State Refresh in PIM-DM

<draft-kouvelas-pim-refresh-00.txt>

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-Timers 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, and all (S,G) pruned interface timers are refreshed.

Stopping transmission of State-Refresh messages is controlled by the expiry of the (S,G) entry timer. The entry timer is not reset upon expiry of the SRT(S,G) timer and is only updated when data packets for the group are received, as in normal PIM-DM operation.

All other routers will send refresh 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 also contains the source and group the message is referring to, the originator address (for debugging purposes), routing information required by the assert mechanism, a TTL value for scope control (different from header TTL) and a Prune-Indicator flag described below. The routing information, TTL and branch indicator can be rewritten hop-by-hop.

The TTL value in the message is initially set by the originating router to 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 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.

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:

- 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.

- 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. If the receiving interface corresponds to a LAN the message may still be processed according to the normal PIM Assert rules described in [1].

- 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 normal PIM Assert rules described in [1].

- 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 normal PIM Assert rules described in [1].

- 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.

- 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 Graft is sent upstream.

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 then the configured thresh-
hold for the interface and that the interface does not have multi-
cast 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 mes-
sage 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 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 Ver | Type  | Reserved      |     Checksum      |
|         +-------------------+-------------------+-------------------+
| Encoded-Group Address                        |
|         +-------------------+-------------------+-------------------+
| Encoded-Unicast-Source Address               |
|         +-------------------+-------------------+-------------------+
| Encoded-Unicast-Originator Address           |
|         +-------------------+-------------------+-------------------+
| R | Metric Preference                             |
|     +-------------------+-------------------+
| Metric                                     |
|     +-------------------+-------------------+
| Masklen | TTL   | P | Reserved |
|     +-------------------+-------------------+
```

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

Type
State-Refresh message type value is TBD. 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 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.

Reserved
Set to zero and ignored upon receipt.

5. Handling Router Failures

PIM Hello messages will contain a Generation ID (GenID) in a Hello option. 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.

6. 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 can be created when grafts are not originated from legacy routers that have timed out prune state whereas State-Refresh capable routers higher up the tree have maintained prune state for the branch.

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 only protocol modification that is required to enable interoperability is in the procedures for packet reception:

- 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) prune timers are refreshed except those leading to 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 routed 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.
7. References


8. Acknowledgments

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

9. Author Information

Dino Farinacci
cisco Systems, Inc.
dino@cisco.com

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

Kurt Windisch
Advanced Network Technology Center
University of Oregon
kurtw@antc.uoregon.edu