Network Working Group
Internet Draft
Expiration Date: May, 2001

Dino Farinacci
Procket Networks
Isidor Kouvelas
cisco Systems
Kurt Windisch
cisco Systems
November 22, 2000

State Refresh in PIM-DM
<draft-ietf-pim-refresh-02.txt>

Status of this Memo

1. Introduction

This proposal extends the PIM-DM [1] protocol specification by introducing the PIM State-Refresh control message.

When an (S,G) entry is created in a router for a directly connected source, if the interface directly connected to the source is the incoming interface for the entry, a new timer is started: the State-Refresh-Timer [SRT(S,G)]. The State-Refresh-Timer controls periodic transmission of the PIM State-Refresh message, which is propagated hop-by-hop down the (S,G) RPF tree. When received by a router on the RPF interface, the State-Refresh message causes existing prune state to be refreshed.

Addition of this heartbeat message solves many of the current problems with PIM-DM. It prevents the periodic timeout of prune state in routers, greatly reducing the re-flooding of multicast traffic down the pruned branches that expire periodically. It also causes topology changes to be realised quicker than the traditional 3 minute timeout.

2. Sending State-Refresh

For a given (S,G) tree, State-Refresh messages will be originated by all routers that use an interface directly connected to the source as the RPF interface for the source. Upon expiry of their (S,G) State-Refresh-Timer the PIM State-Refresh message will be sent on all PIM-DM interfaces with active PIM neighbors, except the interface connecting the source.

In addition, when the SRT(S,G) expires, the following timers are refreshed: SRT(S,G) is restarted with it's default value [Refresh-Interval], and all (S,G) pruned interface timers are refreshed.

The first-hop router will no longer originate state refresh messages when the (S,G) entry times out. The (S,G) entry timer for the first-hop router is updated only by the receipt of data and not upon expiry of the SRT(S,G) timer.

All other routers will forward state refresh messages only when receiving one from a neighbor, as described below.

State-Refresh messages are multicast using address 224.0.0.13 (ALL-PIM ROUTERS group) with protocol number equal to PIMv2 and a TTL of 1. The IP source address is set to the outgoing interface address and is rewritten hop-by-hop when forwarding.

The State-Refresh message contains the source and group the message is referring to, the originator address (for debugging purposes), routing information required by the LAN assert mechanism, a TTL value for scope control (different from header TTL), the state-refresh origination interval and a number of flags described below. The routing information, TTL and flags can be rewritten hop-by-hop.

The TTL value in the message is initialised by the originating router and can be either the result of local configuration, or the value of the largest TTL observed in data packets from the source so far. The TTL value will be decremented by downstream routers forwarding the State-Refresh message. Routers will only forward the State-Refresh message if the value of the TTL in the message is greater than 0 and larger than the configured local threshold. This will prevent State-Refresh messages from reaching areas of the network where data packets have not already created (S,G) state.

The flags in the message consist of the Prune-Indicator, Prune-Now and Assert-Override flags. The Prune-Indicator flag is cleared when the message is transmitted on an outgoing interface in forwarding state and set when the message is transmitted on a pruned interface. This mechanism is required to recover from situations where loss of consecutive refresh messages has caused an inconsistency in prune state on a branch of the (S,G) tree. The Prune-Now flag is required to provide a mechanism for rate-limiting control traffic on multi-access LANs. The Assert-Override flag is used to recover from assert winner failures.

3. Receiving State-Refresh

PIM State-Refresh messages are RPF flooded down the (S,G) tree using the data source address included in the message to determine the RPF neighbor. When a PIM State-Refresh message is received for a given (S,G), the following steps are taken:

- o Whenever a (S,G) State-Refresh message is received on the interface for RPF(S) by a router with no existing (S,G) entry, an (S,G) entry should be created. If the Prune-Indicator flag in the message indicates a forwarding branch, then all non-iif interfaces with PIM neighbors are set to forwarding state in the new entry. Otherwise, the new entry is created with prune state on all non-iif interfaces.

- o If the (S,G) State-Refresh message was received on an interface other than RPF(S) by a router with no existing (S,G) entry, then the message is ignored.

- o If the State-Refresh message was received on a (S,G) non-iif interface then the message is ignored. If the receiving interface corresponds to a LAN the message may still be processed according to the modified PIM Assert rules described in section 4.

- o If the State-Refresh was received on the (S,G) incoming interface from a PIM router other than the upstream neighbor (i.e, RPF neighbor or Assert winner), then the State-Refresh message is ignored. However, the message is still processed according to the modified PIM Assert rules described in section 4.

- o If the State-Refresh was received on the (S,G) incoming interface from the upstream neighbor (i.e, RPF neighbor or Assert winner), then all (S,G) pruned interface timers are refreshed. Further, if (S,G) is a negative cache entry, then the entry timer is also refreshed to its default value.

- o If the State-Refresh was received on the (S,G) incoming interface

from the upstream neighbor (i.e, RPF neighbor or Assert winner) and the Prune-Indicator flag in the message is set, indicating that it was forwarded down a pruned branch, but the local (S,G) entry is not a negative cache entry, then the Prune-Indicator flag in the message is cleared and a Join is sent upstream. To avoid duplicate Join generation from different downstream routers responding to a State-Refresh message, sending the Join is delayed by a random interval smaller than 3 seconds and a scheduled Join is canceled if one is received from another router on the LAN.

- o If the State-Refresh was received on the (S,G) incoming interface from the upstream neighbor (i.e, RPF neighbor or Assert winner) and the Prune-Indicator flag in the message is not set, indicating that it was forwarded down a forwarding branch, but the local (S,G) entry is a negative cache entry, then the Prune-Indicator flag in the message is set and a Prune is sent upstream. To avoid duplicate Prune generation from different downstream routers responding to a State-Refresh message, sending the Prune is delayed by a random interval smaller than 3 seconds and a scheduled Prune is canceled if one is received from another router on the LAN.

  In a scenario where there are multiple downstream routers, some with forwarding and some with negative cache entries, the routers with the negative caches will generate a prune on each State-Refresh message and the routers with the forwarding entries will have to Join override. To reduce the amount of control traffic created by such behavior, it is mandatory for a negative cache router to respond with a Prune to a State-Refresh message with a clear Prune-Indicator if the Prune-Now flag is set in the State-Refresh message. This flag will be set by the State-Refresh originator in one out of 3 messages transmitted. Downstream routers may also respond with a Prune to State-Refresh messages with the Prune-Now flag cleared.

- o If the State-Refresh was received on the (S,G) incoming interface from the upstream neighbor (i.e, RPF neighbor or Assert winner), then the Refresh message is retransmitted on all PIM interfaces other than the (S,G) incoming interface, provided that the TTL in the message is greater than 0 and larger than the configured threshold for the interface and that the interface does not have multicast boundary addresses configured for the group specified in the message. The IP header specifies the outgoing interface address as the source and the Refresh Packet is rewritten with the local router's preference, metric and mask for reaching S. If the (S,G) entry has prune state for the interface on which the refresh message is being sent, the Prune-Indicator flag in the message is set to indicate a pruned branch. The TTL in the forwarded message is one less than that of the received message.

4. State-Refresh processing on LANs

On multi-access LANs, State-Refresh messages double as Asserts. Possible forwarders and downstream routers use the routing metric information in the State-Refresh messages to decide who is the assert winner. In most ways the processing of such messages is identical to the assert processing rules described in [1].

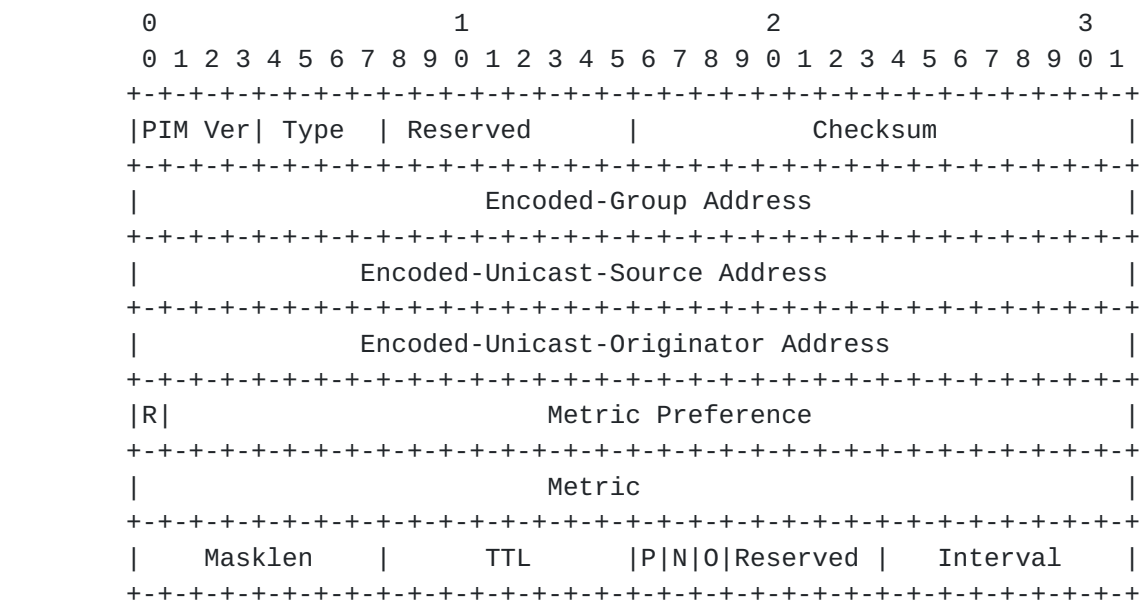
The assert rules described in [1] rely on the periodic timeout of prune state in routers to recover from situations where the assert winner on a LAN goes away. When operating under State-Refresh this no longer happens. In particular on a leaf LAN with multiple forwarders there are no downstream routers to timeout and join towards the new forwarder if the assert winner dies. Possible remaining forwarders that keep receiving State-Refresh messages will refresh their outgoing interface prune timers and will not time out and start forwarding.

To recover from this scenario, the assert processing needs to be slightly modified when operating under State-Refresh. Assert losers need to remember the last time they have heard a State-Refresh from a router on the LAN that has a better routing metric to the source. If a period of three times the [Refresh-Interval] elapses with no such report, then the Assert-Override flag will be set in the next forwarded State-Refresh message. If there are directly connected members reported by IGMP, the interface to the LAN will transition into forwarding state. The value of the Refresh-Interval used for timing out the winner, is extracted from the forwarded message (see section 5).

Downstream routers on a LAN that receive a State-Refresh message with the Assert-Override flag set, will discard the stored routing metric values for the assert winner and use the State-Refresh sender as their new RPF neighbor.

5. State-Refresh Message Packet Format

This section described the details of the packet format for the PIM DM State-Refresh Message. As with all PIM control messages, the State-Refresh message uses protocol number 103. It is multicast hop-by-hop to the `ALL-PIM-ROUTERS' group `224.0.0.13'.



PIM Version, Reserved, Checksum
Described in [2].

Type
State-Refresh message type value is 9. See [2] for types of other PIM control messages.

Encoded-Group Address
The group address to which the data packets were addressed, and which triggered the State-Refresh-Timer. Format described in [2].

Encoded-Unicast-Source Address
The address of the data packet source. Format described in [2].

Encoded-Unicast-Originator Address
The address of the first hop router that originated the State-Refresh message. Format described in [2].

Metric Preference, Metric, Masklen
Preference value assigned to the unicast routing protocol that provided the route to Host address, the metric in units applicable to the unicast routing protocol and the mask length used (needed for assert logic as described in [1]).

TTL
This is set by the originating router to either a locally configured value or the TTL observed in the data packets for the group and is decremented each time the State-Refresh

message is forwarded.

P
The Prune-Indicator flag. This is set if the State-Refresh message was forwarded on a pruned interface and cleared otherwise.

N
The Prune-Now flag. This is set by the State-Refresh originator on one out of three transmitted messages and is used by downstream routers on LANs to rate-control Prune transmission.

O
The Assert-Override flag. This is set by candidate forwarders on a LAN if a State-Refresh message has not been heard by the assert winner over the period of three times the [Refresh-Interval].

Reserved
Set to zero and ignored upon receipt.

Interval
Set by the originating router to the interval (in seconds) between consecutive State-Refresh messages for this source [Refresh-Interval].

6. Handling Router Failures

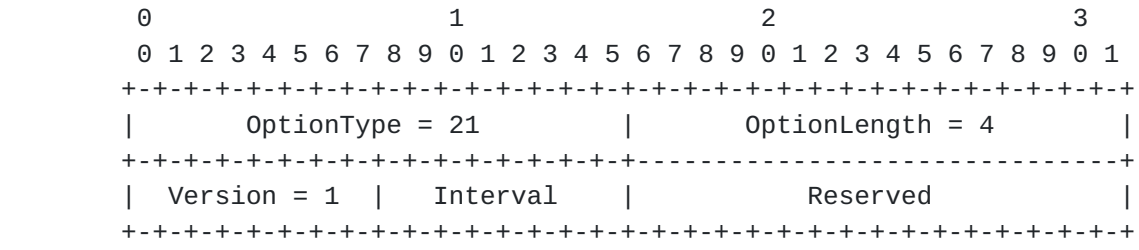
PIM Hello messages will contain a Generation ID (GenID) in a Hello option [3]. When a PIM Hello is received from an existing neighbor and the GenID differs from the previous ID, the neighbor has restarted and may not contain (S,G) state. In order to recreate the missing state, for each (S,G), all routers upstream of the failed router (i.e. those receiving the Hello on a non-iif) can send a new (S,G) PIM State-Refresh message on the interface that the Hello message was received. In order to avoid a burst of incoming State-Refresh messages at the recovering router, transmission of messages for different (S,G) entries has to be randomly spaced over a period of time. The duration of this period can be configured locally and a default value of 3 seconds is recommended. The Prune-Indicator flag of the State-Refresh message should be set to indicate if the recovering router is on a forwarding or pruned branch of the (S,G) tree.

7. Compatibility with Legacy PIM Routers

In order to enable incremental deployment of State-Refresh capable routers, additional mechanisms have to be used to prevent holes in

the distribution tree. These holes can be created because downstream routers without the State-Refresh capability will not send PIM grafts when (S,G) prune state times out. Upstream state-refresh capable routers will maintain (S,G) prune state. If a new receiver joins on a legacy branch, data will never reach this receiver.

Legacy routers are detected through the use of a new capability indicator in PIM Hello messages that can be used to inform neighbors whether a router is State-Refresh capable. The format of this option is as follows:



The Interval field is used to advertise the [Refresh-Interval] used by the router for originating SR messages for directly connected sources on this interface. Using this field, inconsistencies in origination intervals between first-hop routers for the same source can be detected.

The only protocol modification that is required to enable interoperability with detected legacy routers is in the procedures for packet reception:

o When a State-Refresh message is received on the (S,G) incoming interface from the upstream neighbor (i.e, RPF neighbor or Assert winner), then all (S,G) outgoing interface prune timers are refreshed except those leading to directly connected legacy routers. Further if all outgoing interfaces leading to State-Refresh capable routers are pruned then the entry timer is refreshed to its default value.

This will allow the prune state of the outgoing interface leading to the legacy router to timeout and change to forwarding state. As the entry timer will be updated by State-Refresh messages, the entry will persist even after the transition. If the entry was a negative cache entry a graft will be sent upstream as a result.

The above modifications will enable prune state to persist in subtrees of a source distribution tree that fulfill the following two conditions:

a) The subtree is entirely State-Refresh capable.

b) The path from the source to the subtree in entirely State-Refresh capable.

A subtree of the source distribution tree rooted at a legacy router as well as the path from the source to the subtree will not benefit from State-Refresh messages and will experience traditional dense mode flood and prune behavior.

8. References

[1] Deering, et al., "Protocol Independent Multicast Version 2 Dense Mode Specification", draft-ietf-pim-v2-dm-01.txt, November 1998.
[2] Estrin, et al., "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June 1998.
[3] Li, et al., "PIM Neighbor Hello GenId Option", draft-ietf-idmr-pim-hello-genid-00.txt, February 1999.

9. Acknowledgments

The authors would like to acknowledge Liming Wei (cisco), Tony Speakman (cisco) and John Zwiebel (cisco) for their comments and contributions to this specification.

10. Author Information

Dino Farinacci
Procket Networks
dino@procket.com

Isidor Kouvelas
cisco Systems, Inc.
kouvelas@cisco.com

Kurt Windisch
cisco Systems, Inc.
kurtw@cisco.com