Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 4, 2014

R. Raszuk
NTT I3
W. Kumari
Google
J. Mitchell
Microsoft Corporation
K. Patel
Cisco Systems
J. Scudder
Juniper Networks

January 31, 2014

BGP Auto Discovery
draft-raszuk-idr-bgp-auto-discovery-00

Abstract

This document describes a method for automating portions of a router's BGP configuration via discovery of BGP peers with which to establish further sessions from an initial "bootstrap" router. This method can apply for establishment of either Internal or External BGP peering sessions.

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].

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 http://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 August 4, 2014.

Table of Contents

1. History

An idea for IBGP Auto Mesh [I-D.raszuk-idr-ibgp-auto-mesh] was originally presented at IETF 57. The concept made use of an IGP (either ISIS or OSPF) for flooding BGP auto discovery information. In this proposal both auto-discovery/bootstrapping and propagation of BGP configuration parameters occur within the BGP4 protocol itself.

The IGP based IBGP discovery mechanism presented was well fitted to the native IP switching, in which all nodes in the IGP need to participate in BGP mesh. However, it also came with a number of drawbacks, some of which include the requirement for leaking between area boundaries or possible race conditions between disjoint flooding paths from which the information arrived.

The BGP peer auto discovery mechanism described in this document was conceived initially in 2008 as a way to distribute peering session

establishment information via BGP for IBGP applications which are only active on the edge of the network. For example, these applications include BGP MPLS IP VPNs [RFC4364], rt-constrain [RFC4684], flow-spec [RFC5575], or Multicast VPNs [RFC6513]. However the idea was not documented for the community to discuss further at that time.

In 2011, another solution for BGP peer discovery that targeted EBGP peer discovery for Internet Exchange Point (IXP) participants was described in [I-D.wkumari-idr-socialite]. This idea was useful as a potential alternative solution for operators who wished to maintain individual peering sessions with other IXP participants, rather than receiving information through route-servers operated by the IXP operator without the associated administrative burden of configuring and maintaining sessions with all the other participants. This draft distributed the participant sessions information utilizing a BGP capability code [RFC5492] that was ill-suited for updating the information after initial session establishment.

This draft represents an attempt by the authors of both drafts to provide a solution that can be used in multiple IBGP or EBGP applications when the operator desires to automatically collect and distribute basic BGP session establishment information from a centralized BGP speaker.

2. Introduction

The base BGP-4 specification [RFC4271] utilizes TCP for session establishment between peers, which requires prior knowledge of the endpoint's address to which a BGP session should be targeted. This endpoint in most deployments is configured manually by the operator at each end of pair of network elements. In numerous applications, the list of all valid endpoints may be available centrally; however, the task of configuring or updating all of the network elements that require this information becomes a much larger task.

The most typical application of this in most networks is the establishment of a full mesh of IBGP routers to distribute standard IPv4 and IPv6 unicast routing information, such as the Internet route table, within an Autonomous System (AS). This was one of the reasons that lead to the introduction of BGP Route Reflection [RFC4456]. The most common benefits/drawbacks associated with route reflection are listed below:

- Configuration ease when adding or deleting new IBGP peers

- Reduction number of TCP sessions to be handled by ASBRs/PEs

- Information reduction - best path propagation only

- Limitation for new applications that require more than best path propagation

- Route instabilities caused by information reduction (ex: oscillations) etc. ...

Another application which requires prior knowledge of a large number of BGP endpoints is at Internet Exchange Points (IXP). These networks are specifically built and operated as locations for different networks to peer and exchange traffic. Multilateral Interconnection at an IXP [I-D.ietf-grow-ix-bgp-route-server-operations] is utilized to avoid having each participant at the IXP having to contact all of the other participants to enter into peering relationships, utilizing a Route Server (RS). Some of the reasons why participants peer with route-servers at IXPs include:

- reducing the administrative burden of arranging and configuring BGP sessions with all the other participants

- not wanting (or being able) to carry views from all the participants

- relying on the IXP operator to implement routing policy decisions (see [I-D.ietf-idr-ix-bgp-route-server])

This document describes an alternate solution for BGP peering session endpoint information discovery. This alternate solution reduces the administrative burden of configuring and maintaining BGP sessions in both IBGP applications (such as the full or partial mesh) and EBGP applications (such as at an IXP) as described above. This document does not address the other reasons why operators may choose to take alternative approaches that still require manual configuration or relying other devices for routing information distribution; however, auto-discovery and manual configuration are not mutually exclusive, and it is expected that some network elements will utilize both approaches.

In many cases existing route reflectors (in the IBGP use case) or route-servers (in the IXP) case may be utilized for the bootstrapping discovery mechanism in this document. This has several advantages:

- Re-use of already deployed devices for an add on and incremental automated BGP peer discovery

- Current place and operation in the network is optimal for session establishment for the relevant subset of clients that need the information.

- A verification only mode to analyze and generate a warning only message when manual IBGP peering configuration mistakes are detected.

3. Auto discovery mechanism

The amount of discovery information distributed via this mechanism is likely to be orders of magnitude less than the amount of underlying prefix (or other information) distributed today by existing route reflectors or route servers, so scalability for this mechanism should not be a concern.

This mechanism is designed to work on a per AFI/SAFI basis. For example, a currently deployed route reflector, providing route reflection for IPv4 unicast routes could continue in that function and at the same time provide a BGP peer discovery functionality for that or other address families. That could have a very positive effect for the deployment of any of the new address families as core RRs would not need to be upgraded to support new address families yet could still serve as information brokers for them.

In order to propagate information describing their BGP active configuration (activated AFI/SAFIs) we propose to define a new address family with the NLRI format of <Group_ID:Router_ID>.

The new address family will inherit current BGP update & msg formats as well as all necessary attributes used for normal and loop free BGP route distribution.

The Group Identifier Group_ID is a four octet value, and Router_ID is a four octet value [RFC6286].

The new type code for the new BGP Peer Discovery AFI/SAFI will be TBD1.

The role of the Group_ID is to allow scoped group creation in the same ASN/AFI/SAFI tuple. If not set by the operator, implying all peers will be in the same group, this value will be all zeros.

The way to group mesh interconnectivity is left to the operator. The Group_ID could be used for instance to group sub-AS or RR clients (if the RR is not doing client to client reflection), or for tying sets of EBGP peers to specific policy. A similar model takes place today for interconnecting confederation Sub-ASes as described in [RFC5065].

A new BGP Peer Discovery Attribute is defined to carry information about all activated and flagged for automatic provisioning AFI/SAFIs by a given BGP speaker. The format of the new BGP Peer Discovery Attribute is defined below in Figure 1:

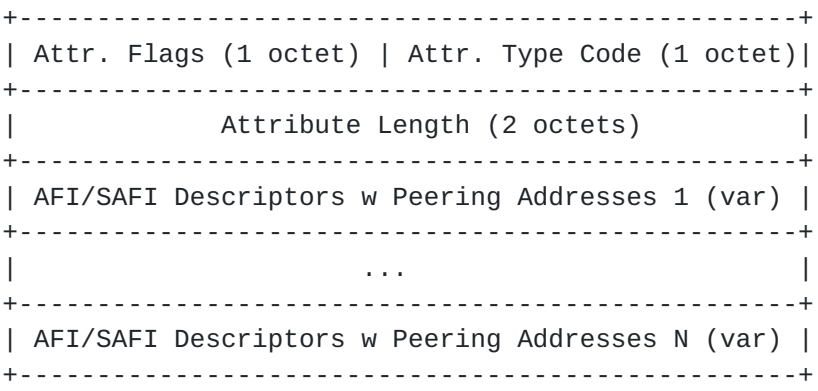

Figure 1: BGP Peer Discovery Attribute

The attribute flags and type code fields are detailed in Figure 2:

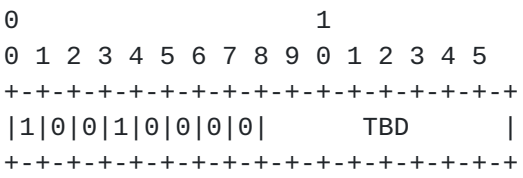

Figure 2: Flags & Type Code Fields

- Bit 0 - Optional attribute (value 1)
- Bit 1 - Non transitive attribute (value 0)
- Bit 2 - Partial bit (value 0 for optional non transitive attributes)
- Bit 3 - Extended length of two octets (value 1)
- Bit 4-7 - Unused (value all zeros)
- Type code - Attribute type code TBD2

Each BGP Peer Discovery Attribute contains one or more of the AFI/SAFI Descriptors as shown in Figure 3:

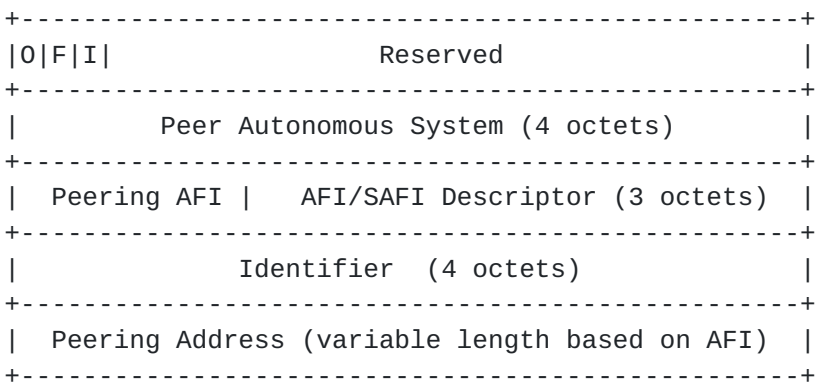

Figure 3: AFI/SAFI Descriptor

AFI/SAFI Descriptor Flags (1 octet):

- O bit - Route originator or EBGP speaker (Yes - 1, No - 0)

- F bit - Force new peering. Default not set - 0, set - 1.

- I bit - Informational only (Do not attempt to establish a BGP connection)

Peer Autonomous System Number:

- This is the neighbor's BGP Autonomous System Number (ASN), as described in [RFC6793], that should be expected for peering, iBGP if it matches the local router ASN, eBGP otherwise.

Identifier:

- This field is set to 0. If a non-zero value is set then the peer connection should be viewed as a tuple of <AFI/SAFI/Identifier>. Also at the same time the peer connection should be viewed as <AFI /SAFI/Identifier> and a separate connection should be initiated if the peer connection is not yet established.

Peering Address:

- Depending on the value of Peering AFI peering address on which BGP speaker is expecting to receive BGP session OPEN messages.

The special value of AFI/SAFI Descriptor can be all zeros. That will indicate that the information contained in the Group_id applies to all AFI/SAFIs given receiver supports. In those cases BGP OPEN msg will negotiate the subset of AFI/SAFIs to be established between given BGP peers.

It is expected that when Router_ID is changed on the BGP speaker sessions are restarted and therefore NLRI received with the former Router_ID withdrawn. When sessions restart, the new Router_ID will be sent in the NLRI corresponding to the BGP speaker with the reconfigured Router_ID. It is highly advised to change Router_ID only when critical as the impact to BGP is for any AFI/SAFI sever. An implementation may force the user to configure BGP Router_ID explicitly, before activating the new BGP Peer Discovery AFI/SAFI.

From the RR perspective as each BGP speaker can have only one Router_ID value, there would be only a single BGP Peer Discovery NLRI originated by one. It was a conscious design decision not to create a new BGP attribute for the reflector and require route reflector to build an aggregate list of AFI/SAFI descriptors common to given set of BGP Peer Discovery NLRIs in such a new attribute. We prefer to allow RR to remain simple with no additional code changes required for the price of no update packing possibility when it handles BGP Peer Discovery NLRIs in an atomic way.

Implementations MAY support local configuration of all possible remote peering address ranges, autonomous system numbers or other filters expected to be received via BGP Peer Discovery, or on a per group basis. Implementations SHOULD allow operators to group specific auto-discovered peers with specific groups based on Group_ID.

On the receive side, a persistent cache SHOULD be maintained by BGP with all received information about other BGP speakers announcing their BGP Peer Discovery information in a given Group's scope.

BGP Peer Discovery implementation should allow for per address family, subsequent address family and Group_ID disjoint topologies granularity.

When multiple AFI/SAFI pairs match on any two BGP speakers and value of the Identifier passed on AFI/SAFI Descriptor field is set to all zeros only one BGP session should be attempted. Regular BGP capabilities will be used to negotiate given AFI/SAFI mutual set. AFI/SAFI granularity is required to allow for disjoint topologies of different information being distributed by BGP.

BGP speakers "O" flag eligible may establish session with any other BGP speaker if passing all peering criteria for a given AFI/SAFI.

BGP speakers "O" flag not eligible (ex: P routers) should not establish IBGP peering to any other "O" flag not eligible BGP speakers.

When peering address changes for an existing AFI/SAFI and new BGP update is received with the new peering address old peering should remain intact when "F" flag is not set (default = 0). When session is cleared manually or goes down for any other reason, the new peering address should be used.

When "F" flag is set new peering address should be used immediately and current BGP session to the peer restarted for given AFI/SAFI.

4. Deployment Considerations

All implementations SHOULD still allow manual neighbor establishments which in fact could be complimentary and co-existing to the BGP Peer Auto Discovery neighbors.

In addition BGP Peer Auto Discovery exchange can be enabled just for informational purposes while provisioning would remain manual before operational teams get familiar with new capability and verify it's mechanics.

Within each Group_ID upon which auto-discovery is enabled, it is expected that neighbors will form sessions with all peers received within the group. This allows the building of full-mesh or partial-mesh topologies of peers for iBGP by varying the Group_ID field.

Incremental deployment with enabling just a few routers to advertise BGP Peer Discovery AF while maintaining manual configuration based peering with the rest of the network is supported.

Another key aspect of today's BGP deployment, other then peer to peer filtering push via ORF [RFC5292], is outbound customization of BGP information to be distributed among various peers. The most common tools for such customization could be peer templates, peer groups or any other similar local configuration grouping. Individual members of such groups can still be added to them manually, and BGP auto-discovery peers can be grouped to such groups using the Group_ID. The Peer Discovery implementation supports the ability to specify peer ranges which could automatically achieve addition or deletion of BGP peers to such groups. This can save a lot of manual configuration and customization for outbound policies shared by multiple peers. Individual session customization would be still possible by manual provisioning.

5. Capability Advertisement

A BGP speaker that wishes to exchange BGP Peer Discovery Information must use the the BGP Multiprotocol Extensions Capability Code as

defined in [RFC4760], to advertise the corresponding (AFI, SAFI) pair.

6. IANA Considerations

This document defines a new BGP Auto Discovery SAFI type code TBD1 which will be used to carry local BGP peering configuration data. That value will need to be assigned by IANA from BGP SAFI Type Code space.

This document defines a new NLRI format, called BGP Auto Discovery NLRI, to be carried in BGP Auto Discovery SAFI using BGP multiprotocol extensions. This document defines a new BGP optional transitive attribute, called BGP Peer Discovery Attribute. A new attribute type code TBD2 is to be assigned by IANA from the BGP path attribute Type Code space.

This document defines a new BGP Capability Type code (TBD3) to be allocated by IANA.

Once TBD1, TBD2, and TBD3 values are allocated please replace them in the above text.

7. Security Considerations

This document allows for local configuration of BGP authentication mechanisms such as BGP-MD5 [RFC2385] or TCP-AO [RFC5925] and these are highly recommended for deployment on the BGP peer auto-discovery neighbor sessions. Similar authentication could be configured on a per peer or peer-group basis based on the auto-discovery information received before session establishment, however no exchange of authentication information occurs within the protocol itself. Operators SHOULD NOT use peer auto-discovery with untrusted peers as attacks on implementation scalability could be triggered by overwhelming the router with a larger number of auto-discovery peers then can be supported. Operators should also use caution on what addresses and AFI/SAFI combinations they want to allow reception of auto-discovery information for.

8. Contributors

The BGP auto-discovery idea contained in this document was originally developed by Pedro Roque Margues and Robert Raszuk in 2008 to cover the IBGP full mesh use case however it was not publicly documented at that time.

9. Acknowledgments

TBD

10. References

10.1. Normative References

[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.

[RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, December 2012.

10.2. Informative References

[I-D.ietf-grow-ix-bgp-route-server-operations] Hilliard, N., Jasinska, E., Raszuk, R., and N. Bakker, "Internet Exchange Route Server Operations", draft-ietf-grow-ix-bgp-route-server-operations-01 (work in progress), August 2013.

[I-D.ietf-idr-ix-bgp-route-server] Jasinska, E., Hilliard, N., Raszuk, R., and N. Bakker, "Internet Exchange Route Server", draft-ietf-idr-ix-bgp-route-server-03 (work in progress), August 2013.

[I-D.raszuk-idr-ibgp-auto-mesh] Raszuk, R., "IBGP Auto Mesh", draft-raszuk-idr-ibgp-auto-mesh-00 (work in progress), June 2003.

[I-D.wkumari-idr-socialite] Kumari, W., Patel, K., and J. Scudder, "Automagic peering at IXPs.", draft-wkumari-idr-socialite-02 (work in progress), October 2012.

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.

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

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007.

[RFC5292] Chen, E. and S. Sangli, "Address-Prefix-Based Outbound Route Filter for BGP-4", RFC 5292, August 2008.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

[RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, August 2009.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

[RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, June 2011.

[RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.

Authors' Addresses

Robert Raszuk
NTT I3
101 S Ellsworth Ave
San Mateo, CA 94401
US

Email: robert@raszuk.net

Warren Kumari
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: warren@kumari.net

Jon Mitchell
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
USA

Email: Jon.Mitchell@microsoft.com

Keyur Patel
Cisco Systems
170 West Tasman Dr.
San Jose, CA 95135
US

Email: keyupate@cisco.com

John Scudder
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale CA
USA

Email: jgs@juniper.net