Network Working Group
Internet Draft

Dan Li
Jianhua Gao
Huawei

Arun Satyanarayana
Cisco

Intended Status: Informational
Expires: December 2007

June, 2007

Description of the RSVP-TE Graceful Restart Procedures
draft-li-ccamp-gr-description-00.txt

Status of this Memo

Abstract

The Hello message for the Resource Reservation Protocol (RSVP) has been defined to establish and maintain basic signaling node adjacencies for Label Switching Routers (LSRs) participating in a Multiprotocol Label Switching (MPLS) traffic engineered (TE) network. The Hello message has been extended for use in Generalized MPLS (GMPLS) network for state recovery of control channel or nodal faults.

GMPLS protocol definitions for RSVP also allow a restarting node to learn the label that it previously allocated for use on a Label Switching Path (LSP).

Further RSVP protocol extensions have been defined to enable a restarting node to recover full control plane state by exchanging RSVP messages with its upstream and downstream neighbors.

This document provides an informational clarification of the control plane procedures for a GMPLS network when there are multiple node failures, and describes how full control plane state can be recovered in different scenarios where the order in which the nodes restart is different.

This document does not define any new processes or procedures. All protocol mechanisms are already defined in the referenced documents.

Table of Contents

1. Introduction....3
2. Existing Procedures for Single Node Restart....4
2.1. Procedures defined in [RFC3473]....4
2.2. Procedures defined in [GR-EXT]....5
3. Multiple Node Restart Scenarios....5
4. RSVP State....6
5. Procedures for Multiple Node Restart....7
5.1. Procedures for the Normal Node....7
5.2. Procedures for the Restarting Node....7
5.2.1. Procedures for Scenario 1....7
5.2.2. Procedures for Scenario 2....9
5.2.3. Procedures for scenario 3....10
5.2.4. Procedures for scenario 4....11
5.2.5. Procedures for scenario 5....11
5.3. Consideration of Re-Use of Data Plane Resources....12
5.4. Consideration of Management Plane Intervention....12
6. Security Considerations....12
7. IANA Considerations....13
8. Acknowledgments....13
9. References....13
9.1. Normative References....13
10. Authors' Addresses....14
11. Full Copyright Statement....14
12. Intellectual Property Statement....15

1. Introduction

The Hello message for the Resource Reservation Protocol (RSVP) has been defined to establish and maintain basic signaling node adjacencies for Label Switching Routers (LSRs) participating in a Multiprotocol Label Switching (MPLS) traffic engineered (TE) network [RFC3209]. The Hello message has been extended for use in Generalized MPLS (GMPLS) network for state recovery of control channel or nodal faults through the exchange of the Restart Capabilities object [RFC3473].

GMPLS protocol definitions for RSVP [RFC3473] also allow a restarting node to learn the label that it previously allocated for use on a Label Switching Path (LSP) through the Recovery Label object carried on a Path message sent to a restarting node from its upstream neighbor.

Further RSVP protocol extensions have been defined [GR-EXT] to perform graceful restart and to enable a restarting node to recover full control plane state by exchanging RSVP messages with its upstream and downstream neighbors. State previously transmitted to the upstream neighbor (principally the downstream label) is recovered from the upstream neighbor on a Path message (using the Recovery Label object as described in [RFC3473]). State previously transmitted to the downstream neighbor (including the upstream label, interface identifiers, and the explicit route) is recovered from the downstream neighbor using a RecoveryPath message.

[GR-EXT] also extends the Hello message to exchange information about the ability to support the RecoveryPath message.

The examples and procedures in [RFC3473] and [GR-EXT] focus on the description of a single node restart when adjacent network nodes are operative. Although the procedures are equally applicable to multi-node restarts, no detailed explanation is provided.

This document provides and informational clarification of the control plane procedures for a GMPLS network when there are multiple node failures, and describes how full control plane state can be recovered in different scenarios where the order in which the nodes restart is different.

This document does not define any new processes or procedures. All protocol mechanisms already defined in [RFC3473] and [GR-EXT] are definitive.

2. Existing Procedures for Single Node Restart

This section documents for information the existing procedures defined in [RFC3473] and [GR-EXT]. Those documents are definitive, and the description here is non-normative. It is provided for informational clarification only.

2.1. Procedures defined in [RFC3473]

In the case of nodal faults, the procedures for the restarting node and the procedures for the neighbor of a restarting node are applied to the corresponding nodes. These procedures described in [RFC3473] are summarized as follows:

For the Restarting Node:

1) Tells its neighbors that state recovery is supported using the Hello message;

2) Recovers its RSVP state with the help of a Path message received from its upstream neighbor carrying the RECOVERY_LABEL object;

3) For bidirectional LSPs, the UPSTREAM_LABEL object on the received Path message is used to recover the corresponding RSVP state;

4) If the corresponding forwarding state in data plane is not existed, the node treats this as a setup for a new LSP. If the forwarding state in data plane is existed, the forwarding state is bound to the LSP associated with the message, and related forwarding state should be considered as valid and refreshed. In addition, if the node is not the tail-end of the LSP, the corresponding outgoing Path messages is sent with the incoming label from that entry carried in the UPSTREAM_LABEL object.

For the Neighbor of a restarting node:

1) Sends the Path message with RECOVERY_LABEL object containing a label value corresponding to the label value received in the most recently received corresponding Resv message;

2) Resumes refreshing Path state with the restarting node;

3) Resumes refreshing Resv state with the restarting node.

2.2. Procedures defined in [GR-EXT]

A new message is introduced in [GR-EXT] which is called the RecoveryPath message. The message is sent by the downstream neighbor of a restarting node to convey the contents of the last received Path message back to the restarting node.

The restarting node will receive the Path message with the RECOVERY_LABEL object from its upstream neighbor, and/or the RecoveryPath message from its downstream neighbor. The full RSVP state of the restarting node can be recovered from these two messages.

From the received Path message the following state can be recovered:

o Upstream data interface (from RSVP_HOP object)

o Label on the upstream data interface (from RECOVERY_LABEL object)

o Upstream label for bidirectional LSP (from UPSTREAM_LABEL object)

From the received RecoveryPath message the following state can be recovered:

o Downstream data interface (from RSVP_HOP object)

o Label on the downstream data interface (from RECOVERY_LABEL object)

o Upstream direction label for bidirectional LSP (from UPSTREAM_LABEL object)

The other objects also can be recovered either by regular Path message or RecoveryPath message, and Resv message.

3. Multiple Node Restart Scenarios

We define the following terms for the different node types:

Restarting - The node has restarted; communication with its neighbor nodes is restored, its RSVP state is under recovery.

Delayed Restarting - The node has restarted, but the communication with a neighbor node is interrupted (for example, the neighbor node needs to restart).

Normal - The normal node is the fully operational neighbor of a restarting or delayed restarting node.

There are five scenarios for multi-node restart. We will focus on the different positions of a restarting node. As shown in Figure 1, an LSP starts from Node A, traverses Nodes B and C, and ends at Node D.

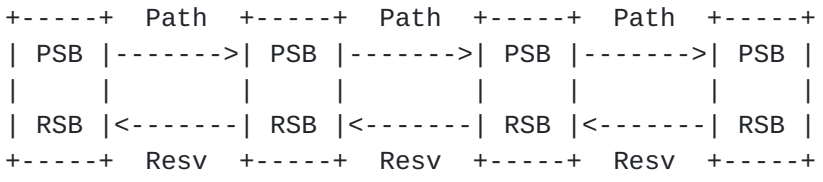

Figure 1 Two neighbor nodes restart

1) A Restarting node with downstream Delayed Restarting node. For example, in Figure 1, Nodes A and D are Normal nodes, Node B is a Restarting node, and Node C is a Delayed Restarting node.

2) A Restarting node with upstream Delayed Restarting node. For example, in Figure 1, Nodes A and D are Normal nodes, Node B is a Delayed Restarting node, and Node C is a Restarting node.

3) A Restarting node with downstream and upstream Delayed Restarting nodes. For example, in Figure 1, Node A is a Normal node, Nodes B and D are Delayed Restarting nodes, and Node C is a Restarting node.

4) A Restarting Ingress node with downstream Delayed Restarting node. For example, in Figure 1, Node A is a Restarting node, and Node B is a Delayed Restarting node. Nodes C and D are Normal nodes.

5) A Restarting Egress node with upstream Delayed Restarting node. For example, in Figure 1, Nodes A and B are Normal nodes, Node C is a Delayed Restarting node, and Node D is a Restarting node.

If the communication between two nodes is interrupted, the upstream node may think the downstream node is a Delayed Restarting node, or vice versa.

4. RSVP State

For each scenario, the RSVP state needs to be recovered at the restarting nodes are Path State Block (PSB) and Resv State Block (RSB), which are created when the node receives the corresponding Path message and Resv message.

According to [RFC2209], how to construct the PSB and RSB is really an implementation issue. In fact, there is no requirement to maintain separate PSB and RSB data structures. And in GMPLS, there

is a much closer tie between Path and Resv state so it is possible to combine the information into a single state block (the LSP state block). On the other hand, if P2MP is supported, it may be convenient to maintain separate upstream and downstream state. Note that the PSB and RSB are not upstream and downstream state since the PSB is responsible for receiving a Path from upstream and sending a Path to downstream.

Regardless of how the RSVP state is implemented, on recovery there are two logical pieces of state to be recovered and these correspond to the PSB and RSB.

5. Procedures for Multiple Node Restart

In this document, all the nodes are assumed to have the graceful restart capabilities which are described in [RFC3473] and [GR-EXT].

5.1. Procedures for the Normal Node

When the downstream Normal node detects its neighbor restarting, it must send a RecoveryPath message for each LSP associated with the restarting node for which it has previously sent a Resv message and which has not been torn down.

When the upstream Normal node detects its neighbor restarting, it must send a Path message with RECOVERY_LABEL object containing a label value corresponding to the label value received in the most recently received corresponding Resv message.

This document does not modify the procedures for the Normal node which are described in [RFC3473] and [GR-EXT].

5.2. Procedures for the Restarting Node

This document does not modify the procedures for the Restarting node which are described in [RFC3473] and [GR-EXT].

5.2.1. Procedures for Scenario 1

After the Restarting node restarts, it starts a Recovery Timer. Any RSVP state that has not been resynchronized when the Recovery Timer expires, should be cleared.

At the Restarting node (Node B in the example), full resynchronization with the upstream neighbor (Node A) is possible because Node A is a Normal node. The upstream Path information is recovered from the Path message received from Node A. Node B also

recovers the upstream Resv information (that it had previously sent to Node A) from the RECOVERY_LABEL object carried in the Path message received from Node A, but, obviously, some information (like the Recorded Route Object) will be missing from the new Resv message generated by Node B, and can not be supplied until the downstream Delayed Restarting node (Node C) restarts and sends a Resv.

After the upstream Path information and upstream Resv information has been recovered by Node B, the normal refresh procedure with the upstream Node A should be started.

As per [GR-EXT], the Restarting node (Node B) would normally expect to receive a RecoveryPath message from its downstream neighbor (Node C). It would use this to recover the downstream Path information, and would subsequently send a Path message to its downstream neighbor and receive a Resv message. But in this scenario, because the downstream neighbor has not restarted yet, Node B detects the communication with Node C is interrupted and must wait before resynchronizing with its downstream neighbor.

In this case, the Restarting node (Node B) follows the procedures in section 9.3 of [RFC3473] and may run a Restart Timer to wait for the downstream neighbor (Node C) to restart. If its downstream neighbor (Node C) has not restarted before the timer expires the corresponding LSPs may be torn down according to local policy [RFC3473]. Note, however, that the Restart Time value suggested in [RFC3473] is based on the previous Hello message exchanged with the node that has not restarted yet (Node C). Since this time value is unlikely to be available to the restarting node (Node B), a configured time value must be used if the timer is operated.

The RSVP state must be reconciled with the retained data plane state if the cross-connect information can be retrieved from the data plane. In the event of any mismatches, local policy will dictate the action that must be taken which could include:

- reprogramming the data plane

- sending an alert to the management plane

- tearing down the control plane state for the LSP.

In the case that the Delayed Restarting node never comes back, and where a Restart Timer is not used to automatically tear down LSPs, the LSPs can be tidied up through the control plane using a PathTear from the upstream node (Node A). Note that if Node C

restarts after this operation, the RecoveryPath message that it sends to Node B will not be matched with any state on Node B and will receive a PathTear as its response resulting in the teardown of the LSP at all downstream nodes.

5.2.2. Procedures for Scenario 2

In this case, the Restarting node (Node C) can recover full downstream state from its downstream neighbor (Node D) which is a Normal node. The downstream Path state can be recovered from the RecoveryPath message which is sent by Node D. This allows Node C to send a Path refresh message to Node D, and Node D will respond with a Resv message from which Node C can reconstruct the downstream Resv state.

After the downstream Path information and downstream Resv information has been recovered in Node C, the normal refresh procedure with downstream Node D should be started.

The Restarting node would normally expect to resynchronize with its upstream neighbor to re-learn the upstream Path and Resv state, but in this scenario, because the upstream neighbor (Node B) has not restarted yet, the Restarting node (Node C) detects that the communication with upstream neighbor (Node B) is interrupted. The Restarting node (Node C) follows the procedures in section 9.3 of [RFC3473] and may run a Restart Timer to wait the upstream neighbor (Node B) to restart. If its upstream neighbor (Node B) has not restarted before the Restart Timer expires, the corresponding LSPs may be torn down according to local policy [RFC3473]. Note, however, that the Restart Time value suggested in [RFC3473] is based on the previous Hello message exchanged with the node that has not restarted yet (Node B). Since this time value is unlikely to be available to the restarting node (Node C), a configured time value must be used if the timer is operated.

Note that no Resv message is sent to the upstream neighbor (Node B) because it has not restarted.

The RSVP state must be reconciled with the retained data plane state if the cross-connect information can be retrieved from the data plane.

In the event of any mismatches, local policy will dictate the action that must be taken which could include:

- reprogramming the data plane

- sending an alert to the management plane

- tearing down the control plane state for the LSP.

In the case that the Delayed Restarting node never comes back, and where a Restart Timer is not used to automatically tear down LSPs, the LSPs cannot be tidied up through the control plane using a PathTear from the upstream node(Node A), because there is no control plane connectivity to Node C from the upstream direction. There are two possibilities in [RFC3473]:

- Management action may be taken at the Restarting node to tear the LSP. This will result in the LSP being removed from Node C, and a PathTear being sent downstream to Node D.

- Management action may be taken at any downstream node (for example, Node D) resulting in a PathErr message with the Path_State_Reomved flag set being sent to Node C to tear the LSP state.

Note that if Node B restarts after this operation, the Path message that it sends to Node C will not be matched with any state on Node C and will be treated as a new Path message resulting in LSP setup. Node C should use the labels carried in the Path message (in the UPSTREAM_LABEL object and in the RECOVERY_LABEL object) to drive its label allocation, but may use other labels according to normal LSP setup rules.

5.2.3. Procedures for scenario 3

In this example, the Restarting node (Node C) is isolated. It's upstream and downstream neighbors have not restarted.

The Restarting node (Node C) follows the procedures in section 9.3 of [RFC3473] and may run a Restart Timer for each of its neighbors (Nodes B and D). If a neighbor has not restarted before its Restart Timer expires, the corresponding LSPs may be torn down according to local policy [RFC3473]. Note, however, that the Restart Time values suggested in [RFC3473] are based on the previous Hello message exchanged with the nodes that have not restarted yet. Since these time values are unlikely to be available to the restarting node (Node C), a configured time value must be used if the timer is operated.

During the Recovery Time, if the upstream Delayed Restarting node has restarted, the procedure for scenario 1 can be applied.

During the Recovery Time, if the downstream Delayed Restarting node has restarted, the procedure for scenario 2 can be applied.

In the case that neither Delayed Restarting node ever comes back, and where a Restart Timer is not used to automatically tear down LSPs, management intervention is required to tidy up the control plane and the data plane on the nodes that are waiting for the failed device to restart.

If the downstream Delayed Restarting node restarts after the cleanup of LSPs at Node C, the RecoveryPath message from Node D will be responded with a PathTear message. If the upstream Delayed Restarting node restarts after the cleanup of LSPs at Node C, the Path message from Node B will be treated as a new LSP setup request, but the setup will fail because Node D cannot be reached - Node C will respond with a PathErr message. Since this happens to Node B during its restart processing, it should follow the rules of [GR-EXT] and tear down the LSP.

5.2.4. Procedures for scenario 4

When the Ingress node (Node A) restarts, it does not know which LSPs it caused to be created. Usually, however, this information is retrieved from the management plane or from the configuration requests stored in non-volatile form in the node in order to recover the LSP state.

Furthermore, if the downstream node (Node B) is a Normal node, according to the procedures in [GR-EXT], the ingress will receive a RecoveryPath message and will understand that it was the ingress of the LSP.

However, in this scenario, the downstream node is a Delayed Restarting node, so Node A must rely on the information from the management plane or stored configuration, or it must wait for Node B to restart.

In the event that Node B never restarts, management plane intervention may be used at Node A to clean up any LSP state restored from the management plane or from local configuration.

5.2.5. Procedures for scenario 5

In this scenario the Egress node (Node D) restarts, and its upstream neighbor (Node C) has not restarted. In this case, the Egress node is completely unaware of the LSPs. It has no downstream neighbor to help it, and no management plane or configuration

information. The Egress node must simply wait until its upstream neighbor restarts and gives it the information as Path messages carrying RECOVERY_LABEL objects.

5.3. Consideration of Re-Use of Data Plane Resources

Fundamental to the processes described above is an understanding that data plane resources may remain in use (allocated and cross-connected) when control plane state has not been fully resynchronized because some control plane nodes have not restarted.

It is assumed that these data plane resources might be carrying traffic and should not be reconfigured except through application of operator-configured policy, or as a direct result of operator action.

In particular, new LSP setup requests from the control plane or the management plane should not be allowed to use data plane resources that are still in use. Specific action must first be taken to release the resources.

5.4. Consideration of Management Plane Intervention

The management plane must always retain the ability to control data plane resources and to over-ride the control plane. In this context, the management plane must always be able to release data plane resources that were previously in place for use by control-plane established LSPs. Further, the management plane must always be able to instruct any control plane node to tear down any LSP.

Operators should be aware of the risks of misconnection that could be caused by careless manipulation from the management plane of in-use data plane resources.

6. Security Considerations

This document clarifies the procedures to be performed on RSVP agents that neighbor one or more restarting RSVP agents. In the case of the control plane in general, and the RSVP agent in particular, where one or more nodes carrying one or more LSPs are restarted due to external attacks, the procedures defined in [GR-EXT] and described in this document provide the ability for the restarting RSVP agents to recover the RSVP state in each restarting node corresponding to the LSPs, with the least possible perturbation to the rest of the network. Ideally, only the neighboring RSVP agents should notice the restart and hence need to perform additional processing. This allows for a network with

active LSPs to recover LSP state gracefully from an external attack, without perturbing the data/forwarding plane state.

7. IANA Considerations

This document defines no new protocols or extensions and makes no requests to IANA for registry management.

8. Acknowledgments

We would like to thank Adrian Farrel, Dimitri Papadimitriou, and Lou Berger for their useful comments.

9. References

9.1. Normative References

[RFC2209] R. Braden, L. Zhang, "Resource ReSerVation Protocol (RSVP) -- Version 1 Message Processing Rules", RFC 2209, September 1997.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[GR-EXT] A. Satyanarayana, R. Rahman, "Extensions to GMPLS RSVP Graceful Restart", Internet Draft, work in progress, draft-ietf-ccamp-rsvp-restart-ext-08.txt, January 2007.

10. Authors' Addresses

Dan Li
Huawei Technologies Co., Ltd.
F3-5-B R&D Center, Huawei Base,
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28972910
Email: danli@huawei.com

Jianhua Gao
Huawei Technologies Co., Ltd.
F3-5-B R&D Center, Huawei Base,
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28972902
Email: gjhhit@huawei.com

Arun Satyanarayana
Cisco Systems, Inc.
170 West Tasman Dr.
San Jose, CA 95134, USA

Phone: +1 408 853-3206
Email: asatyana@cisco.com

11. Full Copyright Statement

12. Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

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".