TOC

Network Working Group	D. Farinacci
Internet-Draft	IJ. Wijnands
Intended status: Experimental	S. Venaas
Expires: January 11, 2010	cisco Systems
	M. Napierala
	AT&T Labs
	July 10, 2009

A Reliable Transport Mechanism for PIM
draft-ietf-pim-port-01.txt

Status of this Memo

Copyright Notice

Abstract

This draft describes how a reliable transport mechanism can be used by the PIM protocol to optimize CPU and bandwidth resource utilization by eliminating periodic Join/Prune message transmission. This draft proposes a modular extension to PIM to use either the TCP or SCTP transport protocol.

Table of Contents

1. Introduction

TOC

The goals of this specification are:

- To create a simple incremental mechanism to provide reliable PIM message delivery in PIM version 2.

- The reliable transport mechanism will be used for Join-Prune message transmission only.

- Can be used for link-local transmission of Join-Prune messages or multi-hop for use in a multicast VPN environments.

- When a router supports this specification, it need not use the reliable transport mechanism on every interface. That is, negotiation on per interface basis (or MDT basis) will occur.

The explicit non-goals of this specification are:

- Changes to the PIM protocol machinery as defined in [RFC4601] (Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, “Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised),” August 2006.). The reliable transport mechanism will be used as a plugin layer so the PIM component does not know it is really there.

- Provide support for both Datagram mode and Transport mode (see Section 1.2 (Definitions) for definitions) on the same physical interface or MDT.

This document will specify how periodic JP message transmission can be eliminated by using TCP [RFC0761] (Postel, J., “DoD standard Transmission Control Protocol,” January 1980.) or SCTP [RFC4960] (Stewart, R., “Stream Control Transmission Protocol,” September 2007.) as the reliable transport mechanism for JP messages.
This specification enables greater scalability in multicast deployment since the processing required for protocol state maintenance can be reduced. These enhancements to PIMv2 are applicable to IP multicast over routed services and VPNs [MCAST-VPN] (Rosen and Aggarwal, “Multicast in MPLS/BGP VPNs,” July 2007.). In addition to reduced processing on PIM enabled routers, another important feature is the reduced join and leave latency provided through a reliable transport. In many existing and emerging networks, particularly wireless and mobile satellite systems, link degradation due to weather, interference, and other impairments can result in temporary spikes in the packet loss. In these environments, periodic PIM joining can cause join latency when messages are lost causing a retransmission only 60 seconds later. By applying a reliable transport, a lost join is retransmitted rapidly. Furthermore, when the last user leaves a multicast group, any lost prune is similarly repaired and the multicast stream is quickly removed from the wireless/satellite link. Without a reliable transport, the multicast transmission could otherwise continue until it timed out, roughly 3 minutes later. As network resources are at a premium in many of these environments, rapid termination of the multicast stream is critical to maintaining efficient use of bandwidth.

1.1. Requirements Notation

TOC

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] (Bradner, S., “Key words for use in RFCs to Indicate Requirement Levels,” March 1997.).

1.2. Definitions

TOC

PORT: Stands for PIM Over Reliable Transport. Which is the short form for describing the mechanism in this specification where PIM can use the TCP or SCTP transport protocol.

JP Message: An abbreviation for a Join-Prune message.

Periodic JP: A JP message sent periodically to refresh state.

Incremental JP: A JP message sent as a result of state creation or deletion events. Also known as a triggered message.

Native JP: A JP message which is carried with an IP protocol type of PIM.

Reliable JP: A JP message using TCP or SCTP for transport.

Datagram Mode: The current procedures PIM uses by encapsulating JP messages in IP packets sent either triggered or periodically.

PORT Mode: Procedures used by PIM defined in this specification for sending JP messages over the TCP or SCTP transport layer.

MDT/PMSI: Used interchangeably in this document. An MDT tunnel is one used between PE router to provide support for a Multicast VPN. The new standards term for an MDT tunnel is a Provider-Network Multicast Service Interface or PMSI.

Segmented Multi-Access LAN: A segmented (or partitioned) LAN is like a virtual overlay network using the physical LAN to realize control and data packets. Multiple overlay networks may be created using the physical LAN, much like how VLANs or PMSI overlays are configured over a multi-access phsyical LAN. The interface associated with the partitioned LAN is like an NBMA interface type so explicit tracking can be accomplished. Each partitioned or segmented LAN has its own data-link encapsulation and link-layer multicast is still used to avoid head-end replication. This concept also applies to MDTs/PMSIs and is called "Segmented MDTs/PMSIs". A Segmented MDT/PMSI is a MDT/PMSI that has a single forwarder (i.e. a single ingress PE) for any multicast stream.

2. Protocol Overview

TOC

PIM Over Reliable Transport (PORT) is a simple extension to PIMv2 for refresh reduction of PIM JP messages. It involves sending incremental rather than periodic JPs over a TCP/SCTP connection between PIM neighbors.

This document only specifies PORT for the following physical or logical link types: point-to-point, segmented multi-access LAN, segmented MDT, PMSI [MCAST-VPN] (Rosen and Aggarwal, “Multicast in MPLS/BGP VPNs,” July 2007.), and point-to-point or point-to-multipoint GRE tunnel. For all other link types, such as multi-access LANs, Datagram Mode is used.

PORT can be incrementally used on a link between PORT capable neighbors. Routers which are not PORT capable can continue to use PIM in Datagram Mode. PORT capability is detected using new PORT Capable PIM Hello Options.

Once PORT is enabled on an interface and a PIM neighbor also announces that it is PORT enabled, only Reliable JP messages will be used. That is, only Reliable JP messages are accepted from, and sent to, that particular neighbor. Native JP messages may still be used for other neighbors.

Reliable JP messages are sent using a TCP/SCTP connection. When two PIM neighbors are PORT enabled, both for TCP or both for SCTP, they will immediately, or on-demand, establish a connection. If the connection goes down, they will again immediately, or on-demand, try to reestablish the connection. No JP messages (neither Native nor Reliable) are sent while there is no connection.

When PORT is used, only incremental JPs are sent from downstream routers to upstream routers. As such, downstream routers do not generate periodic JPs for routes which RPF to a PORT-capable neighbor.

For Joins and Prunes, which are received over a TCP/SCTP connection, the upstream router does not start or maintain timers on the outgoing interface entry. Instead, it explicitly tracks downstream routers which have expressed interest. An interface is deleted from the outgoing interface list only when all downstream routers on the interface, no longer wish to receive traffic.

There is no change proposed for the PIM JP packet format. However, for JPs sent over TCP/SCTP connections, no IP Header is included. The message begins with the PIM common header, followed by the JP message. See section Section 5 (Common Header Definition) for details on the common header.

TOC

3. New PIM Hello Options

3.1. PIM over the TCP Transport Protocol

TOC

Option Type: PIM-over-TCP Capable

```
 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 = 27           |         Length = X + 8        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    TCP Connection ID AFI      |          Reserved             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       TCP Connection ID                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Interface ID                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Allocated Hello Type values can be found in [HELLO-OPT] (IANA, “PIM Hello Options,” March 2007.).
When a router is configured to use PIM over TCP on a given interface, it MUST include the PIM-over-TCP Capable hello option in its Hello messages for that interface. If a router is explicitly disabled from using JP over TCP it MUST NOT include the PIM-over-TCP Capable hello option in its Hello messages. When the router cannot setup a TCP connection, it will refrain from including this option.
This option is only used when a physical or logical interface is a point-to-point, a segmented multi-access LAN, a segmented MDT, a PMSI [MCAST-VPN] (Rosen and Aggarwal, “Multicast in MPLS/BGP VPNs,” July 2007.), a point-to-point or point-to-multipoint GRE tunnel. In all other cases, such as multi-access LANs, Datagram Mode is used.
Implementations may provide a configuration option to enable or disable PORT functionality. We recommend that this capability be disabled by default.

Length: In bytes for the value part of the Type/Length/Value encoding. Where X is 4 bytes if AFI of value 1 (IPv4) is used and 16 bytes when AFI of value 2 (IPv6) is used [AFI] (IANA, “Address Family Indicators (AFIs),” February 2007.).

TCP Connection ID AFI: The AFI value to describe the address-family of the address of the TCP Connection ID field.

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

TCP Connection ID:

An IPv4 or IPv6 address used to establish the TCP connection. When this field is 0, a mechanism outside the scope of this spec is used to obtain the addresses used to establish the TCP connection.

Interface ID: An Interface ID is used to associate the connection a JP message is received over with an interface which is added or removed from an oif-list. When unnumbered interfaces are used or when a single Transport connection is used for sending and receiving JP messages over multiple interfaces, the Interface ID is used convey the interface from JP message sender to JP message receiver. When a PIM router sets a locally generated value for the Interface ID in the Hello TLV, it must send the same Interface ID value in all JP messages it is sending to the PIM neighbor.

3.2. PIM over the SCTP Transport Protocol

TOC

Option Type: PIM-over-SCTP Capable

```
 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 = 28           |         Length = X + 8        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   SCTP Connection ID AFI      |          Reserved             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      SCTP Connection ID                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Interface ID                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Allocated Hello Type values can be found in [HELLO-OPT] (IANA, “PIM Hello Options,” March 2007.).
When a router is configured to use PIM over SCTP on a given interface, it MUST include the PIM-over-SCTP Capable hello option in its Hello messages for that interface. If a router is explicitly disabled from using JP over SCTP it MUST NOT include the PIM-over-SCTP Capable hello option in its Hello messages. When the router cannot setup a SCTP connection, it will refrain from including this option.
This option is only used when a physical or logical interface is a point-to-point, a segmented multi-access LAN, a segmented MDT, a PMSI [MCAST-VPN] (Rosen and Aggarwal, “Multicast in MPLS/BGP VPNs,”

July 2007.), a point-to-point or point-to-multipoint GRE tunnel. In all other cases, such as multi-access LANs, Datagram Mode is used. Implementations may provide a configuration option to enable or disable PORT functionality. We recommend that this capability be disabled by default.

Length: In bytes for the value part of the Type/Length/Value encoding. Where X is 4 bytes if AFI of value 1 (IPv4) is used and 16 bytes when AFI of value 2 (IPv6) is used [AFI] (IANA, “Address Family Indicators (AFIs),” February 2007.).

SCTP Connection ID AFI: The AFI value to describe the address-family of the address of the SCTP Connection ID field.

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

SCTP Connection ID: An IPv4 or IPv6 address used to establish the SCTP connection. When this field is 0, a mechanism outside the scope of this spec is used to obtain the addresses used to establish the SCTP connection.

Interface ID: An Interface ID is used to associate the connection a JP message is received over with an interface which is added or removed from an oif-list. When unnumbered interfaces are used or when a single Transport connection is used for sending and receiving JP messages over multiple interfaces, the Interface ID is used convey the interface from JP message sender to JP message receiver. When a PIM router sets a locally generated value for the Interface ID in the Hello TLV, it must send the same Interface ID value in all JP messages it is sending to the PIM neighbor.

4. Establishing Transport Connections

TOC

While a router interface is PORT enabled, a PIM-over-TCP or a PIM-over-SCTP option is included in the PIM Hello messages sent on that interface. When a router on a PORT-enabled interface receives a Hello message containing a PIM-over-TCP/PIM-over-SCTP Option from a new neighbor, or an existing neighbor that did not previously include the option, it switches to PORT mode for that particular neighbor.
When a router switches to PORT mode for a neighbor, it stops sending and accepting Native JP messages for that neighbor. Any state from previous Native JP messages is left to expire as normal. It will also attempt to establish a Transport connection (TCP or SCTP) with the neighbor.

When the router is using TCP it will compare the TCP Connection ID it announced in the PIM-over-TCP Capable Option with the TCP Connection ID in the Hello received from the neighbor. The router with the lower Connection ID will do an active Transport open to the neighbor Connection ID. The router with the higher Connection ID will do a passive Transport open. An implementation may open connections only on-demand, in that case it may be that the neighbor with the higher Connection ID does the active open, see Section 4.3 (On-demand versus Pre-configured Connections). Note that the source address of the active open must be the announced Connection ID.

When the router is using SCTP, the IP address comparison need not be done since the SCTP protocol can handle call collision.

If PORT is used both for IPv4 and IPv6, both IPv4 and IPv6 PIM Hello messages are sent, both containing PORT Hello options. If two neighbors announce the same transport (TCP or SCTP) and the same Connection ID in the IPv4 and IPv6 Hello messages, then only one connection is established and is shared. Otherwise, two connections are established and are used separately.

The PIM router that performs the active open initiates the connection with a locally generated source transport port number and a well-known destination transport port number. The PIM router that performs the passive open listens on the well-known local transport port number and does not qualify the remote transport port number. See Section 5 (Common Header Definition) for well-known port number assignment for PORT.

When a Transport connection is established (or reestablished), the two routers MUST both send a full set of JP messages for which the other router is the upstream neighbor. This is needed to ensure that the upstream neighbor has the correct state. When moving from Datagram mode, or when the connection has gone down, the router cannot be sure that all the previous JP data was received by the neighbor. Any state received while in Datagram mode that is not refreshed, will be left to expire.

When a Transport connection goes down, Join or Prune state that was sent over the Transport connection is still retained. The neighbor should not be considered down until the neighbor timer has expired. This allows routers to do a control-plane switchover without disrupting the network. If a Transport connection is reestablished before the neighbor timer expires, the previous state is intact and any new JP messages sent cause state to be created or removed (depending on if it was a Join or Prune). If the neighbor timer does expire, only the upstream router, that has oif-list state, to the expired downstream neighbor will need to clear state. A downstream router, when an upstream neighboring router has expired, will simply RPF to a new neighbor where it would trigger JP messages like it would in [RFC4601] (Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, “Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised),” August 2006.). It is required of a PIM router to clear its neighbor table for a neighbor who has timed out due to neighbor holdtime expiration.

Note, since JP messages are sent over a Transport connection, no Prune Override or Join Suppression are possible for these messages.

4.1. TCP Connection Maintenance

TOC

TCP is designed to keep connections up indefinitely during a period of network disconnection. If a PIM-over-TCP router fails, the TCP connection may stay up until the neighbor actually reboots, and even then it may continue to stay up until you actually try to send the neighbor some information. This is particularly relevant to PIM, since the flow of JPs might be in only one direction, and the downstream neighbor might never get any indication via TCP that the other end of the connection isn't really there.

Most applications using TCP want to detect when a neighbor is no longer there, so that the associated application state can be released. Also, one wants to clean up the TCP state, and not keep half-open connections around indefinitely. This is accomplished by using PIM Hellos and by not introducing an application-specific or new PIM keep-alive message. Therefore, when a GENID changes from a received PIM Hello message, and a TCP connection is established or attempting to be established, the local side will tear down the connection and attempt to reopen a new one for the new instance of the neighbor coming up. However, if the connection is shared by multiple interfaces and the GENID changes only for one of them, then there was not a full reboot and the connection is likely to still work. In that case, the router should just resend all JP state for that particular neighbor. This is similar to how state is refreshed when GENID changes for PIM in datagram mode.

There may be situations where a router ignores some joins or prunes. E.g. due to wrong RP information or receiving joins on an RPF interface. A router may try to cache such messages and apply them later if only a temporary error. It may however also ignore the message, and later change its GENID for that interface to make the neighbor resend all state, including any that may have been previously ignored. It is possible that one receives JP messages for an interface/link that is down. As long as the neighbor has not expired, we recommend processing those messages as usual. If they are ignored, then the router should change the GENID for that interface when it comes back up, in order to get a full update.

TOC

4.2. Moving from PORT to Datagram Mode

There may be situations where an administrator decides to stop using PORT. If PORT is disabled on a router interface, we start expiry timers with the respective neighbor holdtimes as the initial values. Similarly if we receive a Hello message without a PORT Capable option from a neighbor, we start expiry timers for all JP state we have for that particular neighbor. The Transport connection should be shut down as soon as there are no more PIM neighborships using it. That is, for the connection we have associated local and remote Connection IDs. When there is no PIM neighbor with that particular remote connection ID on any interface where we announce the local connection ID, the connection should be shut down.

4.3. On-demand versus Pre-configured Connections

TOC

Transport connections could be established when they are needed or when a router interface to other PIM neighbors has come up. The advantage of on-demand Transport connection establishment is the reduction of router resources. Especially in the case where there is no need for n^2 connections on a network interface or MDT tunnel. The disadvantage is additional delay and queueing when a JP message needs to be sent and a Transport connection is not established yet.

If a router interface has become operational and PIM neighbors are learned from Hello messages, at that time, Transport connections may be established. The advantage is that a connection is ready to transport data by the time a JP messages needs to be sent. The disadvantage is there can be more connections established than needed. This can occur when there is a small set of RPF neighbors for the active distribution trees compared to the total number of neighbors. Even when Transport connections are pre-established before they are needed, a connection can go down and an implementation will have to deal with an on-demand situation.

Note that for TCP, it is the router with the lower Connection ID that decides whether to open a connection immediately, or on-demand. The router with the higher Connection ID should only initiate a connection on-demand. That is, if it needs to send a JP message and there is no currently established connection.

Therefore, this specification recommends but does not mandate the use of on-demand Transport connection establishment.

TOC

4.4. Possible Hello Suppression Considerations

This specification indicates that a Transport connection cannot be established until a Hello message is received. One reason for this is to determine if the PIM neighbor supports this specification and the other is to determine the remote address to use to establish the Transport connection.

There are cases where it is desirable to suppress entirely the transmission of Hello messages. In this case, it is outside the scope of this document on how to determine if the PIM neighbor supports this specification as well as an out-of-band (outside of the PIM protocol) method to determine the remote address to establish the Transport connection.

4.5. Avoiding a Pair of Connections between Neighbors

TOC

To ensure there are not two connections between a pair of PIM neighbors, the following set of rules must be followed. Let A and B be two PIM neighbors where A's Connection ID is numerically smaller than B's Connection ID, and each is known to the other as having a potential PIM adjacency relationship.

At node A:

- If there is already an established TCP connection to B, on the PIM-over-TCP port, then A MUST NOT attempt to establish a new connection to B. Rather it uses the established connection to send JPs to B. (This is independent of which node initiated the connection.)

- If A has initiated a connection to B, but the connection is still in the process of being established, then A MUST refuse any connection on the PIM-over-TCP port from B.

- At any time when A does not have a connection to B which is either established or in the process of being established, A MUST accept connections from B.

At node B:

- If there is already an established TCP connection to A, on the PIM-over-TCP port, then B MUST NOT attempt to establish a new connection to A. Rather it uses the established connection to send JPs to A. (This is independent of which node initiated the connection.)

- If B has initiated a connection to A, but the connection is still in the process of being established, then if A initiates a connection too, B MUST accept the connection initiated by A and must release the connection which it (B) initiated.

5. Common Header Definition

TOC

It may be desirable for scaling purposes to allow JP messages from different PIM protocol instances to be sent over the same Transport connection. Also, it may be desirable to have a set of JP messages for one address-family sent over a Transport connection that is established over a different address-family network layer.

To be able to do this we need a common header that is inserted and parsed for each PIM JP message that is sent on a Transport connection. This common header will provide both record boundary and demux points when sending over a stream protocol like Transport.

Each JP message will have in front of it the following common header in Type/Length/Value format. And multiple different TLV types can be sent over the same Transport connection.

To make sure PIM JP messages are delivered as soon as the TCP transport layer receives the JP buffer, the TCP Push flag will be set in all outgoing JP messages sent over a TCP transport connection.

PIM messages will be sent using destination TCP port number 8471. When using SCTP as the reliable transport, destination port number 8471 will be used. See Section 10 (IANA Considerations) for IANA considerations.

JP messages are error checked. This includes a bad PIM checksum, illegal type fields, illegal addresses or a truncated message. If any parsing errors occur in a JP message, it is skipped, and we proceed processing any following TLVs.

The current list of defined TLVs are:

IPv4 JP Message

```
 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 = 1             |        Length = X + 16        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Reserved                       |I-Type |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Interface ID                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Instance ID . . .                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     . . . Instance ID                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       PIMv2 JP Message                        |
|                               .                               |
|                               .                               |
|                               .                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The IPv4 JP common header is used when a JP message is sent that has all IPv4 encoded addresses in the PIM payload.

Length: In bytes for the value part of the Type/Length/Value encoding. Where X is the number of bytes that make up the PIMv2 JP message.

I-Type: Defines the encoding and semantics of the Instance ID field. Instance Type 0 means Instance ID is not used. Other values are not defined in this specification.

Interface ID: This is the Interface ID from the Hello TLV, defined in this specification, the PIM router is sending to the PIM neighbor. It indicates to the PIM neighbor what interface to associate the JP Join or Prune with.

Instance ID: This can be a VPN-ID. This field could also be a BGP Route Target (RT) or BGP Route Distinguisher (RD) as defined in [RFC4364] (Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)," February 2006.). This document only defines this for Instance Type 0. For type 0 the field should be set to zero on transmission and ignored on receipt.

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

PIMv2 JP Message: PIMv2 Join/Prune message and payload with no IP header in front of it. As you can see from the packet format diagram, multiple JP messages can go into one TCP/SCTP stream from the same or different Interface and Instance IDs.

IPv6 JP Message

```
     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 = 2             |        Length = X + 16        |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                        Reserved                       |I-Type |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                         Interface ID                          |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                     Instance ID . . .                         |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                     . . . Instance ID                         |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                       PIMv2 JP Message                        |
    |                               .                               |
    |                               .                               |
    |                               .                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The IPv6 JP common header is used when a JP message is sent that has all IPv6 encoded addresses in the PIM payload.

Length: In bytes for the value part of the Type/Length/Value encoding. Where X is the number of bytes that make up the PIMv2 JP message.

I-Type: Defines the encoding and semantics of the Instance ID field. Instance Type 0 means Instance ID is not used. Other values are not defined in this specification.

Interface ID: This is the Interface ID from the Hello TLV, defined in this specification, the PIM router is sending to the PIM neighbor. It indicates to the PIM neighbor what interface to associate the JP Join or Prune with.

Instance ID: This can be a VPN-ID, BGP Route Target (RT) or BGP Route Distinguisher (RD). This document only defines this for Instance Type 0. For type 0 the field should be set to zero on transmission and ignored on receipt.

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

PIMv2 JP Message: PIMv2 Join/Prune message and payload with no IP header in front of it. As you can see from the packet format diagram, multiple JP messages can go into one TCP/SCTP stream from the same or different Interface and Instance IDs.

6. Outgoing Interface List Explicit Tracking

TOC

Since this specification indicates the use of TCP/SCTP for PIM JP messages only over point-to-point or NBMA type links, explicit tracking can be achieved by tracking only oif-list state and not per-neighbor per oif-list state. This is true for segmented LANs and in segmented MDT/PMSI environments.

By using explicit tracking of oifs, the router tracks all downstream neighbors which have expressed interest in a route on a given interface. The list of tracked routers is one of the checks used to determine whether traffic needs to be forwarded on a given interface or not.

For (*,G) and (S,G) routes, the router starts forwarding traffic on an interface when a Join is received from a neighbor on such an interface. This is tracking the oif to the neighbor. When the neighbor sends a Prune, the interface is removed and forwarding of traffic stops on the interface.

When all interfaces are removed from the oif-list, the route entry can be removed.

For (S,G,R) routes, typically is tracking Prune state on the shared tree. One at least one downstream neighbor sends a Prune over a Transport connection, the (S,G,R) state is create with a empty outgoing interface list. If a subsequent JP is received over a Transport connection which has (*,G) in the join-list and does not have (S,G,R) in the prune-list, the upstream router will add the interface the JP message was received on to the oif-list. And oif-list based explicit tracking will occur just like in the (*,G) and (S,G) route case above. The only difference in the (S,G,R) route case, is that when the outgoing interface is pruned, the entry must stay in the route table or else forwarding will occur on the interfaces for the (*,G) entry. Therefore, explicit tracking for Prunes must be provided. Only when the (S,G,R) oif-list interfaces match the interfaces in the (*,G) can the (S,G,R) route be removed.

7. Multiple Instances and Address-Family Support

TOC

Multiple instances of the PIM protocol may be used to support multiple VPNs or within a VPN to support multiple address families. Multiple instances can cause a multiplier effect on the number of router resources consumed. To be able to have an option to use router resources more efficiently, muxing JP messages over fewer Transport connections can be performed.

There are two ways this can be accomplished, one using a common header format over a TCP connection and the other using multiple streams over a single SCTP connection.

Using the Common Header format described previously in this specification, using different TLVs, both IPv4 and IPv6 based JP messages can be encoded within a Transport connection. Likewise, within a TLV, multiple occurrences of JP messages can occur and are tagged with an instance-ID so multiple JP messages for different VPNs can use a single Transport connection.

When using SCTP multi-streaming, the common header is still used to convey instance information but an SCTP association is used, on a per-VPN basis, to send data concurrently for multiple instances. When data is sent concurrently, head of line blocking, which can occur when using TCP, is avoided.

8. Miscellany

TOC

No changes expected in processing of other PIM messages like PIM Asserts, Grafts, Graft-Acks, Registers, and Register-Stops. This goes for BSR and Auto-RP type messages as well.

This extension is applicable only to PIM-SM, PIM-SSM and Bidir-PIM. It does not take requirements for PIM-DM into consideration.

9. Security Considerations

TOC

Transport connections can be authenticated using HMACs MD5 and SHA-1 similar to use in BGP [RFC4271] (Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” January 2006.) and MSDP [RFC3618] (Fenner, B. and D. Meyer, “Multicast Source Discovery Protocol (MSDP),” October 2003.).

When using SCTP as the transport protocol, [RFC4895] (Tuexen, M., Stewart, R., Lei, P., and E. Rescorla, “Authenticated Chunks for the Stream Control Transmission Protocol (SCTP),” August 2007.) can be used, on a per SCTP association basis to authenticate PIM data.

10. IANA Considerations

TOC

This specification makes use of a TCP port number and a SCTP port number for the use of PIM-Over-Reliable-Transport that has been allocated by IANA. It also makes use of IANA PIM Hello Options allocations that should be made permanent. In addition, a registry for PORT message types is requested. This document defines two PORT message types. Type 1, IPv4 JP Message; and Type 2, IPv6 JP Message.

11. Contributors

TOC

In addition to the persons listed as authors, significant contributions were provided by Apoorva Karan and Arjen Boers.

12. Acknowledgments

TOC

The authors would like to give a special thank you and appreciation to Nidhi Bhaskar for her initial design and early prototype of this idea. Appreciation goes to Randall Stewart for his authoritative review and recommendation for using SCTP.
Thanks also goes to the following for their ideas and commentary review of this specification, Mike McBride, Toerless Eckert, Yiqun Cai, Albert Tian, Suresh Boddapati, Nataraj Batchu, Daniel Voce, John Zwiebel, Yakov Rekhter, Lenny Giuliano, Gorry Fairhurst and Sameer Gulrajani.
A special thank you goes to Eric Rosen for his very detailed review and commentary. Many of his comments are reflected as text in this specification.

13. References

TOC

13.1. Normative References

TOC

[RFC0761]	Postel, J., "DoD standard Transmission Control Protocol," RFC 761, January 1980 (TXT).
[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels," BCP 14, RFC 2119, March 1997 (TXT, HTML, XML).
[RFC3618]	Fenner, B. and D. Meyer, "Multicast Source Discovery Protocol (MSDP)," RFC 3618, October 2003 (TXT).
[RFC4271]	Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)," RFC 4271, January 2006 (TXT).
[RFC4364]	Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)," RFC 4364, February 2006 (TXT).
[RFC4601]	Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)," RFC 4601, August 2006 (TXT, PDF).
[RFC4895]	

	Tuexen, M., Stewart, R., Lei, P., and E. Rescorla, “Authenticated Chunks for the Stream Control Transmission Protocol (SCTP),” RFC 4895, August 2007 (TXT).
[RFC4960]	Stewart, R., “Stream Control Transmission Protocol,” RFC 4960, September 2007 (TXT).

13.2. Informative References

TOC

[AFI]	IANA, “Address Family Indicators (AFIs),” ADDRESS FAMILY NUMBERS http://www.iana.org/numbers.html, February 2007.
[HELLO-OPT]	IANA, “PIM Hello Options,” PIM-HELLO-OPTIONS per RFC4601 http://www.iana.org/assignments/pim-hello-options, March 2007.
[MCAST-VPN]	Rosen and Aggarwal, “Multicast in MPLS/BGP VPNs,” Internet Draft draft-ietf-l3vpn-2547bis-mcast-05.txt, July 2007.

Authors' Addresses

TOC

Dino Farinacci
cisco Systems
Tasman Drive
San Jose, CA 95134
USA
Email: dino@cisco.com

IJsbrand Wijnands
cisco Systems
Tasman Drive
San Jose, CA 95134
USA
Email: ice@cisco.com

Stig Venaas
cisco Systems
Tasman Drive
San Jose, CA 95134
USA
Email: stig@cisco.com

Maria Napierala
AT&T Labs
200 Laurel Drive
Middletown, New Jersey 07748>
USA

Email:	mnapierala@att.com