Network Working Group Yakov Rekhter (Juniper Networks)
Internet Draft Rahul Aggarwal (Juniper Networks)
Expiration Date: February 2006

Graceful Restart Mechanism for BGP with MPLS

draft-ietf-mpls-bgp-mpls-restart-05.txt

Status of this Memo

IPR Disclosure Acknowledgement

Abstract

A mechanism for BGP that helps minimize the negative effects on routing caused by BGP restart has already been developed an is described in a separate document ("Graceful Restart Mechanism for BGP"). This document extends this mechanism to also minimize the negative effects on MPLS forwarding caused by the Label Switching Router's (LSR's) control plane restart, and specifically by the restart of its BGP component when BGP is used to carry MPLS labels and the LSR is capable of preserving the MPLS forwarding state across the restart.

The mechanism described in this document is agnostic with respect to the types of the addresses carried in the BGP Network Layer Reachability Information (NLRI) field. As such it works in conjunction with any of the address famililies that could be carried in BGP (e.g., IPv4, IPv6, etc...)

The mechanism described in this document is applicable to all LSRs, both those with the ability to preserve their forwarding state during BGP restart and those without (although the latter need to implement only a subset of the mechanism described in this document). Supporting a subset of the mechanism described here by the LSRs that can not preserve their MPLS forwarding state across the restart would not reduce the negative impact on MPLS traffic caused by their control plane restart, but it would minimize the impact if their neighbor(s) are capable of preserving the forwarding state across the restart of their control plane and implement the mechanism described here.

Specification of Requirements

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].

1. Introduction

For the sake of brevity in the context of this document by "MPLS forwarding state" we mean either <incoming label -> (outgoing label, next hop)>, or <Forwarding Equivalence Class (FEC) -> (outgoing label, next hop)>, or <incoming label -> label pop, next hop>, or <incoming label, label pop> mapping. In the context of this document the forwarding state that is referred to in [1] means MPLS forwarding state, as defined above. In the context of this document the term "next hop" refers to the next hop as advertised in BGP.

In the case where a Label Switching Router (LSR) could preserve its MPLS forwarding state across restart of its control plane, and specifically its BGP component, and BGP is used to carry MPLS labels (e.g., as specified in [RFC3107]), it may be desirable not to perturb the LSPs going through that LSR (and specifically, the LSPs established by BGP) after failure of or restart of the BGP component of the control plane. In this document, we describe a mechanism that allows this goal to be accomplished. The mechanism described in this document works in conjunction with the mechanism specified in [1]. The mechanism described in this document places no restrictions on the types of addresses (address families) that it can support.

The mechanism described in this document is applicable to all LSRs, both those with the ability to preserve forwarding state during BGP restart and those without (although the latter need to implement only a subset of the mechanism described in this document). Supporting a subset of the mechanism described here by the LSRs that can not preserve their MPLS forwarding state across the restart would not reduce the negative impact on MPLS traffic caused by their control plane restart, but it would minimize the impact if their neighbor(s) are capable of preserving the forwarding state across the restart of their control plane and implement the mechanism described here. The subset includes all the procedures described in this document, except the procedures in Sections 4.1, 4.2, 4.3 and 5.

2. General requirements

First of all an LSR MUST implement the Graceful Restart Mechanism for BGP, as specified in [1]. Second, the LSR SHOULD be capable of preserving its MPLS forwarding state across the restart of its control plane (including the restart of BGP). Third, for the <Forwarding Equivalence Class (FEC) -> label> bindings distributed via BGP the LSR SHOULD be able either (a) to reconstruct the same bindings as the LSR had prior to the restart (see Section 4), or (b) to create new <FEC -> label> bindings after restart, while temporarily maintaining MPLS forwarding state corresponding to both the bindings prior to the restart, as well as to the newly created bindings (see Section 5). Fourth, as long as the LSR retains the MPLS forwarding state that the LSR preserved across the restart, the labels from that state can not be used to create new local label bindings (but could be used to reconstruct the existing bindings, as per procedures in Section 4). Finally, for each next hop, if the next hop is reachable via a Label Switched Path (LSP), then the restarting LSR MUST be able to preserve the MPLS forwarding state associated with that LSP across the restart.

In the scenario where label binding on an LSR is created/maintained not just by the BGP component of the control plane, but by other protocol components as well (e.g., LDP, RSVP-TE), and the LSR supports restart of the individual components of the control plane that create/maintain label binding (e.g., restart of BGP, but no restart of LDP) the LSR MUST be able to preserve across the restart the information about which protocol has assigned which labels.

After the LSR restarts, it MUST follow the procedures as specified in [1]. In addition, if the LSR is able to preserve its MPLS forwarding state across the restart, the LSR SHOULD advertise this to its neighbors by appropriately setting the Flag for Address Family field in the Graceful Restart Capability for all applicable AFI/SAFI pairs.

3. Capability Advertisement

An LSR that supports the mechanism described in this document advertises this to its peer by using the Graceful Restart Capability, as specified in [1]. The Subsequent Address Family Identifier (SAFI) in the advertised capability MUST indicate that the Network Layer Reachability Information (NLRI) field carries not just addressing Information, but labels as well (see [RFC3107] as an example of where NLRI carries labels).

4. Procedures for the restarting LSR

Procedures in this section apply when a restarting LSR is able to reconstruct the same <FEC -> label> bindings as the LSR had prior to the restart.

The procedures described in this section are conceptual and do not have to be implemented precisely as described here, as long as the implementations support the described functionality and their externally visible behavior is the same.

Once the LSR completes its route selection (as specified in Section "Procedures for the Restarting Speaker" of [1]), then in addition to the procedures specified in [1], the LSR performs one of the following:

4.1. Case 1

The following applies when (a) the best route selected by the LSR was received with a label, (b) that label is not an Implicit NULL, and (c) the LSR advertises this route with itself as the next hop.

In this case the LSR searches its MPLS forwarding state (the one preserved across the restart) for an entry with <outgoing label, next hop> equal to the one in the received route. If such an entry is found, the LSR no longer marks the entry as stale. In addition if the entry is of type <incoming label, (outgoing label, next hop)> rather than <Forwarding Equivalence Class (FEC), (outgoing label, next hop)>, the LSR uses the incoming label from the entry when advertising the route to its neighbors. If the found entry has no incoming label, or if no such entry is found, the LSR allocates a new label when advertising the route to its neighbors (assuming that there are neighbors to which the LSR has to advertise the route with a label).

4.2. Case 2

The following applies when (a) the best route selected by the LSR was received either without a label, or with an Implicit NULL label, or the route is originated by the LSR, (b) the LSR advertises this route with itself as the next hop, and (c) the LSR has to generate a (non Implicit NULL) label for the route.

In this case the LSR searches its MPLS forwarding state for an entry that indicates that the LSR has to perform label pop, and the next hop equal to the next hop of the route in consideration. If such an entry is found, then the LSR uses the incoming label from the entry when advertising the route to its neighbors. If no such entry is found, the LSR allocates a new label when advertising the route to its neighbors.

The description in the above paragraph assumes that the LSR generates the same label for all the routes with the same next hop. If this is not the case, and the LSR generates a unique label per each such route, then the LSR needs to preserve across the restart not just <incoming label, (outgoing label, next hop)> mapping, but also the Forwarding Equivalence Class (FEC) associated with this mapping. In such case the LSR would search its MPLS forwarding state for an entry that (a) indicates Label pop (means no outgoing label), (b) the next hop equal to the next hop of the route and (c) has the same FEC as the route. If such an entry is found, then the LSR uses the incoming label from the entry when advertising the route to its neighbors. If no such entry is found, the LSR allocates a new label when advertising the route to its neighbors.

4.3. Case 3

The following applies when the LSR does not set BGP next hop to self.

In this case the LSR, when advertising its best route for a particular NLRI just uses the label that was received with that route. And if the route was received with no label, the LSR advertises the route with no label as well. Either way, the LSR does not allocate a label for that route.

5. Alternative procedures for the restarting LSR

In this section we describe an alternative to the procedures described in Section "Procedures for the restarting LSR".

Procedures in this section apply when a restarting LSR does not reconstruct the same <FEC -> label> bindings as the LSR had prior to the restart, but instead creates new <FEC -> label> bindings after restart, while temporarily maintaining MPLS forwarding state corresponding to both the bindings prior to the restart, as well as to the newly created bindings.

The procedures described in this section require that for the use by BGP graceful restart the LSR SHOULD have (at least) as many unallocated labels as labels allocated for the <FEC -> label> bindings distributed by BGP. The latter forms the MPLS forwarding state that the LSR managed to preserve across the restart. The former is used for allocating labels after the restart.

To create (new) local label bindings after the restart the LSR uses unallocated labels (this is pretty much the normal procedure).

The LSR SHOULD retain the MPLS forwarding state that the LSR preserved across the restart at least until the LSR sends End-of-RIB marker to all of its neighbors (by that time the LSR already completed its route selection process, and also advertised its Adj-RIB-Out to its neighbors). The LSR MAY retain the forwarding state even a bit longer (the amount of extra time MAY be controlled by configuration on the LSR), as to allow the neighbors to receive and process the routes that have been advertised by the LSR. After that, the LSR SHOULD delete the MPLS forwarding state that it preserved across the restart.

Note that while an LSR is in the process of restarting, the LSR may have not one, but two local label bindings for a given BGP route - one that was retained from prior to restart, and another that was created after the restart. Once the LSR completes its restart, the former will be deleted. Both of these bindings though would have the same outgoing label (and the same next hop).

6. Procedures for a neighbor of a restarting LSR

The neighbor of a restarting LSR (the receiving router in terminology used in [1]) follows the procedures specified in [1]. In addition, the neighbor treats the MPLS labels received from the restarting LSR the same way as it treats the routes received from the restarting LSR (both prior and after the restart).

Replacing the stale routes by the routing updates received from the restarting LSR involves replacing/updating the appropriate MPLS labels.

In addition, if the Flags in the Graceful Restart Capability received from the restarting LSR indicate that the LSR wasn't able to retain its MPLS state across the restart, the neighbor SHOULD immediately remove all the NLRI and the associated MPLS labels that it previously acquired via BGP from the restarting LSR.

An LSR, once it creates a binding between a label and a Forwarding Equivalence Class (FEC), SHOULD keep the value of the label in this binding for as long as the LSR has a route to the FEC in the binding. If the route to the FEC disappears, and then re-appears again later, then this may result in using a different label value, as when the route re-appears, the LSR would create a new <label, FEC> binding.

To minimize the potential mis-routing caused by the label change, when creating a new <label, FEC> binding the LSR SHOULD pick up the least recently used label. Once an LSR releases a label, the LSR SHALL NOT re-use this label for advertising a <label, FEC> binding to a neighbor that supports graceful restart for at least the Restart Time, as advertised by the neighbor to the LSR. This rule SHALL apply to any label release at any time.

7. Comparison between alternative procedures for the restarting LSR

Procedures described in Section 4 involve more computational overhead on the restarting router relative to the procedures described in Section 5.

Procedures described in Section 5 requires twice as many labels as the procedures described in Section 4.

Procedures described in Section 4 cause fewer changes to the MPLS forwarding state in the neighbors of the restarting router than the procedures described in Section 5.

In principle it is possible for an LSR to use procedures described in

Section 4 for some AFI/SAFI(s) and procedures described in Section 5 for other AFI/SAFI(s).

8. Security Consideration

The security considerations pertaining to the original BGP protocol remain relevant.

In addition, the mechanism described here renders LSRs that implement it vulnerable to additional denial-of-service attacks as follows:

> An intruder may impersonate a BGP peer in order to force a failure and reconnection of the TCP connection, but where the intruder sets the Forwarding State (F) bit (as defined in [1]) to 0 on reconnection. This forces all labels received from the peer to be released.
>
> An intruder could intercept the traffic between BGP peers and override the setting of the Forwarding State (F) bit to be set to 0. This forces all labels received from the peer to be released.

All of these attacks may be countered by use of an authentication scheme between BGP peers, such as the scheme outlined in [RFC2385].

As with BGP carrying labels, a security issue may exist if a BGP implementation continues to use labels after expiration of the BGP session that first caused them to be used. This may arise if the upstream LSR detects the session failure after the downstream LSR has released and re-used the label. The problem is most obvious with the platform-wide label space and could result in mis-routing of data to other than intended destinations and it is conceivable that these behaviors may be deliberately exploited to either obtain services without authorization or to deny services to others.

In this document, the validity of the BGP session may be extended by the Restart Time, and the session may be re-established in this period. After the expiry of the Restart Time the session must be considered to have failed and the same security issue applies as described above.

However, the downstream LSR may declare the session as failed before the expiration of its Restart Time. This increases the period during which the downstream LSR might reallocate the label while the upstream LSR continues to transmit data using the old usage of the label. To reduce this issue, this document requires that labels are not re-used until for at least the Restart Time.

9. Intellectual Property Considerations

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.

10. Copyright Notice

11. Acknowledgments

We would like to thank Chaitanya Kodeboyina and Loa Andersson for their review and comments. The approach described in Section "Alternative procedures for the restarting LSR" is based on the idea suggested by Manoj Leelanivas.

12. Normative References

[1] "Graceful Restart Mechanism for BGP", work in progress

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

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC2385

[RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", RFC2026

13. Non-normative References

[RFC3107] Rekhter, Y., Rosen, E., "Carrying Label Information in BGP-4", RFC3107

14. Author Information

Yakov Rekhter
Juniper Networks
1194 N.Mathilda Ave
Sunnyvale, CA 94089
e-mail: yakov@juniper.net

Rahul Aggarwal
Juniper Networks
1194 N.Mathilda Ave
Sunnyvale, CA 94089
e-mail: rahul@juniper.net