. The number of branches in a distribution tree.
2. The membership type of the distribution tree, that is SSM or ASM.
3. Routing domain and time zone boundary information.
4. On-tree node and tree diameter counters.
5. Effective MTU and bandwidth.

This draft adds a new PIM Join Attribute type [RFC5384] to the Join/Prune message as well as a new Hello TLV. The mechanism is applicable to IPv4 and IPv6 multicast.

2.1. Terminology

This section defines the terms used in this draft.

Multicast Route: A (S,G) or (*,G) entry regardless if the route is in ASM, SSM, or Bidir mode of operation.

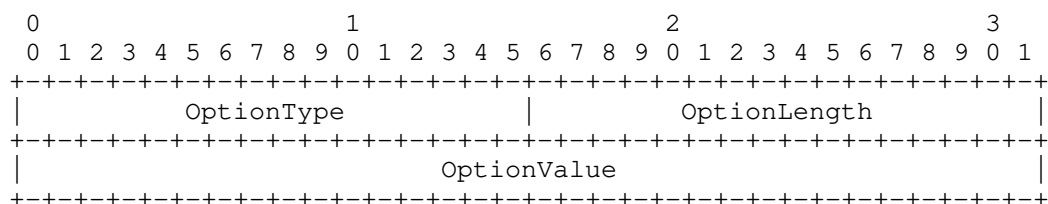
Stub Link: A link with members joined to the group via IGMP or MLD.

Transit Link: A link put in the oif-list for a multicast route because it was joined by PIM routers.

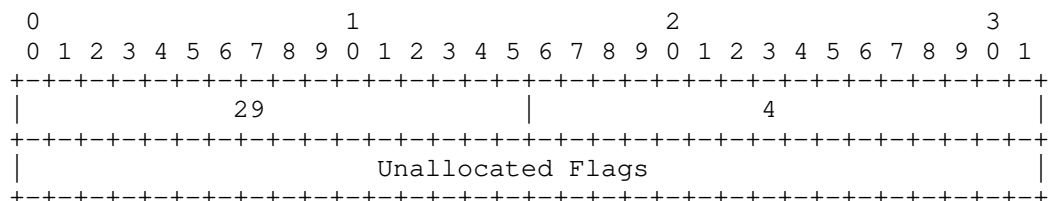
Note that a link can be both a Stub Link and a Transit Link at the same time.

3. New Hello TLV Pop-Count Support

When a PIM router sends a Join/Prune message to a neighbor, it will encode the data in a new PIM Join Attribute type (described in this draft) when the PIM router determines the neighbor can support this draft. If a PIM router supports this draft, it must send the Pop-Count-Supported TLV. The format of the TLV is defined to be:



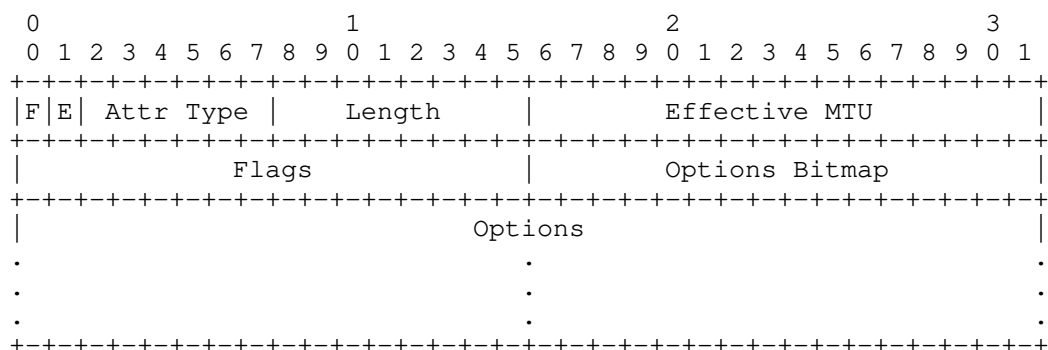
OptionType = 29, OptionLength = 4, there is no OptionValue semantics defined at this time but will be included for expandability and be defined in future revisions of this draft. The format will look like:



Unallocated Flags: for now should be sent as 0 and ignored on receipt.

4. New Pop-Count Join Attribute Format

When a PIM router supports this draft and has determined from a received Hello, the neighbor supports this draft, it will send Join/Prune messages that MAY include a Pop-Count attribute. The mechanism to process PIM Join Attribute is described in [RFC5384]. The format of the new attribute is described in the following.



The above format is used only for entries in the join-list section of the Join/Prune message.

F bit: 0 Non-Transitive Attribute.

E bit: As specified by [RFC5384].

Attr Type: 2.

Length: The minimum length is 6.

Effective MTU: This contains the minimum MTU for any link in the oif-list. The sender of Join/Prune message takes the minimum value for the MTU (in bytes) from each link in the oif-list. If this value is less than the value stored for the multicast route (the one received from downstream joiners) then the value should be reset and sent in Join/Prune message. Otherwise, the value should remain unchanged.

This provides one to obtain the MTU supported by multicast distribution tree when examined at the first-hop router(s) or for sub-tree for any router on the distribution tree.

Flags: The flags field has the following format:

```

      0                               1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  | Unalloc/Reserved | P | a | t | A | S |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

```

Unallocated Flags: The flags which are currently not defined.

If a new flag is defined and sent by a new implementation, an old implementation should preserve the bit settings. This means that if a bit was set in a PIM Join message from any of the downstream routers, then it MUST also be set in any PIM Join sent upstream.

S flag: If an IGMPv3 or MLDv2 report was received on any oif-list entry or the bit was set from any PIM Join message. This bit should only be cleared when the above becomes untrue.

A flag: If an IGMPv1, IGMPv2, or MLDv1 report was received on any oif-list entry or the bit was set from any PIM Join message. This bit should only be cleared when the above becomes untrue.

A combination of settings for these bits indicate:

A-flag	S-flag	Description
0	0	There are no members for the group ('Stub Oif-List Count' is 0)
0	1	All group members are only SSM capable
1	0	All group members are only ASM capable
1	1	There is a mixture of SSM and ASM capable

t flag: If there are any tunnels on the distribution tree. If a tunnel is in the oif-list, a router should set this bit in its Join/Prune messages. Otherwise, it propagates the bit setting from downstream joiners.

a flag: If there are any auto-tunnels on the distribution tree. If an auto-tunnel is in the oif-list, a router should set this bit in its Join/Prune messages. Otherwise, it propagates the bit setting from downstream joiners. An example of an auto-tunnel is an tunnel setup by the AMT [AMT] protocol.

P flag: This flag remains set if all downstream routers support this specification. That is, they are PIM pop-count capable. This allows one to tell if the entire sub-tree is completely accounting capable.

Options Bitmap: This is a bitmap that shows which options are present. The format of the bitmap is as follows:

```

      0                               1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+---+---+---+---+---+---+---+---+---+
|T|s|m|M|d|n|D|z| Unalloc/Rsrvd |
+---+---+---+---+---+---+---+---+

```

Each one of the bits T, s, m, M, d, n, D and z is associated with one option, where the option is included if and only if the respective bit is set. Included options MUST be in the same order as these bits are listed. The bits denote the following options:

bit	Option
T	Transit Oif-List Count
s	Stub Oif-List Count
m	Minimum Speed Link
M	Maximum Speed Link
d	Domain Count
n	Node Count
D	Diameter Count
z	TZ Count

See Section 4.1 for details on the different options. The unallocated bits are reserved. Any unknown bits MUST be set to 0 when a message is sent, and treated as 0 (ignored) when received. This means that unknown options which are denoted by unknown bits are ignored.

By using this bitmap we can specify at most 16 options. If there becomes a need for more than 16 options, one can define a new option that contains a bitmap, which can then be used to specify which further options are present. The last bit in the current bitmap could be used for that option. The exact definition of this is however left for future documents.

Options: This field contains options. Which options are present are determined by the flag bits. As new flags and options may be defined in the future, any unknown/reserved flags MUST be ignored, and any additional trailing options MUST be ignored. See Section 4.1 for details on the options defined in this document.

4.1. Options

There are several options defined in this document. For each option, there is also a related flag that shows whether the option is present. See the Options Bitmap above for a list of the options and their respective bits. Each option has a fixed size.

Transit Oif-List Count: This is filled in by a router sending a Join/Prune message which is equal to the number of oifs for the multicast route that has been joined by PIM. This indicates the transit branches on a multicast distribution tree (no members on the links between this router and joining routers). This is added to the value advertised by all downstream PIM routers that have joined on this oif. Length 2 octets.

Stub Oif-List Count: This is filled in by a router sending a Join/Prune message which is equal to the number of oifs for the multicast route that has been joined by IGMP or MLD. This indicates the links where there are host members for the multicast route. This is added to the value advertised by all downstream PIM routers that have joined on this oif. Length 2 octets.

Minimum Speed Link: This contains the minimum bandwidth rate for any link in the oif-list and is encoded as specified in Section 4.1.1. The sender of Join/Prune message takes the minimum value for each link in the oif-list for the multicast route. If this value is less than the value stored for the multicast route (the one received from downstream joiners) then the value should be reset and sent in Join/Prune message. Otherwise, the value should remain unchanged. This together with the Maximum Speed Link option provides a way to obtain the lowest and highest speed link for the multicast distribution tree. Length 2 octets.

Maximum Speed Link: This contains the maximum bandwidth rate for any link in the oif-list and is encoded as specified in Section 4.1.1. The sender of Join/Prune message takes the maximum value for each link in the oif-list for the multicast route. If this value is greater than the value stored for the multicast route (the one received from downstream joiners) then the value should be reset and sent in Join/Prune message. Otherwise, the value should remain unchanged. This together with the Minimum Speed Link option provides a way to obtain the lowest and highest

speed link for the multicast distribution tree. Length 2 octets.

Domain Count: This indicates the number of routing domains the distribution tree traverses. A router should increment this value if it is sending a Join/Prune message over a link which traverses a domain boundary. Length 1 octet.

Node Count: This indicates the number of routers on the distribution tree. Each router will sum up all the Node Counts from all joiners on all oifs and increment by 1 before including this value in the Join/Prune message. Length 1 octet.

Diameter Count: This indicates the longest length of any given branch of the tree in router hops. Each router that sends a Join increments the max value received by all downstream joiners by 1. Length 1 octet.

TZ Count: This indicates the number of timezones the distribution tree traverses. A router should increment this value if it is sending a Join/Prune message over a link which traverses a time zone. This can be a configured link attribute or use other means to determine the timezone is acceptable. Length 1 octet.

4.1.1. Link Speed Encoding

The speed is encoded using 2 octets as follows:

```

      0                               1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
| Exponent |      Significand      |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+

```

Using this format, the speed of the link is Significand * 10 ^ Exponent kbps. This allows specifying link speeds with up to 3 decimal digits precision and speeds from 1 kbps to 10 ^ 67 kbps. A computed speed of 0 kbps means the link speed is < 1 kbps.

Here are some examples how this is used:

Link Speed	Exponent	Significand
500 kbps	0	500
500 kbps	2	5
155 Mbps	3	155
40 Gpbs	6	40
100 Gpbs	6	100
100 Gpbs	8	1

4.2. Example message layouts

We will here give a few examples to illustrate the use of flags and options.

A minimum size message has no option flags set, and looks like this:

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|F|E| Attr Type | Length = 6 | Effective MTU |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Unalloc/Reserved | P|a|t|A|S|0|0|0|0|0|0|0|0|0|0| Unalloc/Rsrvd |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

A message containing all the options defined in this document would look like this:

```

<figure>
<preamble></preamble>
<artwork><![CDATA[
0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|F|E| Attr Type | Length = 18 | Effective MTU |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Unalloc/Reserved | P|a|t|A|S|1|1|1|1|1|1|1|1|1|1| Unalloc/Rsrvd |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Transit Oif-List Count | Stub Oif-List Count |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Minimum Speed Link | Maximum Speed Link |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Domain Count | Node Count | Diameter Count | TZ Count |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

A message containing only Stub Oif-List Count and Node Count would look like this:

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|F|E| Attr Type | Length = 9 | Effective MTU |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Unalloc/Reserved | P|a|t|A|S|0|1|0|0|0|1|0|0| Unalloc/Rsrvd |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Stub Oif-List Count | Node count |
+-----+-----+-----+-----+-----+-----+-----+

```

5. How to use Pop-Count Encoding

A router supporting this draft MUST include PIM Join Attribute TLV in its PIM Hellos. See [RFC5384] and [HELLO] for details.

It is very important to note that any changes to the values maintained in this draft MUST NOT trigger a new Join/Prune message. Due to the periodic nature of PIM, the values can be accurately obtained at 1 minute intervals (or whatever Join/Prune interval used).

When a router removes a link from an oif-list, it must be able to reevaluate the values that it will advertise upstream. This happens when an oif-list entry is timed out or a Prune is received.

It is recommended that the Join Attribute defined in this draft be used for entries in the join-list part of the Join/Prune message. If the new encoding is used in the prune-list or an Assert message, an implementation must ignore them but still process the Prune as if it was in the original encoding described in [RFC4601].

It is also recommended that join suppression be disabled on a LAN when Pop-Count is used.

6. Implementation Approaches

An implementation can decide how the accounting attributes are maintained. The values can be stored as part of the multicast route data structure by combining the local information it has with the joined information on a per oif basis. So when it is time to send a Join/Prune message, the values stored in the multicast route can be copied to the message.

Or, an implementation could store the accounting values per oif and when a Join/Prune message is sent, it can combine the oifs with its local information. Then the combined information can be copied to the message.

When a downstream joiner stops joining, accounting values cached must be evaluated. There are two approaches which can be taken. One is to keep values learned from each joiner so when the joiner goes away the count/max/min values are known and the combined value can be adjusted. The other approach is to set the value to 0 for the oif, and then start accumulating new values as subsequent Joins are received.

The same issue arises when an oif is removed from the oif-list. Keeping per-oif values allows you to adjust the per-route values when an oif goes away. Or, alternatively, a delay for reporting the new set a values from the route can occur while all oif values are zeroed (where accumulation of new values from subsequent Joins cause re-population of values and a new max/min/ count can be reevaluated for the route).

It is recommended that when triggered Join/Prune messages are sent by a downstream router, that the accounting information not be included in the message. This way when convergence is important, avoiding the processing time to build an accounting record in a downstream router and processing time to parse the message in the upstream router will help reduce convergence time. An upstream router should not interpret a Join/Prune message received with no accounting data to mean clearing or resetting what accounting data it has cached.

7. Caveats

This draft requires each router on a multicast distribution tree to support this draft or else the accounting attributes for the tree will not be known.

However, if there are a contiguous set of routers downstream in the distribution tree, they can maintain accounting information for the sub-tree.

If there are a set of contiguous routers supporting this draft upstream on the multicast distribution tree, accounting information will be available but it will not represent an accurate assessment of the entire tree. Also, it will not be clear for how much of the distribution tree the accounting information covers.

8. IANA Considerations

A new PIM Hello Option type, 29, has been assigned. See [HELLO] for details.

A new PIM Join Attribute type needs to be assigned. 2 is proposed in this draft.

9. Security Considerations

There are no security considerations for this design other than what is already in the main PIM specification [RFC4601].

10. Acknowledgments

The authors would like to thank John Zwiebel, Amit Jain, and Clayton Wagar for their review comments on the initial versions of this draft. Further review and comments were provided by Thomas Morin and Zhaohui (Jeffrey) Zhang.

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.
- [RFC5384] Boers, A., Wijnands, I., and E. Rosen, "The Protocol Independent Multicast (PIM) Join Attribute Format", RFC 5384, November 2008.

11.2. Informative References

- [AMT] Thaler, D., Talwar, M., Aggarwal, A., Vicisano, L., and T. Pusateri, "Automatic IP Multicast Without Explicit Tunnels (AMT)", draft-ietf-mboned-auto-multicast-10.txt (work in progress), March 2010.
- [HELLO] IANA, "PIM Hello Options", PIM-HELLO-OPTIONS per RFC4601 <http://www.iana.org/assignments/pim-hello-options>, March 2007.

Authors' Addresses

Dino Farinacci
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: dino@cisco.com

Greg Shepherd
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: gjshep@gmail.com

Yiqun Cai
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: ycai@cisco.com

Stig Venaas
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: stig@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 12, 2011

IJ. Wijnands, Ed.
Y. Cai
Cisco Systems, Inc.
February 8, 2011

PIM neighbor reduction for transit LAN's.
draft-wijnands-pim-neighbor-reduction-02

Abstract

PIM establishes a neighbor relationship with other routers directly connected to it on startup. Networks that are LANs or behave like a LAN, potentially create many PIM neighbors depending on how many routers are attached to it. If such a LAN is also a transit network (no directly connected source or receiver), many of the PIM procedures don't apply. This proposal describes a procedure to reduce the amount of neighbors established over a transit LAN.

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 August 12, 2011.

Copyright Notice

Copyright (c) 2011 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 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	3
1.1. Conventions used in this document	3
1.2. Terminology	3
2. Reducing the number of PIM neighbors	3
3. Bidir support	4
4. Generation ID Hello option	4
5. Hello suppression and options	5
5.1. PIM suppress Hello option	5
5.2. Backwards compatibility	6
6. Security Considerations	6
7. IANA considerations	6
8. Acknowledgments	7
9. Contributing authors	7
10. References	7
10.1. Normative References	7
10.2. Informative References	8
Authors' Addresses	8

1. Introduction

PIM sends hello messages to discover other PIM enabled routers that are directly connected on a particular interface and form a PIM neighbor relationship with them. Various PIM procedures depend on having a PIM neighbor elected as Designated Router (DR), like for PIM register messages [RFC4601] and processing IGMP reports [RFC4604]. Most of these procedures are specific to either directly connected receivers or senders and do not apply to transit networks. Networks that are LANs or behave like a LAN (Mi-PMSI) [I-D.ietf-l3vpn-2547bis-mcast] create as many PIM neighbors as there are PIM enabled routers directly connected to that LAN. For networks where the sources and/or RPs are only in few locations, which is a very typical deployment, it's very likely that many of these PIM neighbors are never used as a target in any PIM J/P message. Combined with the fact that on transit networks there are no directly connected receivers or senders, having a PIM neighbor relationship with all the PIM routers over a transit LAN network seems unnecessary. It is however still useful to have a PIM neighbor relationship with PIM routers that are used as target in the PIM Join or Prune (J/P) messages. We'll discuss these later in this draft.

The proposal is to not form unnecessary PIM neighbor relations by creating PIM neighbors dynamically on demand. Only PIM routers forwarding multicast data or on the path to the RP will be seen as a PIM neighbor. Other PIM routers on that LAN that act as receivers will stay passive and not form neighbor relationships. This will significantly reduce the number of PIM neighbors established over a LAN network where there are more passive receivers than there are senders. Networks that have directly connected senders and/or receivers are outside the scope of this draft.

1.1. 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].

1.2. Terminology

2. Reducing the number of PIM neighbors

PIM uses a unicast RIB to lookup the path to an upstream router for a particular Source or Rendezvous-Point (RP) [RFC4601]. The result of that lookup provides a directly connected next-hop that is used as the target in a PIM J/P message. [RFC4601] currently states that this next-hop also needs to be a PIM neighbor in order to send a PIM

J/P to it. However, whether you're not sending a PIM message because it is not a PIM neighbor or this upstream router is unable to parse the join, functionally does not make a big difference. The multicast tree can't be formed and traffic is interrupted. This draft proposes to send a PIM J/P to a target upstream router even if it is not a PIM neighbor. We also propose that a router accepts the PIM J/P and processes it as if it was received from a PIM neighbor. In most multicast deployments it is very likely that a next-hop for a source and an RP is also a PIM enabled router, so this is not considered to be a big issue. However, we do want to form a one-way PIM neighbor relationship with the target upstream router.

If a PIM router has a desire to send a PIM J/P to a non-PIM neighbor U, we propose to take one bit out of the PIM Join/Prune header 'Reserved' range and set it to 1 before we send the J/P packet. We call this bit the 'Hello Request' bit. A router that receives a PIM J/P with the 'Hello Request' bit on, sends a PIM hello out over the interface the PIM J/P was received on. The other routers on the LAN will receive the PIM hello and MAY form a one-way PIM neighbor relationship with U. A router that receives the Hello from U and has no interest in it MAY ignore the Hello to limit the amount of neighbor state. In the next PIM J/P the 'Hello Request' bit will be off because the PIM neighbor is known by the router sending the Join. Router U will continue to send periodic PIM Hello's out the interface as long as there is at least one downstream router joined over that interface for either a (*,G) or (S,G) state.

3. Bidir support

The support for PIM Bidir [RFC5015] on a LAN depends on the election of the Designated Forwarder (DF). The DF election mechanism has a few dependencies on PIM neighbors. [RFC5015] section 3.5.5 also describes a PIM Hello dependency on the DF election. For that reason routers that are bidir capable and a Candidate DF will send out a PIM Hello over that LAN. A PIM neighbor relationship will be established among the candidate DF routers. Note that a candidate DF router on a LAN is a router that has an RPF interface towards the RPA that is NOT on that same LAN. Please see [RFC5015] section 2.1 and 3.5.2 for details. It is expected that there are few Candidate DF routers and it's very likely these routers are already on the path to the RPA for the Sparse-Mode groups. We don't expect this procedure to add to the number of PIM neighbors that is established over that LAN.

4. Generation ID Hello option

PIM routers may use the generation ID in a PIM hello to make

downstream router trigger PIM J/P's to it. This feature is used for upstream router High Availability (HA) and when a router or interface becomes active. Using this feature there is no need to wait for the next periodic PIM J/P interval to (re)populate the forwarding state on the upstream router. If a Router or LAN becomes active, it is allowed to send a PIM hello on that LAN interface to speed up convergence, but it SHOULD not continue to send hello's periodically. Note, it's up to the downstream router(s) to either respond to this PIM Hello or ignore it if there is no interest in this PIM neighbor.

5. Hello suppression and options

PIM includes options in its Hello packets. We can group these options in two categories, options that are significant per neighbor or per LAN. For example, the GenID option is significant to the neighbor originating it, the Bidir option is significant to the LAN. Options that are significant per neighbor are learned as soon as a node has any interest that the neighbor. For these we don't need any special procedures. However, options relevant to the LAN, like Bidir capable or DR priority may not be learned because nodes on the LAN may suppress their Hello's using the procedures described in this draft. Its not important to know which nodes on the LAN support it or not, is good enough to know that at least one node does not support it. In order to discover the LAN specific options without creating PIM hello neighbor relationship between all the nodes we introduce the procedure below.

5.1. PIM suppress Hello option

We introduce a new PIM hello option called the PIM suppress option that is included in Hello's sent on the LAN. A PIM node on the LAN that receives this option in the PIM Hello (and supports it) will suppress its Hello if the set of included options match the options of this node. If this node has no interest in the sender of the Hello, no PIM neighbor relationship is created. The option set that this PIM neighbor advertised will be stored with an expire timer set to the advertised PIM hello holdtime. If this option set did already exist, only the option set expire timer is updated. The PIM hello periodic interval timer is started at the PIM hello interval time plus a random delay between 0 and 3 seconds. After the timer expires a PIM Hello is originated, unless a PIM Hello with the same set of options was received before the timer expired. This is similar to how PIM Join suppression works. With these procedures we are suppressing PIM hello's that share the same option set. Its likely that the PIM nodes on the LAN have the same option set, or at least have a limited set of option combinations. Below is the proposed PIM Hello suppress option encoding;

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|                                     Type = TDB                             Length = 0
|                                     |                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Type indicates PIM Hello suppression is supported.

5.2. Backwards compatibility

PIM nodes on the LAN that don't understand the suppress capability will obviously not suppress their Hello. They will just ignore the capability and create a PIM neighbor relation with the sender. This node does not expect other nodes to suppress their Hello so will assume that an upstream neighbor is not enabled with PIM. This may prevent PIM from sending PIM Join/Prunes. How this situation should be handled depends on the PIM implementation. Some implementations deployed in the field already ignore PIM neighbors for sending PIM Join/Prunes. For these implementations no special procedures are needed. Implementations that depend on PIM neighbors may only apply Hello suppression if all the PIM nodes on that LAN support the PIM suppress option. We propose the following two options to be supported;

As soon as one PIM node on the LAN does not support the suppress option all routers on the LAN will default back to sending periodic PIM hello's. Routers on the LAN continue to include the suppress option. As soon as all the routers on the LAN support the suppress option, PIM Hello suppression will be activated.

PIM hello suppression is always one and will not fall back to sending periodic PIM hello's.

6. Security Considerations

For securing PIM J/P messages please see the security section in [RFC4601].

7. IANA considerations

This document requests the reservation of a bit from the PIM Join/Prune header reserved field. This bit field is called 'Hello Request' bit.

8. Acknowledgments

Thanks to Stig Venaas, Eric Rosen and Maria Napierala for their comments on the draft.

9. Contributing authors

Below is a list of the contributing authors in alphabetical order:

Yiqun Cai
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134
USA
E-mail: ycai@cisco.com

IJsbrand Wijnands
Cisco Systems, Inc.
De kleetlaan 6a
1831 Diegem
Belgium
E-mail: ice@cisco.com

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, August 2006.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.
- [RFC5384] Boers, A., Wijnands, I., and E. Rosen, "The Protocol Independent Multicast (PIM) Join Attribute Format",

RFC 5384, November 2008.

10.2. Informative References

[I-D.ietf-l3vpn-2547bis-mcast]
Aggarwal, R., Bandi, S., Cai, Y., Morin, T., Rekhter, Y.,
Rosen, E., Wijnands, I., and S. Yasukawa, "Multicast in
MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10 (work
in progress), January 2010.

Authors' Addresses

IJsbrand Wijnands (editor)
Cisco Systems, Inc.
De kleetlaan 6a
Diegem 1831
Belgium

Email: ice@cisco.com

Yiqun Cai
Cisco Systems, Inc.
170 Tasman Drive
San Jose CA, 95134
USA

Email: ycai@cisco.com

