Network Working Group
Internet-Draft
Intended status: Experimental
Expires: August 29, 2008

Dino Farinacci
Greg Shepherd
cisco Systems
February 26, 2008

Population Count Extensions to PIM
draft-farinacci-pim-pop-count-02.txt

Status of this Memo

Copyright Notice

Abstract

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

Table of Contents

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

2.1. Status of Draft

This draft was originally written in May of 2004 and presented at the PIM working group in San Diego in August of 2004. At that time, the working group was not interested in pursuing this work. At the Vancouver IETF in December of 2007, there seem to be renewed interest in the design. So it is being submitted for the working group to decide if it should become a working group document.

2.2. Overview

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 Encoded-Source-Address format to the Join/Prune message as well as a new Hello TLV. The mechanism is applicable to IPv4 and IPv6 multicast. See [HELLO] for details.

2.3. 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.

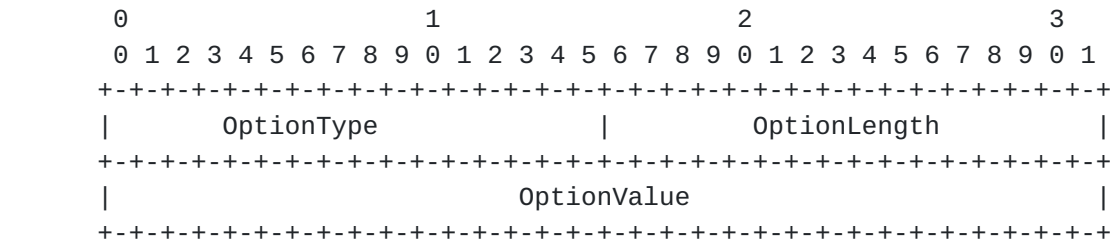
Stub Link: A link with only members joined to the group via IGMP or MLD. Which means there are no PIM routers joining for the multicast route on the link.

Transit Link: A link put in the oif-list for a multicast route because it was joined by PIM routers only (no IGMP or MLD reports were received on the link).

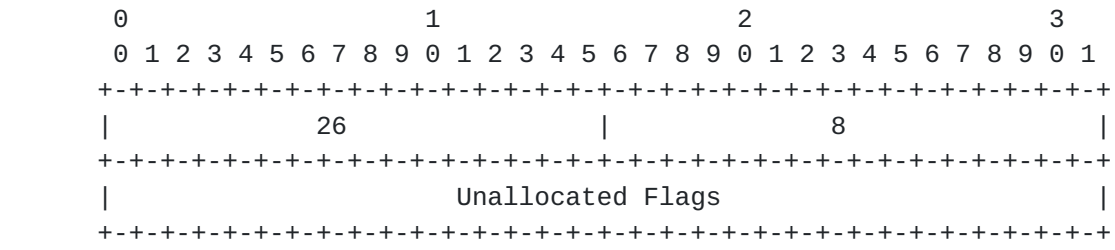
Dual Link: Is a link in the oif-list which is has the attributes of a Stub Link *and* Transit Link. That is, there are IGMP and MLD members as well as PIM joiners for the multicast route on the link.

3. New Hello TLV Pop-Count Support

When a PIM router sends a Join/Prune message to a neighbor, it will use the new form of the Encoded-Source-Address (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:



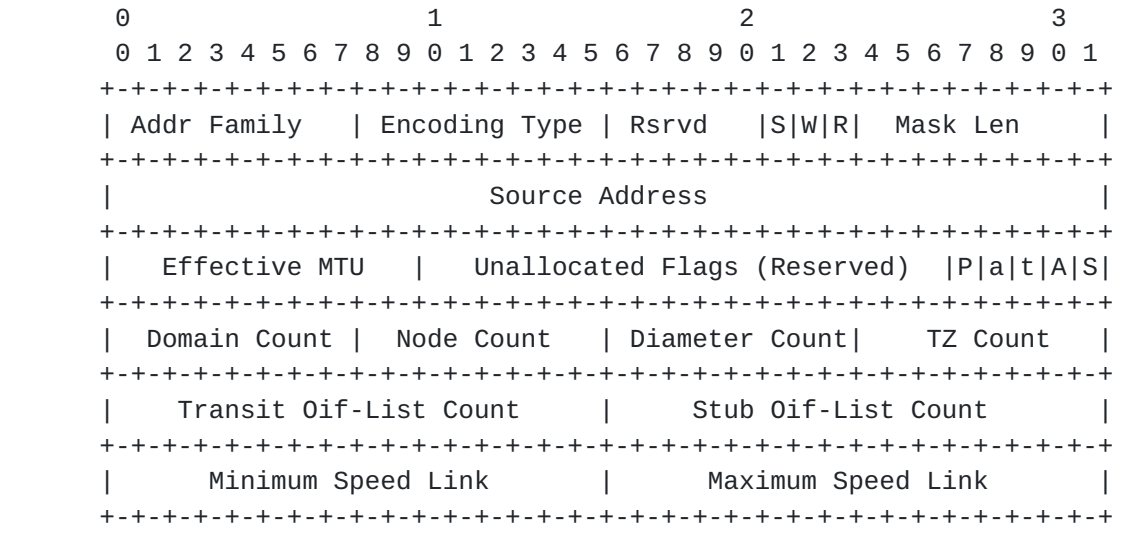
OptionType = 26, OptionLength = 8, there is no OptionValue semantics defined at this time but will be included for expandability and be defined in future revisions of this draft. The format will look like:



Unallocated Flags: for now should be sent as 0 and ignored on receipt. This field could be used to enable the use of future flags in the Unallocated Flags field of the new Encoded-Source-Address format defined below.

4. New Encoded-Source-Address 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 with the following format:



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

The format is identical to the format defined for the Address-Family [AFI] independent Encoded-Source-Address in [RFC4601] except there are additional fields appended. What distinguishes the above format from the format in [RFC4601] is the use of a different Encoding Type format. If the Encoding Type value is 1, the above format will be used.

The additional fields are:

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. The value is in units of 10s of bytes (i.e. so the value for a traditional Ethernet MTU would be 150).

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.

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.

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 it's 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 it's 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.

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.

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.

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.

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.

Transit Oif-List Count: 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 only. 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.

Stub Oif-List Count: 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 only 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.

Minimum Speed Link: this contains the minimum bandwidth rate (in mbps) for any link in the oif-list. 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. A value of 0 means the link speed is < 1 mbps.

Maximum Speed Link: this contains the maximum bandwidth rate (in mbps) for any link in the oif-list. 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. A value of 0 means the link speed is < 1 mbps.

This provides a way to obtain the lowest and highest speed link for the multicast distribution tree.

5. How to use Pop-Count Encoding

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 Encoded-Source-Address 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, an implementation must ignore them but still process the Prune as if it was in the original encoding described in [RFC4601].

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 it's 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 accconting 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 acccounting 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. Security Considerations

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

9. 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.

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.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

10.2. Informative References

[AFI] IANA, "Address Family Indicators (AFIs)", ADDRESS FAMILY NUMBERS http://www.iana.org/numbers.html, February 2007.

[AMT] Thaler, D., Talwar, M., Aggarwal, A., Vicisano, L., and T. Pusateri, "Automatic IP Multicast Without Explicit Tunnels (AMT)", draft-ietf-mboned-auto-multicast-08.txt (work in progress), October 2007.

[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: shep@cisco.com

Full Copyright Statement

Intellectual Property

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).