Internet Engineering Task Force
Internet Draft
Updates: 4271 (if approved)
Intended status: Standards Track
Expires: February 13, 2022

E. Chen
J. Yuan
Palo Alto Networks

August 12, 2021

Deterministic Route Redistribution into BGP
draft-chen-bgp-redist-03.txt

Abstract

In this document we present several examples of non-deterministic routing behavior involving route redistribution into BGP. In order to eliminate such non-deterministic behavior, we propose an enhancement to BGP route selection that would take into account the administrative distance under certain conditions. We also recommend that the LOCAL_PREF value be reduced for the redistributed backup route, and be calculated automatically based on the administrative distance.

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on February 13, 2022.

1. Introduction

A routing protocol usually downloads its best (or active) route to the routing table, also known as Routing Information Base (RIB), which in turn selects the best (or active) route to program the forwarding table.

When comparing routes from different routing protocols, RIB typically uses the "administrative distance" [ADMIN-DIS] (abbreviated as "admin-distance" hereafter) as the tie breaker. The convention is that a route with a lower admin-distance is more preferred, and that is assumed in this document when specific admin-distance values are given as examples. The admin-distance associated with a route in RIB is commonly used to implement various routing schemes such as designating primary and backup routes in a network.

On the other hand, the route selection in BGP [RFC4271] involves comparing the LOCAL_PREF, AS_PATH and other BGP attributes. The bestpath in BGP usually becomes the candidate for downloading to the RIB, and for advertising to BGP neighbors.

It is common to redistribute routes from other routing protocols (such as "static routing" [STATIC-R]) into BGP for route propagation. This topic is briefly discussed in [Sect. 9.4, RFC4271]. A redistributed route is usually assigned a fixed LOCAL_PREF value, and has an empty AS_PATH attribute.

The interaction between RIB and BGP follows these general rules:

- A local route may be redistributed into BGP only if it is active in RIB based on the admin-distance.

- Only the bestpath in BGP is downloaded to RIB.

Currently the admin-distance does not play any role in BGP route selection. Due to the lack of such correlation between RIB and BGP, when a backup route (based on the admin-distance) is redistributed into BGP as shown in the next section, routing may converge to different paths depending on the order of path arrivals. Such non-deterministic routing behavior is clearly detrimental to network operations.

In order to eliminate such non-deterministic behavior, we propose an enhancement to BGP route selection that would take into account the admin-distance under certain conditions. We also recommend that the LOCAL_PREF value be reduced for the redistributed backup route, and be calculated automatically based on the admin-distance.

The proposed enhancement and recommendation are backward compatible, and can be deployed on an individual router basis.

Although the static routing is used as examples in the document, the proposed enhancement and recommendation also apply when a route is redistributed from other routing protocols into BGP.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. The Problem

In this section several examples are presented to illustrate the non-deterministic routing behavior involving route redistribution into BGP.

2.1. On a Single Router

Consider an example in which there are two paths for the same destination on a single router. As shown in the following table, the primary path A is received from an external BGP neighbor, and the backup path B is a static route and is configured for redistribution into BGP.

Path	Type	Admin_Distance	LOCAL_PREF	AS_PATH
A	EBGP	20	100	65535
B	Static	150	100	--

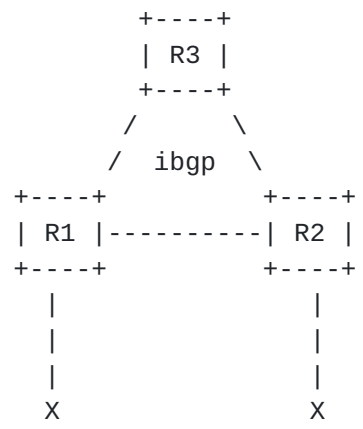
Depending on the order of path arrivals, the path that arrives first would be selected as the bestpath in both RIB and BGP.

More specifically, if Path A is received in BGP and is downloaded to RIB first, it would remain as the best in RIB (due to the admin-distance) even when Path B shows up in RIB later. In this case Path A would be the best one in both RIB and BGP.

If Path B shows up in RIB and is redistributed into BGP first, it would remain as the best in BGP (due to it being a local route or with a shorter AS-PATH) even when Path A is received in BGP later. In this case Path B would be the best one in both RIB and BGP.

2.2. Network-wide Behavior

Consider the following example in which Routers R1, R2 and R3 are part of a provider network and IBGP sessions are maintained among them. There are two customer connections, a primary connection on R1 and a backup connection on R2. The customer route X is statically routed on both R1 and R2, and is redistributed into BGP. On R2, the backup path for X is configured with a less preferred admin-distance than the one for IBGP paths.



While R1 consistently selects the local static route as the best one, the route selection on R2 would be non-deterministic. As shown in the following figure, there are potentially two BGP paths A and B for X on R2, with Path A learned from R1 and Path B locally redistributed.

Path	Type	Admin_Distance	LOCAL_PREF	AS_PATH
A	IBGP	200	100	--
B	Static	210	100	--

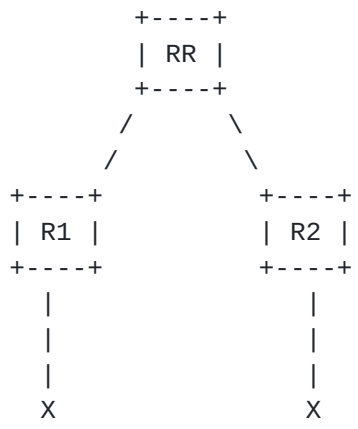
Depending on the order of arrivals of these two paths, the path that arrives first would be selected as the bestpath in both RIB and BGP.

More specifically, if Path A is received in BGP and is downloaded to RIB first, it would remain as the best in RIB (due to the admin-distance) even when Path B shows up in RIB later. In this case A would be the best one in both RIB and BGP.

If Path B shows up in RIB and is redistributed into BGP first, it would remain as the best in BGP (due to it being a local route or with a lower IGP metric) even when Path A is received in BGP later. In this case Path B would be the best one in both RIB and BGP.

The non-deterministic route selection on R2 may cause other nodes (like R3) to converge to different paths as well. The routing behavior in the network would be non-deterministic, and inconsistent with the intended routing design.

A network using BGP route reflection [RFC4456] (or BGP confederation [RFC5065]) may experience additional cases of network-wide "non-deterministic" routing behavior. For example in the following figure, when both R1 and R2 advertise their respective local routes to the route reflector (RR) simultaneously, the RR would use the "IGP metric" to choose the bestpath between the two IBGP paths. As a result the network may or may not converge to the primary path.



3. The Proposed Solution

In order to eliminate the non-deterministic routing behavior involving route redistribution into BGP, we propose an enhancement to BGP route selection that would take into account the admin-distance under certain conditions. We also recommend that the LOCAL_PREF value be reduced for the redistributed backup route, and calculated automatically based on the admin-distance.

3.1. Enhancement to BGP Route Selection

To make it deterministic on a single router regarding the route being sourced and advertised to the network, we propose that the following procedure be added prior to the step that compares the degrees of preference of routes and identifies the route that has the highest degree of preference, as described in Sect. 9.1.2 [RFC4271] for BGP route selection:

> When comparing a locally redistributed route with another route that is either locally aggregated or received from an external neighbor, favor the one with a more preferred admin-distance. The admin-distance for a BGP route is obtained as follows:
>
> > For a locally redistributed route, it is inherited from the route being redistributed from RIB.
> >
> > For a non-redistributed route, it is of the same value as the admin-distance assigned to the route for the purpose of RIB installation (regardless of whether it is actually installed in RIB).

It should be noted that IBGP paths are deliberately excluded from the algorithm. As the admin-distance is not propagated by BGP, involving IBGP paths in the admin-distance comparison can easily result in unintended routing behavior and even route churns. To influence route selection in a network, use the LOCAL_PREF attribute as described in the next section.

3.2. Setting the LOCAL_PREF Value

When a non-BGP route is designated as a backup route in the network, it should be assigned a less preferred admin-distance than the value for IBGP routes. When such a route is redistributed into BGP, the LOCAL_PREF value for the redistributed route SHOULD be set lower than the LOCAL_PREF values of the primary route and other more preferred routes.

Assuming the default LOCAL_PREF value is assigned to the primary route, then the LOCAL_PREF value for the redistributed backup route can be calculated automatically as described by the following pseudo-code:

```
if (redist_admin_distance > ibgp_admin_distance) {
    offset = redist_admin_distance - ibgp_admin_distance;
    if (default_local_pref > offset)
        calculated_local_pref = default_local_pref - offset;
    else
        calculated_local_pref = 0;
}
```

in which

- "redist_admin_distance" is the admin-distance of the route being redistributed.
- "ibgp_admin_distance" is the admin-distance for IBGP routes on the local router.
- "default_local_pref" is the default LOCAL_PREF value in the network.
- "calculated_local_pref" is the calculated LOCAL_PREF value for the redistributed route.

Clearly, in order for the calculated LOCAL_PREF value to truly reflect the intended routing design, the admin-distance needs to be assigned properly. Guideline is provided on assigning the admin-distance in the next section.

This algorithm would not apply if the "default_local_pref" is not assigned to the primary route, in which case manual configuration should be used.

In addition to lowering the LOCAL_PREF value, it may be necessary to modify the parameters for the aforementioned redistributed route pertaining to any vendor-specific route selection criteria preceding the LOCAL_PREF comparison. For example, the "weight" parameter exists in a number of implementations in which case the "weight" for the aforementioned redistributed route should be made equal to the default "weight" for IBGP routes.

3.3. Admin-distance Assignment

In order to achieve the desired routing scheme using the LOCAL_PREF calculated from the admin-distance, coordination would be necessary for the admin-distance assignment when the same destination is redistributed from multiple routers in a network.

While the default LOCAL_PREF value is usually consistent in a network, the default admin-distance for IBGP routes can vary from one node to another in a multi-vendor network.

The coordination of the admin-distance assignment can be simplified by examining the "role" that a non-BGP route is supposed to play (such as being the primary, the secondary or the tertiary), and then associate an "offset" to the route based on its role. Among the routes involved, the less preferred a route is, the higher the offset should be. Then the admin-distance for the route can be assigned as (ibgp_admin_distance + offset), and the desired LOCAL_PREF value would be automatically calculated using the algorithm described in the previous section.

As an example shown in the following table, there are three non-BGP paths for the same destination on separate routers A, B and C in the network and they are designated as the primary, the secondary and the tertiary. The default LOCAL_PREF value is 100 in the network, and the "ibgp_admin_distance" is 200 on the router with the secondary path, and 170 on the router with the tertiary path.

The desired LOCAL_PREF values for the redistributed routes are obtained using the algorithm and procedures described in this document.

Router	Role	Offset	Admin_Distance	LOCAL_PREF
A	Primary	-	50	100 (Default)
B	Secondary	10	200 + 10	100 - 10
C	Tertiary	20	170 + 20	100 - 20

3.4. Configuration Option

Configuration can be used to achieve the equivalent outcome by setting the appropriate LOCAL_PREF value (and also the "weight" parameter if applicable) for the redistributed backup route. It can also be used to override the LOCAL_PREF value calculated based on the admin-distance value of the redistributed route as proposed in this document.

When route redistribution is part of a more complex routing scheme beyond what can be automated with the proposed solution, configuration can also be used following the general principles discussed in this document.

4. IANA Considerations

This document has no request for IANA.

5. Security Considerations

The solution proposed in this document does not change the underlying security or confidentiality issues inherent in the existing BGP [RFC4271].

6. Acknowledgments

The authors would like to thank Naiming Shen, Acee Lindem and Robert Raszuk for inputs and discussions.

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

7.2. Informative References

[STATIC-R] Static routing, Wikipedia, https://en.wikipedia.org/wiki/Static_routing

[ADMIN-DIS] Administrative distance, Wikipedia, https://en.wikipedia.org/wiki/Administrative_distance.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <http://www.rfc-editor.org/info/rfc4456>.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, DOI 10.17487/RFC5065, August 2007, <http://www.rfc-editor.org/info/rfc5065>.

Authors' Addresses

Enke Chen
Palo Alto Networks

Email: enchen@paloaltonetworks.com

Jenny Yuan
Palo Alto Networks

Email: jyuan@paloaltonetworks.com