Internet Engineering Task Force I. Varlashkin
Internet-Draft Easynet Global Services
Intended status: Standards Track R. Raszuk
Expires: September 28, 2012 NTT MCL Inc.
March 27, 2012

Carrying next-hop cost information in BGP
draft-ietf-idr-bgp-nh-cost-01

Abstract

This document describes new BGP SAFI to exchange cost information to next-hops for the purpose of calculating best path from a peer perspective rather than local BGP speaker own perspective.

Table of Contents

1. Motivation

In certain situation route-reflector clients may not get optimum path to certain destinations. ADDPATH solves this problem by letting route-reflector to advertise multiple paths for given prefix. If number of advertised paths sufficiently big, route-reflector clients can choose same route as they would in case of full-mesh. This approach however places additional burden on the control plane. Solutions proposed by [BGP-ORR] use different approach - instead of calculating best path from local speaker own perspective the calculations are done using cost from the client to the next-hops. Although they eliminate need for transmitting redundant routing information between peers, there are scenarios where cost to the next-hop cannot be obtained accurately using this methods. For example, if next-hop information itself has been learned via BGP then simple SPF run on link-state database won't be sufficient to obtain cost information. To address such scenarios this document proposes a solution where cost information to the next-hops is carried within BGP itself using dedicated SAFI.

2. NEXT-HOP INFORMATION BASE

To facilitate further description of the proposed solution we introduce new table for all known next hops and costs to it from various routers on the network.

Next-Hop Information Base (NHIB) stores cost to reach next-hop from arbitrary router on the network. This information is essential for choosing best path from a peer perspective rather than BGP-speaker own perspective. In canonical form NHIB entry is triplet (router, next-hop, cost), however this specification does not impose any restriction on how BGP implementations store that information internally. The cost in NHIB is does not have to be an IGP cost, but all costs in NHIB MUST be comparable with each other.

NHIB can be populated from various sources both static and dynamic. This document focuses on populating NHIB using BGP. However it is possible that protocols other than BGP could be also used to populate NHIB.

3. BGP BEST PATH SELECTION MODIFICATION

This section applies regardless of method used to populate NHIB.

When BGP speaker conforming to this specification selects routes to be advertised to a peer it SHOULD use cost information from NHIB

rather than its own IGP cost to the next-hop after step (d) of 9.1.2.2 in [RFC4271].

4. USING BGP TO POPULATE NHIB

This section describes extension to base BGP specification that allows BGP to be used for exchanging next-hop information between BGP speakers via new SAFI in order to populate NHIB. Although next-hops costs are exchanged via dedicated SAFI, this information is vital to best path selection process for other AFI/SAFI (e.g. IPv4 and IPv6 unicast). It's therefore recommended that next-hop cost information is exchanged before other AFI/SAFI.

4.1. NEXT-HOP SAFI

This document introduces Next-Hop SAFI (NH SAFI) with value to be assigned by IANA and purpose of exchanging information about cost to next-hops.

4.2. CAPABILITY ADVERTISEMENT

A BGP speaker willing to exchange next-hop information MUST advertise this in the OPEN message using BGP Capability Code 1 (Multiprotocol Extensions, see [RFC4760]) setting AFI appropriately to indicate IPv4 or IPv6 and SAFI to the value assigned by IANA for NH SAFI. Note that if BGP speaker whishes to exchange cost information for both IPv4 and IPv6, then it MUST advertise two capabilities: one NH SAFI for IPv4 and one NH SAFI for IPv6.

4.3. INFORMATION ENCODING

Routers use standard BGP UPDATE messages to exchange NH SAFI information. Cost to reachable next-hops is communicated using MP_REACH_NLRI (attribute 14) with NLRI part as described below. Requests are also sent using MP_REACH_NLRI. Informing a neighbour about unreachable next-hop is done using MP_UNREACH_NLRI. All NH SAFI messages MUST contain BGP COMMUNITY attribute with value NO_ADVERTISE (0xFFFFFF02) and their propagation MUST follow normal BGP rules (i.e. they're not to be propagated).

To request cost to a next-hop from peer or to inform peer about cost to a next-hop BGP attribute 14 is used as follow:

1. AFI is set to indicate IPv4 or IPv6 (whichever is appropriate)

2. SAFI is set to NH SAFI

3. Network Address of Next-Hop field is zeroed out

4. NLRI field is encoded as shown in the next figure

Format of NH SAFI NLRI is as follow:

```
 +-----+------+-------+---------+------+
 | AFI | SAFI | Flags | NEXT_HOP | cost |
 +-----+------+-------+---------+------+
```

Flags - 1 octet field. Least significant bit MUST be set to 1 for Request and to zero for Response

AFI/SAFI fields can be set either to one of the registered values to indicate that next-hop cost info applies only to specified AFI/SAFI. Alternatively when both fields are be set to zero, the cost information applies to any compatible AFI/SAFI negotiated with given peer.

Next-hop - IPv4 or IPv6 address for which cost is being communicated or requested. Type is determined from context, and length is inferred from total length of attribute.

Cost is 32-bit unsigned integer (value described below), and NEXT_HOP is AFI-specific address of the next-hop cost to which is being communicated or requested. Size of NEXT_HOP field is inferred from total length of attribute 14.

To inform peer that particular next-hop is unreachable MP_UNREACH_NLRI attribute is used with same NLRI format as described above. In this case cost field SHOULD be set to 0xFFFFFFFF.

4.4. SESSION ESTABLISHMENT

BGP speakers willing to exchange next-hop information SHOULD NOT establish more then one session for given AFI and NH SAFI, even using different transport addresses. This can be ensured for example by checking peer's Router Id.

4.5. INFORMATION EXCHANGE

Typically NH SAFI sessions will be established between route-reflectors and its internal peers (both clients and non-clients). As soon as the NH SAFI session is ESTABLISHED requests for next-hop cost and information information about next-hop costs MAY be sent independently. That is, route-reflector MAY send multiple requests without waiting for response, and its peers MAY send cost information before or after receiving such request. On the other hand, Router Reflectors SHOULD request cost information from their internal peers

as soon as possible (due to reasons stated in section "BGP best path selection modification"). BGP speaker does not need to track outstanding requests to the peer.

When a BGP speaker receives request for cost information it MUST reply with actual cost (not necessarily IGP cost, but whatever has been chosen to be carried in NH SAFI) to given next-hop or with cost set to all-ones indicating that next-hop is unreachable. If next-hop information is obtained from sender's routing table, then sender MUST perform lookup exactly the same way as it would for resolving next-hop in BGP UPDATE message. For example, for non-labelled destinations (e.g. AFI/SAFI 1/1 or 2/1) lookup would be done using longest match, whereas for labelled IPv4 (AFI/SAFI 1/4, 1/128 or 2/4) exact-match would be used.

When a BGP speaker detects change in cost to previously advertised next-hop with delta equal or exceeding configured advertisement threshold, it SHOULD inform peer by sending MP_UNREACH_NLRI as described earlier.

When a BGP speaker discovers new next-hop among candidate routes it SHOULD request cost information from the peer.

4.6. TERMINATION OF NH SAFI SESSION

When BGP speaker terminates (for whatever reason) NH SAFI session with a peer, it SHOULD remove all cost information received from that peer unless instructed by configuration to do otherwise.

4.7. GRACEFUL RESTART AND ROUTE REFRESH

NH SAFI sessions could use graceful restart and route refresh mechanisms in the same way as it's used for IPv4 and IPv6 unicast - preservation and purge of next-hop cost information follows normal GR rules.

5. Security considerations

No new security issues are introduced to the BGP protocol by this specification.

6. IANA Considerations

IANA is requested to allocate value for Next-Hop Subsequent Address Family Identifier.

7. Acknowledgment

Authors would like to thank Keyur Patel, Anton Elita, Nagendra Kumar for critical reviews and feedback.

8. References

8.1. Normative References

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

8.2. Informative References

[I-D.raszuk-bgp-optimal-route-reflection] Raszuk, R., Cassar, C., Aman, E., and B. Decraene, "BGP Optimal Route Reflection (BGP-ORR)", draft-raszuk-bgp-optimal-route-reflection-01 (work in progress), March 2011.

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.

Appendix A. USAGE SCENARIOS

A.1. Trivial case

```
--+---NetA---+--
  |          |
 r1          r2
  |          |
  R1--RR-----R2
  | \        |
  |  +------R4
  R3
```

In this scenario r1 and r3 along with NetA are part of AS1; and R1-R4 along with RR are in AS2.

If RR implements non-optimized route-reflection, then it will choose path to NetA via R1 and advertise it to both R3 and R4. Such choice is good from R3 perspective, but it results in suboptimal traffic

flow from R4 to NetA.

Using NH SAFI the route-reflector will learn that cost from R4 to R1 is 8 whereas to R2 it's only 1. RR will announce NetA to R4 with next-hop set to R2, while its announce to R3 will still have R1 as next-hop. Both R3 and R4 now will send traffic to NetA via closest exit, achieving same behaviour as if full iBGP mesh would have been configured.

A.2. Non-IGP based cost

When it's desirable to direct traffic over an exit other than the one with smallest IGP cost, NH SAFI can be used to convey cost which is not based on IGP. For example, network operator may arrange exit points in order of administrative preference and configure routers to send this instead of IGP cost. Route reflector then will then calculate best path based on administrative preference rather than IGP metrics.

Network operators should excercise care to ensure that all routers up to and including exit point do not devert packets on to a different path, otherwise routing loops may occur. One way to achieve this is to have consistent administrative preference among all routers. Another option is to use a tunneling mechanism (e.g. MPLS-TE tunnel) between source and the exit point, provided that the router serving as exit point will send packets out of the network rather than diverting them to another exit point.

A.3. Multiple route-reflectors

This example demonstrates that NH SAFI peerings are necessary only between routers that already exchange other AFI/SAFI.

```
                           |
R1----R3---------R5----R7--+
      |           |        |
     RR1          |       NetA
      |          RR2       |
      |           |        |
R2----R4---------R6----R8--+
                           |
```

In the above network the routers R1-R4 are clients of RR1, and R5-R8 are clients of RR2. RR1 and RR2 also peer with each other and use ADDPATH.

RR2 learns about NetA from R7 and R8. Since it sends not just best-path but all prefixes to RR1, there is no need for RR2 to learn cost

information from R1 and R2 towards R7 and R8. On the other hand RR1 does exchange NH SAFI information with R1 and R2 so that each of them can receive routes, which are best from their perspective.

As addition to ADDPATH a mechanism could be devised that would allow RR2 to learn how many alternative routes does it need to send to RR1. For example, if NetA would also be connected to R9 (not shown) but all clients of RR1 prefer R7 as exit point and R9 as next-best, then there is no need for RR2 to send NetA routes with next-hop R8 to RR1.

Discussion: authors would like to solicit discussion whether there is sufficient interest in such mechanism.

A.4. Inter-AS MPLS VPN

Previous example could be transposed to Inter-AS MPLS VPN Option C scenario. In this case route reflectors RR1 and RR2 can be from different autonomous system. Essentially the behaviour of routers remains as already described.

A.5. Corner case

```
   --+---NetA--+--
     |         |
RR---R1        R2
       \      /
        R3---R4
```

In the above network cost from R3 to R1 is 10, all other costs are 1. If RR advertises NetA to R3 based on cost information received from R3, but uses its own cost when advertising NetA to R4, there will be a loop formed. This is the reason why section "BGP best path selection modification" requires RR to have next-hop cost information for every next-hop and every peer.

Note that the problem is the same as if RR would not use extensions described in this document and R3 would peer directly with R1 and R2, while R4 would peer only with RR.

Authors' Addresses

Ilya Varlashkin
Easynet Global Services

Email: ilya.varlashkin@easynet.com

Robert Raszuk
NTT MCL Inc.
101 S Ellsworth Avenue Suite 350
San Mateo, CA 94401
US

Email: robert@raszuk.net