

Protocol Independent Multicast (pim)
Internet-Draft
Intended status: Standards Track
Expires: May 25, 2019

A. Peter, Ed.
Infinera
R. Kebler
V. Nagarajan
Juniper Networks, Inc.
T. Eckert
Huawei USA - Futurewei Technologies Inc.
S. Venaas
Cisco Systems, Inc.
November 21, 2018

Reliable Transport For PIM Register States
draft-anish-pim-reliable-registers-00

Abstract

This document introduces a hard-state, reliable transport for the existing PIM-SM registers states. This eliminates the needs for periodic NULL-registers and register-stop in response to each data-register or NULL-registers.

This specification uses the existing PIM reliability mechanisms defined by PIM Over Reliable Transport [[RFC6559](#)]. This is simply a means to transmit reliable PIM messages and does not require the support for Join/Prune messages over PORT as defined in [[RFC6559](#)].

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 [RFC2119](#) [[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 <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

Internet-Draft Reliable Transport For PIM Register States November 2018

This Internet-Draft will expire on May 25, 2019.

Copyright Notice

Copyright (c) 2018 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 (<https://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.	Reliable Register Overview	4
3.	Targeted Hellos	4
3.1.	New Hello Optional TLV's	5
3.2.	Differences from Link-Level hellos	6
3.3.	Address in Hello message	6
3.4.	Timer Values	6
3.5.	Targeted Neighbor	6
4.	Reliable Connection setup	7
4.1.	Active FHR	7
4.2.	Connection setup between two RP's	7
4.3.	Hello Generation ID and reconnect	7
4.4.	Handling Connection or reachability loss	8
5.	Anycast RP's	8
5.1.	Targeted Hellos and Neighbors	8
5.2.	Anycast-RP connection setup	8
5.3.	Anycast-RP state sync	9
5.4.	Anycast-RP change	9
5.5.	Anycast-RP with MSDP	10
6.	PIM-registers and Interoperation with legacy PIM nodes	10
6.1.	Initial packet-loss avoidance with PORT	10
6.2.	First-Hop-Router does not support PORT	10
6.3.	RP does not support PORT	10
6.4.	Data-Register free operations	11

7.	PORT message	11
7.1.	PORT register message TLV	11
7.2.	Sending and receiving PORT register messages	12
7.3.	PORT register-stop message TLV	12
7.4.	Sending and receiving PORT register stop messages	13

Internet-Draft Reliable Transport For PIM Register States November 2018

7.5.	PORT Keep-Alive Message	14
8.	Management Considerations	14
9.	IANA Considerations	14
9.1.	PIM Hello Options TLV	14
9.2.	PIM PORT Message Type	14
10.	Security Considerations	15
10.1.	PIM Register Threats	15
10.2.	Targeted Hello Threats	15
10.3.	TCP or SCTP security threats	15
11.	References	15
11.1.	Normative References	15
11.2.	Informative References	16
	Authors' Addresses	16

[1.](#) Introduction

Protocol Independent Multicast-Sparse Mode Register mechanism serves the following purposes.

- a, With a register, First-Hop-Router (FHR) informs the RP (that way the network) that a particular multicast stream is active
- b, A register helps avoid initial packet loss. (Initial packet loss could happen in an anycast-RP deployment even when packet registers are used.)
- c, Through its periodic refreshes register keeps RP informed about the aliveness of this multicast stream.

As it is defined in [\[RFC7761\]](#) , register mechanisms face limitations, when the number of multicast streams on the network is high, especially when one RP is expected to serve a large number of streams. These problems are mainly due to these factors.

- a, PIM register needs control-plane and data-plane intervention to handle it.

- b, Due to the nature of PIM register, First-Hop-Router and RP now needs to maintain states and timers for each register state entry.
- c, PIM register's requirements for periodic refresh and expiry, is quite aggressive and makes them vulnerable when the PIM speaker could not find cycles to meet these needs

To take for instance a major multicast application the IPTV. With the streaming servers connected to FHR. A restarting FHR would result in a burst of register messages at line rate. The RP may get

overloaded with packet registers. Which will continue untill RP is able to create states and do a register-stop. In the meantime many flows may go unserved due to drops. In addition to affecting multicast streams it may lead to starvation for other processing done by the controlplane application. With Anycast RP, this becomes even more tricky due to the control-plane's job to forward the registers to the rp-set.

In general: PIM registers have limitation in connections across WAN. It has no flow-control mechanisms, making PIM not compatible with IETF transport/congestion control expectations. It is challenging to deployed it over WAN or other bandwidth limited networks. High amount of state: periodic retransmission creates undesirable processing load. Especially with larger mesh-groups (re-send same (S,G) N-times, periodically).

[2.](#) Reliable Register Overview

Reliable PIM register extends PIM PORT [[RFC6559](#)] to have PIM register states to be sent over a reliable transport.

This document introduces 'targeted' hellos between any two PIM peers. This helps in capability negotiation and discovery between two PIM speakers (FHR and RP in the context of this document). Once this discovery happens, First-Hop-Router would setup a reliable transport connection based on the negotiated parameters.

Over this reliable connection, First-Hop-Router would start sending to RP the source and group addresses of the multicast streams active

with it. When any of this stream stops, First-Hop-Router would sent an update to RP about the streams that have stopped. This way once a reliable connection is setup, First-Hop-Router would update RP with its existing active multicast streams. Subsequently it would sent incremental updates about the change to RP.

For a multicast application that may demand initial packets or for bursty sources existing data-registers may be used. For them the RP would now respond with a 'reliable'-register-stop, which could persist until the First-Hop-Router withdraws the register-state.

3. Targeted Hellos

PIM hellos defined in PIM-SM [[RFC7761](#)] confines them to link level. This document extends these hellos to support 'targeted' hellos.

Targeted hellos are identical to existing hellos messages except that they would have an unicast address as its destination address. It

would traverse multiple hops using the unicast routing to reach the targeted hello neighbor.

3.1. New Hello Optional TLV's

Option Type: Targeted hello

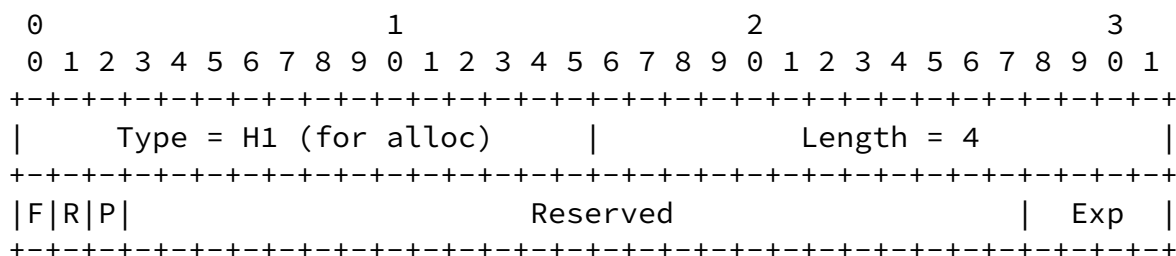


Figure 1: PIM Hello Optional TLV

Assigned Hello Type values can be found in IANA PIM registry.

Type: This is subject to IANA allocation. Its stated as H1 for reference

Length: Length in bytes for the value part of the Type/Length/Value encoding fixed as 4.

F: To be set by a router that wants to be a First-Hop-Router.

R: To be set by a RP that is capable taking the role of an RP as per the current states.

P: To be set by a device to indicate that its capable of handling packet registers. Packet registers can be send only if either of the peers have this bit set on their targeted hellos.

Reserved: Set to zero on transmission and ignored on receipt.

Exp: For experimental use [[RFC3692](#)]. One expected use of these bits would be to signal experimental capabilities. For example, if a router supports an experimental feature, it may set a bit to indicate this. The default behavior, unless a router supports a particular experiment, is to keep these bits reset and ignore the bits on receipt.

[3.2.](#) Differences from Link-Level hellos

The Major differences that Link-Level-Hellos have over Interface hellos are,

1. Destination address would be an unicast address unlike ALL-PIM-ROUTER destination address for link-level hellos
2. TTL value would be the system default TTL
3. Targeted Hellos SHOULD carry Targeted Hello Optional TLV (Defined in this document.)
4. Holdtime SHOULD NOT be set as 0xffff by a targeted hello sender, and such hellos should be discarded up on receive.

[3.3.](#) Address in Hello message

When sending targeted hellos, the sender SHOULD send with its primary reachable address (may be its loopback address) as the source address for the hellos. The other addresses that are relevant SHOULD be added in the secondary address list.

[3.4.](#) Timer Values

The timers relevant to this specification are in relation to PIM hello. The recommended timer values are

- 1: PIM Targeted Hello default refresh time : 60s (2 * Default Link-level hello time)
- 2: PIM Targeted Hello default hold time : 210s (3.5 times targeted hello default refresh time)

[3.5.](#) Targeted Neighbor

A Targeted PIM neighbor is a neighbor-ship established by virtue of exchanging targeted hello messages.

A First-Hop-Router (The initiator) that learns the RP's address would start sending hellos to the known RP address (could be anycast-address).

The RP (The Responder) when it receives this hello, would add sender as a targeted neighbor and would respond to this targeted neighbor from its primary address. The responder SHOULD also include its anycast address (If available) in the secondary address list. The

First-Hop-Router when receiving this hello would form a targeted neighbor with the anycast address.

The RP upon hold-time-out for the neighbor would remove this neighbor and its associated states.

The initiator or responder upon having a need to terminate a targeted neighbor MAY send hello with hold-time as 0.

[4.](#) Reliable Connection setup

A reliable connection has to be setup between the First-Hop-Router and RP for reliable registers to happen. Targeted hellos works as the medium for discovery and capability-negotiation between the two peers.

[4.1.](#) Active FHR

Once First-Hop-Router and RP discover each other, First-Hop-Router takes the active role. First-Hop-Router would listen for RP to connect once it forms targeted neighbor-ship with RP. The RP would be expected to use its primary address, which it would have used as the source address in its PIM hellos.

[4.2.](#) Connection setup between two RP's

In a network if there happens to be two RP's which are First-Hop-Router's too, then the mechanism could result in two connections getting established. It's desirable to have just one connection instead of two. First-Hop-Router could detect this condition when it receives hello with targeted hello header identifying that the RP wants to be First-Hop-Router too.

In this condition the connection setup could use the procedures stated in PIM over reliable transport [[RFC6559](#)]

[4.3.](#) Hello Generation ID and reconnect

If RP or First-Hop-Router gets into a situation needing for capability-renegotiation or reconnect, it would change the hello generation ID (gen-ID) to notify its peer to reset all the states and re-init this peering. The trigger for this could be configuration change or local operating parameter change, restart, etc. . .

[4.4.](#) Handling Connection or reachability loss

Connection loss or reachability loss could happen for one or more of the following reasons

- 1: PORT Keep-alive time out
- 2: Targeted neighbor loss
- 3: Reliable Connection close

Upon detecting one of these conditions, the connection with the peer SHOULD be closed immediately and the states created by the peer SHOULD be cleared after a grace-period, long enough for the peer to re-establish connection and re-sync the states.

This interval for re-sync would involve the initial time needed for re-establishing the connection, followed by transmission and reception of the states from FHR to RP over the reliable connection.

The ideal interval for this re-sync period could not be predicted, hence this should be a configurable parameter with default value as 300s.

[5.](#) Anycast RP's

PIM uses Anycast-RP [[RFC4610](#)] as a mechanism for RP redundancy. This section describes how anycast-RP would work with this specification.

[5.1.](#) Targeted Hellos and Neighbors

An RP that serves an anycast RP address, would know the primary addresses of other RP's serving the anycast address. These anycast-RP's would form a full mesh of targeted hello-neighbor-ships. In its targeted hello options tlv, the R bit MUST be set. The secondary address list in the PIM hello message SHOULD include the anycast-addresses that the sender is servicing.

[5.2.](#) Anycast-RP connection setup

A full mesh of connection is needed among the anycast-RP's of the same anycast address. Once targeted neighbor-ship is established, it would use the PIM PORT [[RFC6559](#)] procedures to setup reliable connection among them.

[5.3.](#) Anycast-RP state sync

An anycast-RP that gets the register state from a peer who's address is in the RP-set of address for the given group would update the register state and would retain the state. If the peer address is not in the RP-set address for the RP-group range, then the RP would replicate the state to all the other RP's in the RP-set. This procedure and forwarding rules are similar to the existing forwarding rules in Anycast-rp [[RFC4610](#)] register specification.

An RP should identify register state as a combinations of (source, group, 'PORT connection'). Where 'PORT connection' is the reliable connection with the PORT peer which had reported this s,g. Following considerations are made for a register-state identity.

- A. Reconnect: Connection between RP and First-Hop-Router could get re-established for various reasons. The register-states would get retransmitted over the new connection. Then it should be possible for RP to identify and timeout register-states on the old connection and retain the right set of states.
- B. DR-change: When DR in the First-Hop-LAN changes, a new First-Hop-Router would be retransmitting the same set of SG's that are already known and the old DR would be withdrawing the states advertised by it.
- C. FHR primary address change: In this case too connection would get re-established and state handling would be similar to case A.
- D. Multi-homed sources (but not on same LAN): In this case different First-Hop-Routers could be sending the same register-states. Then RP should be capable of identifying register-state along with the peer.

[5.4.](#) Anycast-RP change

In the event of nearest anycast-RP changing over to a different router, First-Hop-Router would detect that when it starts receiving PIM hellos with a different primary address for the same anycast address. This can also happen if the primary address of present anycast-RP has changed.

Upon detecting this scenario, the First-Hop-Router would establish connection and transmit its states to the new peer. Subsequently older connection would get terminated due to neighbor timeout.

Internet-Draft Reliable Transport For PIM Register States November 2018

[5.5.](#) Anycast-RP with MSDP

MSDP [[RFC3618](#)] is an alternative mechanism for 'active multicast stream state' synchronization between RP's. When MSDP is used, PIM's anycast synchronization need not be used. An anycast-RP network could use MSDP instead of PIM procedures for state synchronization among anycast-RP's. This document does not state any change in MSDP specification and usage

In such deployments, PIM will not have RP-set configured. As RP-set address is not available PIM procedures for Anycast-RP synchronization does not apply.

MSDP being a soft-state oriented protocol, it depends on frequent state refreshes over the reliable TCP transport. The support for mesh-groups in MSDP could be advantageous in some case.

[6.](#) PIM-registers and Interoperation with legacy PIM nodes

It may not be possible for PIM node to migrate altogether onto a PORT-registers in one go. Also there could be a few nodes in the network, which may not support PORT register states. This section states how both could interoperate.

[6.1.](#) Initial packet-loss avoidance with PORT

If its found that a few streams in the multicast network has to have initial packets to be delivered to the receiver, the existing PIM register mechanism could be used for them. For these streams a PORT register-stop message would be sent by the RP to First-Hop-Router.

[6.2.](#) First-Hop-Router does not support PORT

If the First-Hop-Router is not capable of doing PORT-register, then it would not establish targeted hello neighbor-ship with the RP. Hence reliable connection also would not be established. To handle such scenarios RP should accept PIM register messages and should respond to them with register-stop messages.

6.3. RP does not support PORT

If the RP is not capable of doing PORT-register, then it would not respond to the targeted hellos from the RP. Hence reliable connection also would not be established. In this case First-Hop-Router could sent existing packet registers to RP.

6.4. Data-Register free operations

If initial packet loss is acceptable in a multicast network, then Data-Registers could be avoided altogether in such networks. In such network PORT-Register-state specified in this document alone would be supported.

7. PORT message

This document defines new PORT register state message and PORT register-stop messages, to the existing messages in PORT specification.

7.1. PORT register message TLV

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	2	3	4	5	6	7	8	9																								
Type = P1 (for alloc)																Message Length																																															
Reserved																Exp.																																															
B N A																Reserved-1																																															
																src addr-1																z																															
z																grp addr-1																																															
z																2, 3, . . .																z																															
B N A																Reserved-n																																															

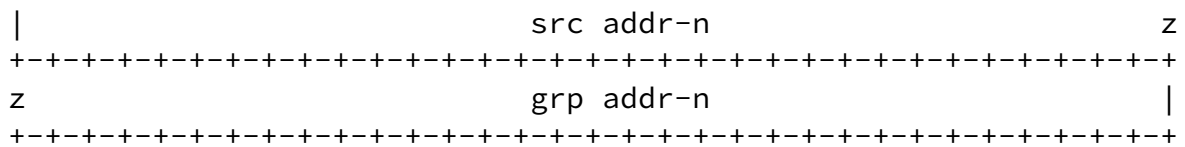


Figure 2: PORT Register State Message

Type: This is subject to IANA allocation. It would be next unallocated value, which is referred until allocation as P1.

Length: Length in bytes for the value part of the Type/Length/Value

B: As specified in [\[RFC7761\]](#) (set as 0 on send and ignore when received)

N: As specified in [\[RFC7761\]](#) (set as 1 on send and ignore when received.)

A: This flag signifies if the SG is to be Added or Deleted. When cleared, it indicates that the First-Hop-Router is withdrawing the SG.

src addr-x : This is the encoded source address of an ipv4/ipv6 stream

grp addr-x : This is the encoded group address of an ipv4/ipv6 stream

Reserved: Set to zero on transmission and ignored on receipt. These reserved bits are for properties that apply to the entire message.

Reserved-n: Set to zero on transmission and ignored on receipt. These reserved bits are for properties that apply to any particular sg.

Exp: : For experimental use.

[7.2.](#) Sending and receiving PORT register messages

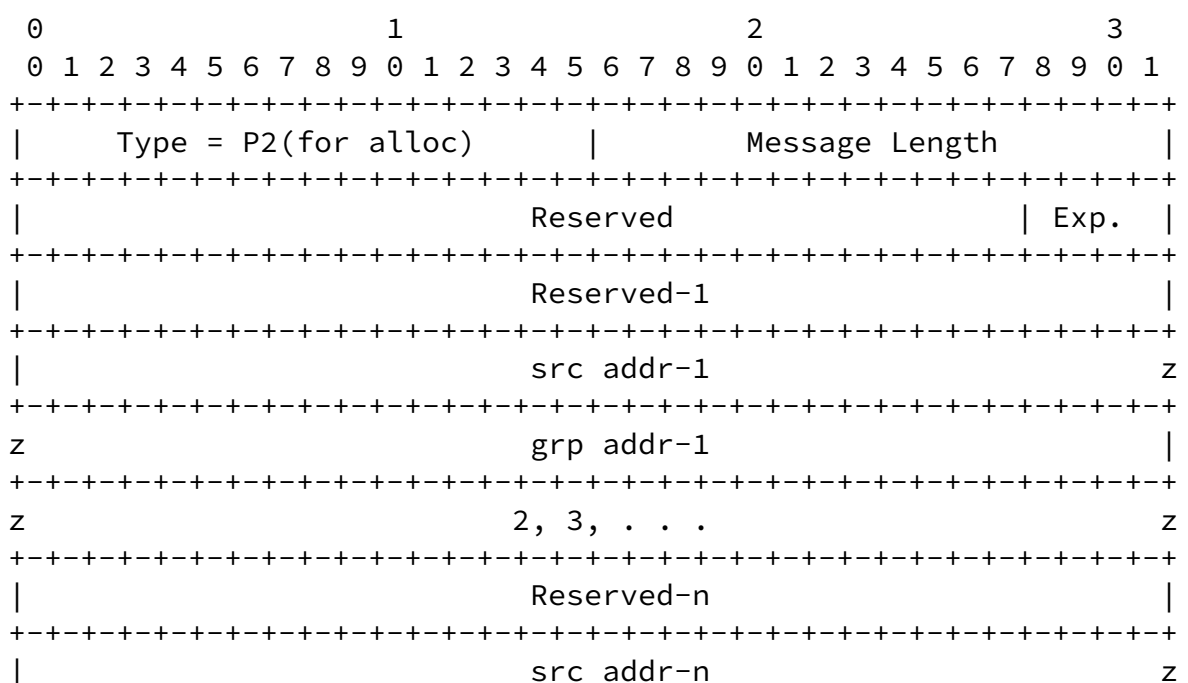
The First-Hop-Router upon learning a new stream would send a register state add message to the corresponding RP. If the reliable

connection got setup later, then once the connection becomes established it would send the entire list of streams active with it.

When KAT timer for a multicast stream expires, it would send an update to RP to remove that stream from its list.

An RP would maintain a database of multicast streams (src, grp) active along with the peer from which it had learned it. If the receiver RP is an anycast RP, it SHOULD re-transmit this register state message to each of the other anycast RP. An RP SHOULD NOT re-transmit a register state message it received from an another anycast-RP.

[7.3.](#) PORT register-stop message TLV



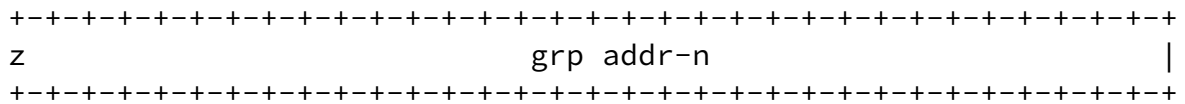


Figure 3: PORT Register Stop Message

Type: This is subject to IANA allocation. It would be next unallocated value, which is referred until allocation as P2.

Length: Length in bytes for the value part of the Type/Length/Value

src addr-x : This is the encoded source address of an ipv4/ipv6 stream

grp addr-x : This is the encoded group address of an ipv4/ipv6 stream

Reserved: Set to zero on transmission and ignored on receipt. These reserved bits are for properties that apply to the entire message.

Reserved-n: Set to zero on transmission and ignored on receipt. These reserved bits are for properties that apply to any particular sg.

Exp: : For experimental use.

[7.4.](#) Sending and receiving PORT register stop messages

PORT register-stop messages are send only as a response to receiving packet registers from a PIM peer, with which a reliable connection has been established. If reliable connection is not available, the

RP should consider the peer as a legacy node and should respond to this PIM register-stop message as defined in PIM-SM [[RFC7761](#)]

The First-Hop-Router up on receiving PORT-Register-Stop message should treat that as an indication from RP that it does not require the packets over the PIM tunnel and should stop sending register messages.

[7.5.](#) PORT Keep-Alive Message

The PORT Keep-alive messages as specified in PIM over Reliable

Transport [[RFC6559](#)] would be used to check the liveness of the peer and the reliable session

8. Management Considerations

PORT-register is capable of configuration free operations. But its recommended to have it as configuration controlled.

Implementation should provide knobs needed to stop supporting data-registers on a router.

9. IANA Considerations

This specification introduces new TLV in PIM hello and in PIM PORT messages. Hence the tlv ids for these needs IANA allocation

9.1. PIM Hello Options TLV

The following Hello TLV types needs IANA allocation. Place holder are kept to differentiate the different types.

Value	Length	Name	Reference
H1 (next-available)	4 (Fixed)	Targeted-Hello-Options	This document

Table 1: Place holder values for PIM Hello TLV type until IANA allocation

9.2. PIM PORT Message Type

The following PIM PORT message TLV types needs IANA allocation. Place holder are kept to differentiate the different types.

Value	Name	Reference
P1 (Next available)	PORT Register-state	This document

Table 2: Place holder values for PIM PORT TLV type for IANA allocation

[10.](#) Security Considerations

[10.1.](#) PIM Register Threats

PIM register is considered as security vulnerability for PIM networks. [\[RFC4609\]](#) The concern arises mainly due to the existing PIM register protocol design where in any remote node could start sending line-rate multicast traffic as PIM registers due to malfunction, mis-configuration or from a malicious remote node.

[10.2.](#) Targeted Hello Threats

This document introduces targeted hellos. This could be seen as a new security threat. Targeted hellos are part of other IETF protocol implementations, which are widely deployed. In future introduction of a mechanism similar to those stated in [RFC 7349](#) [\[RFC7349\]](#) could be used in PIM.

[10.3.](#) TCP or SCTP security threats

The security perception for this is stated in [\[RFC6559\]](#).

[11.](#) References

[11.1.](#) Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

[RFC3618] Fenner, B., Ed. and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", [RFC 3618](#), DOI 10.17487/RFC3618, October 2003, <<https://www.rfc-editor.org/info/rfc3618>>.

- [RFC4609] Savola, P., Lehtonen, R., and D. Meyer, "Protocol Independent Multicast - Sparse Mode (PIM-SM) Multicast Routing Security Issues and Enhancements", [RFC 4609](#), DOI 10.17487/RFC4609, October 2006, <<https://www.rfc-editor.org/info/rfc4609>>.
- [RFC4610] Farinacci, D. and Y. Cai, "Anycast-RP Using Protocol Independent Multicast (PIM)", [RFC 4610](#), DOI 10.17487/RFC4610, August 2006, <<https://www.rfc-editor.org/info/rfc4610>>.
- [RFC6559] Farinacci, D., Wijnands, IJ., Venaas, S., and M. Napierala, "A Reliable Transport Mechanism for PIM", [RFC 6559](#), DOI 10.17487/RFC6559, March 2012, <<https://www.rfc-editor.org/info/rfc6559>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, [RFC 7761](#), DOI 10.17487/RFC7761, March 2016, <<https://www.rfc-editor.org/info/rfc7761>>.

[11.2.](#) Informative References

- [RFC7349] Zheng, L., Chen, M., and M. Bhatia, "LDP Hello Cryptographic Authentication", [RFC 7349](#), DOI 10.17487/RFC7349, August 2014, <<https://www.rfc-editor.org/info/rfc7349>>.

Authors' Addresses

Anish Peter (editor)
Infinera
Brunton Road
Bangalore, KA 560025
India

Email: anish.ietf@gmail.com

Robert Kebler
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: rkebler@juniper.net

Internet-Draft Reliable Transport For PIM Register States November 2018

Vikram Nagarajan
Juniper Networks, Inc.
Electra, Exora Business Park
Bangalore, KA 560103
India

Email: vikramna@juniper.net

Toerless Eckert
Huawei USA - Futurewei Technologies Inc.

Email: tte+ietf@cs.fau.de

Stig Venaas
Cisco Systems, Inc.

Email: stig@cisco.com

Peter, et al.

Expires May 25, 2019

[Page 17]