Advertising Equal Cost Multi-Path (ECMP) routes in BGP

draft-bhatia-ecmp-routes-in-bgp-00.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

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. Internet Drafts may be updated, replaced, or obsoleted by other documents at any time. It is not appropriate to use Internet Drafts as reference material or to cite them other than as a "working draft" or "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.

Abstract

This document describes an extensible mechanism that will allow a BGP [BGP4] speaker to advertise equal cost multi-path (ECMP) routes for a destination to its peers without changing the semantics of the UPDATE message.

A new BGP attribute is introduced that will be used to advertise the multiple next hops for the feasible and the un-feasible ECMP BGP routes to the remote peers.

The mechanisms described in this document are applicable to all routers, both those with the ability to inject multiple routing entries in their forwarding table and those without (although the latter need not implement some extensions described in this document).
1. Motivation

The BGP specification allows only one "best" route to be inserted into its Loc-RIB and to be announced to other BGP speakers. If another route with the same NLRI is announced then it is taken as an implicit withdraw of the previous one. This creates some problems and BGP speakers are thus never able to advertise equal cost multi-path routes to their peers.

The maximum that most of the current implementations do when they receive multiple equal cost BGP routes is to insert all of them (or a subset of them based on their local policies) in their forwarding table and locally do load balancing for the destination, while announcing just one "best" path to their peers. The "best" path selection could be either based on the lower Router ID or the route which has been received first. Selecting the best path based on the Router ID is deterministic and can cause MED churn [BGP-MED] in some topologies while the latter selection criterion is known to be non-deterministic.

This document modifies the Phase 2 and the Phase 3 of the Decision Process to select multiple best routes out of all those available for each distinct destination to be installed in the Loc-RIB and for disseminating multiple routes for one destination in the Loc-RIB to its peers.

The idea is to introduce minimal changes in the BGP protocol to accommodate support for ECMP BGP routes.

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 [KEY-WORDS].

2. Operation

In the following sections, "Local speaker" refers to a router which is advertising these ECMP routes, and the "Receiving Speaker" refers to a router that peers with the former router to accept multiple BGP routes for a destination.

Consider that a BGP session between the Local Speaker and the Receiving Speaker is established. The following sections explain how the Local Speaker may advertise multiple BGP routes to the Receiving Speaker without the latter replacing the routes recieved from the former peer previously.
3. Procedures for the Local Speaker.

When the Local speaker receives multiple routes to the same destination from different (or the same, in case this extension is implemented) peers then it runs its decision process to select the best BGP routes that will be injected into its Loc-Rib table and those that will be advertised to its peers.

Section 9.1.2.2 of [BGP-4] explains the tie breaking procedure for selecting only one of the routes, from the multiple routes present in Adj-Ribs-In, for inclusion in the associated Loc-Rib.

This document modifies this algorithm to support inclusion of multiple routes in the Loc-RIB and subsequently, advertisement of multiple ECMP routes to the peers. The changes introduced are as follows:

After the step (e) in sec 9.1.2.2 whatever candidate BGP routes exist are all considered for inclusion in the Loc-RIB and are announced to the remote BGP speaker supporting this capability.

4. Advertisement of ECMP BGP routes

To provide backward compatibility, as well as to simplify introduction of the ECMP capabilities into BGP, a new BGP attribute, Equal Cost Multi-Path Next Hop (ECMP_NEXT_HOP) is introduced.

This will be used in addition to the existing NEXT_HOP attribute for announcing multiple next-hops to the destinations listed in the Network Layer Reachability Information of the UPDATE message.

The ECMP_NEXT_HOP attribute is kept as optional and non-transitive so that BGP speakers that dont support the ECMP capability will simply ignore the information carried in this attribute, and will not pass it to other BGP speakers.

All prefixes announced using this attribute will not replace the previous advertisement and thus a prefix can be advertised multiple times by the Local Speaker.

If the same prefix is announced by using the NEXT_HOP attribute only then it is taken as an implicit withdraw for all the previous entries advertised by that peer for those destinations listed.

An UPDATE message that contains feasible routes and carries ECMP_NEXT_HOP and no NEXT_HOP attribute will not be considered as implicit withdrawals. The Receiving Speaker will simply add these routes in its Adj-RIBs-In as multiple routes to that destination.
If some of the attributes for one of the ECMP BGP route changes (e.g. IGP cost to reach the next-hop) and it is no longer the preferred route then an implementation MUST send an explicit withdrawal for that particular route.

5. Equal Cost Multi-Path Next Hop - ECMP_NEXT_HOP (Type Code: TBD)

This is an optional non-transitive attribute that can be used for advertising the multiple next-hops associated with a NLRI.

The attribute contains one or more triples <Address Family Information, Next Hop Information>, where each triple is encoded as shown below:

```
+---------------------------------------------------+
| Address Family Identifier (2 octets)              |
+---------------------------------------------------+
| Number of Next Hops (1 octet)                     |
+---------------------------------------------------+
| Length of the First Next Hop (1 octet)            |
+---------------------------------------------------+
| Network Address of First Next Hop (variable)      |
+---------------------------------------------------+
| Length of the Second Next Hop (1 octet)           |
+---------------------------------------------------+
| Network Address of Second Next Hop (variable)     |
+---------------------------------------------------+
| . . .                                             |
+---------------------------------------------------+
| Length of the Nth Next Hop (1 octet)              |
+---------------------------------------------------+
| Network Address of Nth Next Hop (variable)        |
+---------------------------------------------------+
```

The use and meaning of these fields are as follows:

Address Family Identifier:

This field carries the identity of the Network Layer protocol associated with the Network Address that follows. Presently defined values for this field are specified in RFC1700.

Number of Next-Hops:

This field carries a number one less than the total number of ECMP BGP routes for the given NLRI.
Length of Nth Next Hop Network Address:

A 1 octet field whose value expresses the length of the "Network Address of Next Hop" field as measured in octets. For IPv6 routes the value shall be set to 16, when only a global address is present, or 32 if a link-local address is also included in the Next Hop field [BGP-IPv6].

Network Address of Nth Next Hop:

A variable length field that contains the Network Address of the next router on the path to the destination.

The N next-hops listed in the ECMP_NEXT_HOP path attribute defines the Network Layer address of the routers that should be used as next-hop to the destinations listed in the UPDATE message. The N+1th next-hop is carried in the NEXT_HOP attribute.

6. Procedures for the Receiving Speaker

The Receiving Speaker upon receiving the ECMP_NEXT_HOP attribute will understand that the Local Speaker has advertised ECMP BGP routes. It will accept all the routes and they will all be exactly the same except for the next-hop which will be different for each one of them. It will run the modified decision process as explained in the Section 4 and depending upon the result will either

- inject multiple routes into Local-RIB and advertise multiple paths to its peers
- OR
- inject a single prefix which has better path attributes than the ECMP routes

If the Receiving Peer receives some withdrawn routes along with the ECMP_NEXT_HOP attribute then it shall understand that some of the previously advertised ECMP BGP have been removed and an implementation MUST proceed with removing all such paths.

If a peer wants to withdraw all the ECMP BGP routes then it can send a normal BGP UPDATE message listing the NLRI in the WITHDRAWN Routes field. An implementation should then remove all the paths which it had previously received from the Local Speaker for this NLRI.

If the Receiving Speaker receives an UPDATE message with the ECMP_NEXT_HOP attribute which contains both the feasible and the unfeasible routes then it should consider these attributes for the feasible routes. All the destinations listed in the withdrawn routes shall be removed as per [BGP4].
7. Configuring BGP ECMP Support

An implementation MUST provide a configuration option to set and unset this feature irrespective of whether it is capable of injecting multiple routes into its Loc-RIB or not.

It is recommended to advertise BGP ECMP routes to the peers even if the Local Speaker cannot insert multiple entries in its forwarding table. This way it can help its other peers to make better routing decisions.

The default configuration for this option MUST be to announce ECMP BGP routes. However there can be cases when a Local Speaker may not choose to announce such routes, e.g. memory constraints on the remote router with a low amount of memory and especially when its carrying full Internet routing table.

8. Security Considerations

This document introduces no new security concerns to BGP or other specifications referenced in this document.

9. Acknowledgements

The author would like to thank Curtis Villamizar for his valuable comments and suggestions.

10. IANA Considerations

This document uses an attribute type to indicate additional next-hops for the BGP paths. This must be assigned by IANA as per RFC 2842.

11. References


12. Author's Address

Manav Bhatia
Network Systems Division,
Samsung India Software Operations,
Bangalore
INDIA
Email: manav@samsung.com