Network Working Group	J. Uttaro
Internet-Draft	AT&T
Intended status: Standards Track	A. Simpson
Expires: April 22, 2012	Alcatel-Lucent
	R. Shakir
	C&W
	C. Filsfils
	P. Mohapatra
	Cisco Systems
	B. Decraene
	France Telecom
	J. Scudder
	Y. Rekhter
	Juniper Networks
	October 20, 2011

BGP Persistence

draft-uttaro-idr-bgp-persistence-00

Abstract

For certain AFI/SAFI combinations it is desirable that a BGP speaker be able to retain routing state learned over a session that has terminated. By maintaining routing state forwarding may be preserved. This technique works effectively as long as the AFI/SAFI is primarily used to realize services that do not depend on exchanging BGP routing state with peers or customers. There may be exceptions based upon the amount and frequency of route exchange that allow for this technique. Generally the BGP protocol tightly couples the viability of a session and the routing state that is learned over it. This is driven by the history of the protocol and it's application in the internet space as a vehicle to exchange routing state between administrative authorities. This document addresses new services whose requirements for persistence diverge from the Internet routing point of view.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

In certain scenarios, a BGP speaker may maintain forwarding in spite of BGP session termination. Currently all routing state learned between two speakers is flushed upon either normal or abnormal session termination. There are techniques that are useful for maintaining routing when a session abnormally terminates i.e BGR Graceful RestartR (RFC 4724) or normal termination such as increasing timers but they do not change the fundamental problem. The technique of BGP persistence works effectively as long as the expectation is that there is a decoupling of session viability and the correct service delivery, and the delivery uses the routing state learned over that session. This document proposes a modification to BGP's behavior by enabling persistence of BGP learned routing state in spite of normal or abnormal session termination.

1.1. Requirements Language

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].

2. Communities

This memo defines three new communities that are used to identify the capability of a path to persist and whether or not that path is live or stale.

2.1. PERSIST

This memo defines a new transitive BGP community, PERSIST, with value TBD (to be assigned by IANA). Attaching of the PERSIST community SHOULD be controlled by configuration. Attaching the PERSIST community indicates that the peer should maintain forwarding in the case of a session failure. The functionality SHOULD default to being disabled.

2.2. DO_NOT_PERSIST

This memo defines a new transitive BGP community, DO_NOT_PERSIST, with value TBD (to be assigned by IANA). Attaching of the DO_NOT_PERSIST community SHOULD be controlled by configuration. The functionality SHOULD default to being disabled.

2.3. STALE

This memo defines a new transitive BGP community, STALE, with value TBD (to be assigned by IANA). Attaching of the STALE community is limited to a path that currently has the PERSIST community attached

3. Configuration (Persistence Timer, PERSIST and DO_NOT_PERSIST Community)

Persistence must be configured on a per session basis. A speaker configures the ability to persist independently of it's peer. There is no negotiation between the peers. A timer must be configured indicating the time to persist stale state from a peer where the session is no longer viable. This timer is designated as the persist-timer. A speaker must also attach persistence community value indicating if a path to a route should persist.

3.1. Settings for Different Applications

The setting of the persist-timer should be based upon the field of use. BGP is used in a many different applications that each bring a unique requirement for retaining state. The following is not meant as a comprehensive listing but to suggest timer settings for a subset of AFI/SAFIs.

L2VPN This AFI/SAFI requires the exchange of routing state in order to establish PWs to realize a VPLS VPN, or a VPWS PW. This AFI/SAFI does not require exchange of routing state with a customer and there is no eBGP session established. The persist-timer should be set to a large value on the order of days to infinity.

L3VPN This AFI/SAFI requires the exchange of routing state to create a private VPN. This AFI/SAFI requires exchange of state with customers via eBGP and is dynamic. The SP needs to consider the possibility that stale state may not reflect the latest route updates and therefore may be incorrect from the customer perspective. The persist-timer should be set to a large value on the order of hours to a few days. this is built upon the notion some incorrectness is preferable to a large outage.

4. Operation

Assuming a session failure has occurred a BGP persistent router must retain local forwarding state for those paths that are Persistent/Stale and propagate paths to downstream speakers that indicate that a given path is now stale.

4.1. Attaching the STALE Community Value and Propagation of Paths

The following rules must be followed.

- *Identify paths learned over a failed session that have the PERSIST capable community value attached.
- *For those paths attach the STALE community value and propagate to all peers.
- *For those paths learned over the failed session that do not have PERSIST capable community value or are marked with the DO_NOT_PERSIST community follow BGP rules and generate withdrawals to all peers for those paths.

4.2. Forwarding

The following rules must be followed to ensure valid forwarding:

- *All forwarding state must be retained i.e labels for BGP labeled unicast.
- *Forwarding must ensure that the Next Hop to a "stale" route is viable.
- *Forwarding to a "stale" route is only used if there are no other paths available to that route. In other words an active path always wins regardless of path selection. "Stale" state is always considered to be less preferred when compared with an active path.
- *Forwarding should be retained through an advertisement. When the session is re-established forwarding should only change if the new state is either different or better in terms of path selection. A make before break strategy should be employed.
- *Stale state may be retained indefinitely or may be programmed to expire via configuration.
- *The Receiving Speaker MUST replace the stale routes by the routing updates received from the peer. Once the End-of-RIB marker for an address family is received from the peer, it MUST immediately remove any paths from the peer that are still marked as stale for that address family.
- *There is no restriction on whether the session is internal or external.

4.3. Example Behaviour

Upon session establishment a speaker S2 may receive paths from S1 that are marked with PERSIST, DO_NOT_PERSIST or neither. Assume S2 is also peered with a downstream speaker S3.. Implementations MUST follow the specifications outlined below for.
Upon recognition of the failure to S1, S2 will identify paths that had been marked with PERSIST, DO_NOT_PERSIST or neither learned from S1. S2 MUST implement the following behavior:

```
if ( P1 is tagged with PERSIST ) {

Retain Forwarding
  Attach the STALE Community to all paths that were marked with PERSIST
  Advertise STALE paths to all peers including S3
}
else ( P1 is marked with DO_NOT_PERSIST || not marked )

Tear down the forwarding structure for P1
Follow normal BGP rules i.e Best path, withdrawal etc.

fi
```

5. Deployment Considerations

BGP Persistence as described in this document is useful within a single autonomous system or across autonomous systems.

6. Applications

This technique may be useful in a wide array of applications where routing state is either fairly static or, the state is localized within a routing context. Some applications that come immediately to mind are L2 and L3 VPN.

6.1. Persistence in L2VPN (VPLS/VPWS)

VPLS/VPWS VPNs use BGP to exchange routing state between two PEs. This exchange allows for the creation of a PW within a VPN context between those PEs. By definition, L2VPN does not exchange any routing state with customers via BGP. BGP persistence is very useful here as the state is quite constant. The only time state is exchanged is when a PW endpoint is provisioned, deleted or when a speaker reboots.
Referring to Figure 2, PE1 and PE2 have advertised BGP routing state in order to create PWs between PE1 and PE2. The RRs are only responsible to reflect this state between the PEs. The use of a unique RD makes every path unique from the RRs perspective.
Assume that the both RR experience catastrophic failure.
Case 1 - All BGP speakers are persistent capable.

The PWs created between PE1 and PE2 persist. Forwarding uninterrupted.
Case 2 - PE1 and the RRs are persistent capable, PE2 is not.
In this case the path advertised from PE2 via the RRs is persistent at PE1, the PW from PE1 to PE2 is not torn down. PE2 will remove the path from PE1 and tear down the PW from PE2 to PE1. THe effect is that MAC state learned at PE2 is valid as the PW is still valid. MAC state learned at PE1 is removed as the PW is no longer valid. Eventually MAC destinations recursed to the PW at PE1 destined for PE2 over the valid PW will time out.
Assume that the RRs are valid but the iBGP sessions are torn down..
Case 3 - All BGP speakers are persistent capable.
The PWs created between PE1 and PE2 persist. Forwarding uninterrupted.

```
VPNA                 VPNA
 PW+++++++++++++++++++PW

CE1-------PE1--------RR1-------PE2------CE2
  |                    |
  |                    |
  ----------RR2---------

  <--iBGP---><---iBGP-->
```

6.2. Persistence in L3VPN

```
             --------RR1-------
            / A               C \
CE1 ----- PE1 --Forwarding Path-- PE2 ---- CE2
            \ B               D /
             ------- RR2 ------
```

In the case of a Layer 3 VPN topology, during the failure of a route reflector device at the current time, all routing information propagated via BGP is purged from the routing database. In this case, forwarding is interrupted within such a topology due to the lack of signalling information, rather than an outage to the forwarding path between the PE devices. With the addition of BGP persistence, a complete service outage can be avoided.
The topology shown in Figure 3 is a simple L3VPN topology consisting of two customer edge (CE) devices, along with two provider edge (PE), and route reflector (RR) devices. In this case, where an RFC4364 VPN topology is utilised a BGP session exists between PE1 to both RR1 and RR2, and from PE2 to RR1 and RR2, in order to propagate the VPN topology.
Case 1: No BGP speakers are persistence capable:

- *In this scenario, during a simultaneous failure of RR1 and RR2 (which are extremely likely to share route reflector clients) both PE1 and PE2 remove all routing information from the VPN from their RIB, and hence a complete service outage is experienced.

- Where either sessions A and B, or C and D fail simultaneously, routing information from either PE1 (in the case of A and B), or PE2 (in the case of C and D) are withdrawn, and a partial service topology exists.

- Both of the states described reflect a service outage where the forwarding path between the PE devices is not interrupted.

Case 2: All BGP speakers are persistence capable:

- PE1 continues to forward utilising the label information received from PE2 via the working forwarding path for the duration of the persistence timer (and vice versa).

- This condition occurs regardless of the session(s) that fail. In the worst case where sessions A, B, C and D fail simultaneously, the network continues to operate in the state in which it was at the time of the failure.

Case 3: PE1 and RR[12] are persistence capable - PE2 is not.

- During a failure of BGP session A or B, PE1 will continue to forward utilising the routing information received from the RRs for PE2 for the duration of the persistence timer. PE2 will continue to forward utilising the routing information received from the RRs, again for the duration of the persistence timer.

- In the case that either BGP session C or D fails, all routes will be withdrawn by RR[12] towards PE1 since these routes are not valid to be persisted by the RRs. The end result of this will be that the routes advertised by CE2 into the VPN will be withdrawn.

- Where the worst case failure occurs (i.e. sessions A, B, C and D fail) the routes advertised by CE1 into the VPN will be persistently advertised by the RR devices, whereas those advertised by CE2 will be withdrawn. Clearly in the example shown in the figure this results in a service outage, but where multiple PE devices exist within a topology, service is maintained for the subset of CEs attached to PE devices supporting the persistence capability.

Within the Layer 3 VPN deployment it should be noted that routing information is less static than that of the many Layer 2 VPNs since typically multiple routes exist within the topology rather than an individual MAC address or egress interface per CE device on the PE device. As such, the L3VPN operates with the routing databases in the 'core' of the network reflecting those at the time of failure. Should there be re-convergence for any path between the PE and CE devices, this will result in invalid routing information, should the egress PE device not hold alternate routing information for the prefixes undergoing such re-convergence. It is expected that where each PE maintains multiple paths to each egress prefix (where an alternate path is available), it is expected that the egress PE will forward packets towards an alternative egress PE for the prefix in question where the topology is no longer valid.

The lack of convergence within a Layer 3 topology during the persistent state SHOULD be considered since it may adversely affect services, however, an assumption is made that a degraded service is preferable to a complete service outage during a large-scale BGP control plane failure.

7. Security Considerations

The security implications of the persistence mechanism defined within in this document are akin to those incurred by the maintenance of stale routing information within a network. This is particularly relevant when considering the maintenance of routing information that is utilised for service segregation - such as MPLS label entries.

For MPLS VPN services, the effectiveness of the traffic isolation between VPNs relies on the correctness of the MPLS labels between ingress and egress PEs. In particular, when an egress PE withdraws a label L1 allocated to a VPN1 route, this label MUST not be assigned to a VPN route of a different VPN until all ingress PEs stop using the old VPN1 route using L1.

Such a corner case may happen today, if the propagation of VPN routes by BGP messages between PEs takes more time than the label re-allocation delay on a PE. Given that we can generally bound worst case BGP propagation time to a few minutes (e.g. 2-5), the security breach will not occur if PEs are designed to not reallocate a previous used and withdrawn label before a few minutes.

The problem is made worse with BGP GR between PEs as VPN routes can be stalled for a longer period of time (e.g. 20 minutes).

This is further aggravated by the BGP persistent extension proposed in this document as VPN routes can be stalled for a much longer period of time (e.g. 2 hours, 1 day).

Therefore, to avoid VPN breach, before enabling BGP persistence, SPs needs to check how fast a given label can be reused by a PE, taking into account:

- The load of the BGP route churn on a PE (in term of number of VPN label advertised and churn rate).

- The label allocation policy on the PE (possibly depending upon the size of pool of the VPN labels (which can be restricted by hardware consideration or others MPLS usages), the label allocation scheme (e.g. per route or per VRF/CE), the re-allocation policy (e.g. least recently used label...)

In addition to these considerations, the persistence mechanism described within this document is considered to be complex to exploit maliciously - in order to inject packets into a topology, there is a requirement to engineer a specific persistence state between two PE devices, whilst engineering label reallocation to occur in a manner that results in the two topologies overlapping. Such allocation is particularly difficult to engineer (since it is typically an internal mechanism of an LSR).

8. IANA Considerations

IANA shall assigned community values from BGP well-known communities registry for the PERSIST, DO-NOT-PERSIST and STALE communities. No additional IANA action is required.

9. Acknowledgements

We would like to acknowledge Roberto Fragassi (Alcatel-Lucent), John Medamana, (AT&T) Han Nguyen (AT&T), Jeffrey Haas (Juniper), Nabil Bitar (Verizon), Nicolai Leymann (DT) for their contributions to this document.

10. References

[RFC1997]	Chandrasekeran, R., Traina, P. and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.
[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
[RFC4271]	Rekhter, Y., Li, T. and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
[RFC4364]	Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

Authors' Addresses

James Uttaro Uttaro AT&T 200 S. Laurel Avenue Middletown, NJ 07748 USA EMail: ju1738@att.com

Adam Simpson Simpson Alcatel-Lucent 600 March Road Ottawa, Ontario K2K 2E6 Canada EMail: adam.simpson@alcatel-lucent.com

Rob Shakir Shakir Cable&Wireless Worldwide London, UK EMail: rjs@cw.net URI: http://www.cw.com/

Clarence Filsfils Filsfils Cisco Systems Brussels, 1000 BE EMail: cf@cisco.com

Pradosh Mohapatra Mohapatra Cisco Systems 170 W. Tasman Drive San Jose, CA 95134 USA EMail: pmohapat@cisco.com

Bruno Decraene Decraene France Telecom 38-40 Rue de General Leclerc 92794 Issy Moulineaux cedex 9 France EMail: bruno.decraene@orange.com

John Scudder Scudder Juniper Networks 1194 N. Mathilda Ave Sunnyvale, CA 94089 USA EMail: jgs@juniper.net

Yakov Rekhter Rekhter Juniper Networks EMail: yakov@juniper.net