Network Working Group   Vishnu Pavan Beeram
Internet Draft   Juniper Networks
Intended status: Informational   Ina Minei
                                                      Google, Inc
                                                    Yakov Rekhter
                                                 Juniper Networks
                                                      Ebben Aries
                                                         Facebook
                                                   Dante Pacella
                                                          Verizon

Expires: September 08, 2015   March 08, 2015

                   RSVP-TE Scalability - Recommendations
                   draft-beeram-mpls-rsvp-te-scaling-01

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), its areas, and its working groups. Note that
   other groups may also distribute working documents as Internet-
   Drafts.

   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."

   The list of current Internet-Drafts can be accessed at
   http://www.ietf.org/ietf/1id-abstracts.txt

   The list of Internet-Draft Shadow Directories can be accessed at
   http://www.ietf.org/shadow.html

   This Internet-Draft will expire on September 08, 2015.

Copyright Notice

Abstract

   RSVP-TE [RFC3209] describes the use of standard RSVP [RFC2205] to
   establish Label Switched Paths (LSPs). As such, RSVP-TE inherited
   some properties of RSVP that adversely affect its control plane
   scalability. Specifically these properties are (a) reliance on
   periodic refreshes for state synchronization between RSVP neighbors
   and for recovery from lost RSVP messages, (b) reliance on refresh
   timeout for stale state cleanup, and (c) lack of any mechanisms by
   which a receiver of RSVP messages can apply back pressure to the
   sender(s) of these messages.

   Subsequent to [RFC2205] and [RFC3209] further enhancements to RSVP
   and RSVP-TE have been developed. In this document we describe how an
   implementation of RSVP-TE can use these enhancements to address the
   above mentioned properties to improve RSVP-TE control plane
   scalability.

Conventions used in this document

   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 RFC-2119 [RFC2119].

Table of Contents

---

1. Introduction

   RSVP-TE [[RFC3209](#)] describes the use of standard RSVP [[RFC2205](#)] to
   establish Label Switched Paths (LSPs). As such, RSVP-TE inherited
   some properties of RSVP that adversely affect its control plane
   scalability. Specifically these properties are (a) reliance on
   periodic refreshes for state synchronization between RSVP neighbors
   and for recovery from lost RSVP messages, (b) reliance on refresh
   timeout for stale state cleanup, and (c) lack of any mechanisms by
   which a receiver of RSVP messages can apply back pressure to the
   sender(s) of these messages. The following elaborates on this.

1.1. Reliance on refreshes and refresh timeouts

   Standard RSVP [[RFC2205](#)] maintains state via the generation of RSVP
   Path/Resv refresh messages. Refresh messages are used to both
   synchronize state between RSVP neighbors and to recover from lost
   RSVP messages. The use of Refresh messages to cover many possible
   failures has resulted in two operational problems. The first
   relates to scaling, the second relates to the reliability and
   latency of RSVP signaling.

   The scaling problem is linked to the control plane resource
   requirements of running RSVP-TE. The resource requirements increase
   proportionally with the number of LSPs established by RSVP-TE. Each
   such LSP requires the generation, transmission, reception and
   processing of RSVP Path and Resv messages per refresh period.
   Supporting a large number of LSPs and the corresponding volume of
   refresh messages, presents a scaling problem for the RSVP-TE control
   plane.

   The reliability and latency problem occurs when a triggered (non-

refresh) RSVP message such as Path, Resv, or PathTear is lost in
transmission. Standard RSVP [RFC2205] recovers from a lost message
via RSVP refresh messages. In the face of transmission loss of RSVP
messages, the end-to-end latency of RSVP signaling, and thus the
end-to-end latency of RSVP-TE signaled LSP establishment, is tied to
the refresh interval of the Label Switch Router(s) experiencing the
loss. When end-to-end signaling is limited by the refresh interval,
the delay incurred in the establishment or the change of an RSVP-TE
signaled LSP may be beyond the range of what is acceptable in
practice. This is because RSVP-TE ultimately controls establishment

of the forwarding state required to realize RSVP-TE signaled LSPs.
Thus delay incurred in the establishment or the change of such LSPs
results in delaying the data plane convergence, which in turn
adversely impacts the services that rely on the data plane.

One way to address the scaling problem caused by the refresh volume
is to increase the refresh period, "R" as defined in Section 3.7 of
[RFC2205]. Increasing the value of R provides linear improvement on
RSVP-TE signaling overhead, but at the cost of increasing the time
it takes to synchronize state. For the reasons mentioned in the
previous paragraph, in the context of RSVP-TE signaled LSPs,
increasing the time to synchronize state is not an acceptable
option.

One way to address the reliability and latency of RSVP signaling is
to decrease the refresh period R. Decreasing the value of R
increases the probability that state will be installed in the face
of message loss, but at the cost of increasing refresh message rate
and associated processing requirements, which in turn adversely
affects RSVP-TE control plane scalability.

An additional problem is the time to clean up the stale state after
a tear message is lost. RSVP does not retransmit ResvTear or
PathTear messages. If the sole tear message transmitted is lost, the
stale state will only be cleaned up once the refresh timeout has
expired. This may result in resources associated with the stale
state being allocated for an unnecessary period of time. Note that
even when the refresh period is adjusted, the refresh timeout must
still expire since tear messages are not retransmitted. Decreasing
the refresh timeout by decreasing the refresh interval will speed up
timely stale state cleanup, but at the cost of increasing refresh

message rate, which in turn adversely affects RSVP-TE control plane
scalability.

1.2. Lack of back pressure

In standard RSVP, an RSVP speaker sends RSVP messages to a peer with
no regard for whether the peer's RSVP control plane is busy. There
is no control plane mechanism by which an RSVP speaker may apply
back pressure to the peer by asking the peer to reduce the rate of
RSVP messages that the peer sends to the speaker. RSVP-TE inherited
this from standard RSVP. Lack of such a mechanism could result in
RSVP-TE control plane congestion.

RSVP-TE control plane is especially susceptible to congestion during
link/node failures, as such failures produce bursts of RSVP-TE

---

messages: Path/Resv for re-routing LSPs affected by the failures,
Path/Resv for setup of new backup LSPs (as required by RSVP-TE Fast
Reroute [RFC4090]), Tear/Error messages for the affected LSPs. Note
that the load on the RSVP-TE control plane caused by these bursts is
in addition to the load due to the periodic refreshes of Path/Resv
messages for the LSPs not affected by the failures.

RSVP-TE control plane congestion may result in loss of RSVP
messages, which in turn have detrimental effects on the overall
system behavior. Path/Resv refreshes lost by a peer's busy control
plane will cause refresh timeout for some or all of its existing
RSVP-TE state on the peer, thus inadvertently deleting existing LSPs
and disrupting traffic carried over these LSPs. Triggered Path/Resv
lost by a peer's busy control plane may result in failure to
establish new backup LSPs used by RSVP-TE Fast Reroute [RFC4090]
before the state for the corresponding protected primary LSPs times
out, thus defeating the whole purpose of RSVP-TE Fast Reroute.

2. Recommendations

Subsequent to the publication of [RFC2205] and [RFC3209] further
enhancements to RSVP and RSVP-TE have been developed. In this
section we describe how these enhancements could be used to address
the problems listed in Section 1.

2.1. Eliminating reliance on refreshes and refresh timeouts

To eliminate reliance on refreshes for both state synchronization
between RSVP neighbors and for recovery from lost RSVP messages, as
well as to address both the refresh volume and the reliability
issues with RSVP mechanisms other than adjusting refresh rate, this
document RECOMMENDS the following:

-   Implement reliable delivery of Path/Resv messages using the
procedures specified in [RFC2961].

-   Indicate support for RSVP Refresh Overhead Reduction Extensions
(as specified in Section 2 of [RFC2961] by default, with the ability
to override the default via configuration.

-   Make the value of the refresh interval configurable with the
default value of 20 minutes.

To eliminate reliance on refresh timeouts, in addition to the above,
this document RECOMMENDS the following:

-   Implement reliable delivery of Tear/Err messages using the
procedures specified in [RFC2961]

-   Implement coupling the state of individual LSPs with the state of
the corresponding RSVP-TE signaling adjacency. When an RSVP-TE
speaker detects RSVP-TE signaling adjacency failure, the speaker
MUST clean up the LSP state for all LSPs affected by the failed
adjacency. The LSP state is the combination of "path state"
maintained as Path State Block and "reservation state" maintained as
Reservation State Block (see Section 2.1 of [RFC2205]).

-   Use of Node-ID based Hello session ([RFC3209], [RFC4558]) for
detection of RSVP-TE signaling adjacency failures. Make the value of
the node hello_interval [RFC3209] configurable; increase the default
value from 5 ms (as specified in Section 5.3 of [RFC3209]) to 9
seconds.

-   Implement procedures specified in [draft-chandra-mpls-enhanced-
frr-bypass] which describes methods to facilitate FRR that works
independently of the refresh-interval.

2.2. Providing the ability to apply back pressure

   To provide an RSVP speaker with the ability to apply back pressure
   to its peer(s) to reduce/eliminate RSVP-TE control plane congestion,
   in addition to the above, this document RECOMMENDS the following:

   - Use lack of ACKs from a peer as an indication of peer's RSVP-TE
   control plane congestion, in which case the local system SHOULD
   throttle RSVP-TE messages to the affected peer. This has to be done
   on a per-peer basis.

   - Retransmit of all RSVP-TE messages using exponential backoff, as
   specified in Section 6 of [RFC2961].

   - Increase the Retry Limit (Rl), as defined in Section 6.2 of
   [RFC2961], from 3 to 7.

   - Prioritize Tear/Error over trigger Path/Resv sent to a peer when
   the local system detects RSVP-TE control plane congestion in the
   peer.

2.3. Making Acknowledgements mandatory

   The reliable message delivery mechanism specified in [RFC2961]
   states that "Nodes receiving a non-out of order message containing a

   MESSAGE_ID object with the ACK_Desired flag set, SHOULD respond with
   a MESSAGE_ID_ACK object." To improve predictability of the system in
   terms of reliable message delivery this document RECOMMENDS that
   nodes receiving a non-out of order message containing a MESSAGE_ID
   object with the ACK_Desired flag set, MUST respond with a
   MESSAGE_ID_ACK object.

2.4. Clarifications on reaching Rapid Retry Limit (Rl)

   According to section 6 of [RFC2961] "The staged retransmission will
   continue until either an appropriate MESSAGE_ID_ACK object is
   received, or the rapid retry limit, Rl, has been reached." The
   following clarifies what actions, if any, a router should take once
   Rl has been reached.

   If it is the retransmission of Tear/Err messages and Rl has been

reached, the router need not take any further actions.

If it is the retransmission of Path/Resv messages and Rl has been
reached, then the router starts periodic retransmission of these
messages every 30 seconds. The retransmitted messages MUST carry
MESSAGE_ID object with ACK_Desired flag set. This periodic
retransmission SHOULD continue until an appropriate MESSAGE_ID ACK
object is received indicating acknowledgement of the (retransmitted)
Path/Resv message.

2.5. Avoiding use of Router Alert IP Option

In RSVP-TE the Path message is carried in an IP packet that is
addressed to the tail end of the LSP that is signaled using this
message. To make all the intermediate/transit LSRs process this
message, the IP packet carrying the message includes the Router
Alert IP option. The same applies to the PathTear message.

An alternative to relying on the Router Alert IP option is to carry
the Path or PathTear message as a sub-message of a Bundle message
[RFC2961], as Bundle messages are "addressed directly to RSVP
neighbors" and "SHOULD NOT be sent with the Router Alert IP option
in their IP headers" [RFC2961]. Notice that since a Bundle message
could contain only a single sub-message, this approach could be used
to send just a single Path or PathTear message. This document
RECOMMENDS implementing support for Bundle messages [RFC2961], and
carrying Path and PathTear message(s) as sub-message(s) of a Bundle
message.

2.6. Checking Data Plane readiness

In certain scenarios, like Make-Before-Break (MBB), a router needs
to move traffic from an existing LSP to a new LSP in the least
disruptive fashion. To accomplish this the data plane of the new LSP
must be operational before the router moves the traffic.

A possible mechanism by which the router can determine whether the
data plane of the new LSP is operational is specified in [draft-
bonica-mpls-self-ping]. This document RECOMMENDS implementing this
mechanism and using it whenever the ingress of an LSP needs to check

whether the data plane of the LSP is operational.

3. Security Considerations

    This document does not introduce new security issues. The security
    considerations pertaining to the original RSVP protocol [RFC2205]
    and RSVP-TE [RFC3209] remain relevant.

4. IANA Considerations

    This document makes no request of IANA.

    Note to RFC Editor: this section may be removed on publication as an
    RFC

5. Normative References

    [RFC2119]   Bradner, S., "Key words for use in RFCs to Indicate
                 Requirement Levels", BCP 14, RFC 2119, March 1997.

    [RFC2205]   Braden, R., "Resource Reservation Protocol (RSVP)",
                 RFC 2205, September 1997.

    [RFC2961]   Berger, L., "RSVP Refresh Overhead Reduction
                 Extensions", RFC 2961, April 2001.

    [RFC3209]   Awduche, D., "RSVP-TE: Extensions to RSVP for LSP
                 Tunnels", RFC 3209, December 2001.

    [RFC4090]   Pan, P., "Fast Reroute Extensions to RSVP-TE for LSP
                 Tunnels", RFC 4090, May 2005.

    [RFC4558]   Ali, Z., "Node-ID Based Resource Reservation (RSVP)
                 Hello: A Clarification Statement", RFC 4558, June 2006.

    [draft-bonica-mpls-self-ping] Ron Bonica, et al., "LSP Self-Ping",
                 draft-bonica-mpls-self-ping, (work in progress)

    [draft-chandra-mpls-enhanced-frr-bypass] Chandra Ramachandran, et
                 al., "Refresh Interval Independent FRR Facility
                 Protection", draft-chandra-mpls-enhanced-frr-bypass,

(work in progress)

6. Acknowledgments

   Most of the text in Section 1.1 has been taken almost verbatim from
   [RFC2961].

Authors' Addresses

   Vishnu Pavan Beeram
   Juniper Networks
   Email: vbeeram@juniper.net

   Ina Minei
   Google, Inc
   Email: inaminei@google.com

   Yakov Rekhter
   Juniper Networks
   Email: yakov@juniper.net

   Ebben Aries
   Facebook
   Email: exa@fb.com

   Dante Pacella
   Verizon
   Email: dante.j.pacella@verizon.com

   Markus Jork
   Juniper Networks
   Email: mjork@juniper.net