Network Working Group
Internet Draft
Category: Standards Track
Expires: December 2003

Robert Raszuk (Editor)
Chandra Appanna
Cisco Systems, Inc
Pedro Roque Marques
Juniper Networks, Inc

June 2003

IBGP Auto Mesh

draft-raszuk-idr-ibgp-auto-mesh-00.txt

Status of this Memo

Abstract

The distribution of BGP routing information within an autonomous system requires all border routers to be fully meshed. This constitutes a significant operational problem in terms of configuration management.

This has led to the wide-spread adoption of route reflection [RFC2796], primarily in order to reduce the number systems which configuration must be modified in order to introduce or remove a new internal BGP speaker. Route reflection, however, implies with it information reduction which is not always desired.

This document defines a discovery mechanism that is designed to address the problem of introducing (or removing) a BGP speaker into an iBGP mesh without implying any other behavior change when compared

to manual configuration.

1. Specification of Requirements

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 [RFC2119].

2. Introduction

One of the most common complains received from operators is the comment on complexities associated with configuration of BGP meshes.

This draft attempts to make this claim a history by proposing auto discovery of internal BGP peers via configuration information flooding as well as a set of procedures which would allow to establish IBGP sessions automatically.

It should be noted that for easy and fast deployment flooding of information is piggybacked on top of existing IGP mechanisms. This feature is designed with flooding mechanism transparency in mind. When new and more effective flooding protocol is introduced and deployed into production networks it should be with no additional effort on the BGP component to migrate the flooding from IGPs to such a new protocol.

Unlike other attempts in this area it does not relay on any management station. Also it keeps all BGP functional and transport mechanism unchanged.

The particular piece of functionality this draft addresses is distribution of related local configuration to all routers within flooding scope as well as usage of this information in establishment of IBGP peering.

Today we can observe a number of network topologies where IBGP is being used to distribute routing information between BGP speakers:

A - Full iBGP mesh between all routers in the AS
B - Full iBGP mesh between all members of confederation
C - Route reflectors clusters peering
D - Full iBGP mesh between all PEs/ASBRs and BGP free core

In the operation section we will describe how this draft can automate IBGP peering in all of the above scenarios.

Another very important observation is that today BGP often play other roles than just distribution of IPv4 reachability information. With introduction of multiprotocol BGP extensions [RFC2858] BGP speakers may be configured to keep and maintain data belonging to multiple address families.

In multiprotocol environments, the IBGP mesh for one address family may not match the mesh for a different address family as in the case where different route reflectors are used for different applications. This draft will also automate IBGP session establishment with matching AFIs taken into consideration of the auto discovered peers.

Historically introduction of route reflection drastically removed the need for full mesh manual configuration of all BGP speakers in a domain. It also reduced the number of TCP session each BGP speakers needed to handle.

Another characteristics of route reflection is their ability to eliminate number of advertised paths to a given peer to only best path from reflector's point of view.

While this last point was originally thought of a benefit (at least from the perspective of best effort ipv4 reachability) today's applications require a bit more granular path's classification and features like IBGP multipath is becoming already a production requirement. To keep the former benefit of reduced configuration or low number of TCP sessions approaches like add-paths draft [ADD-PATH] are being currently under investigation.

It should be noticed that idea presented in this draft eliminates any manual configuration for full mesh BGP peering. Also modern operating systems implementations can hold and maintain much more TCP sessions that their predecessors could. Therefore by auto meshing all or selected address families on BGP speakers the need for route reflection becomes obsolete in some cases.

In order to reduce amount of information distributed across IBGP sessions to only a required subset outbound route filtering techniques could be employed [IDR-ORF].

Another operational benefit of using IBGP auto mesh is the ease of AS renumbering or merges/migrations. It is generally very difficult to manually change the AS number on all BGP speakers in a short maintenance window. With the automation such task would be much easier to accomplish.

3. The BGP Auto Discovery TLV

This section proposes an encoding to be used in IPv4 & IPv6 networks.

The BGP Auto Discovery TLV is defined as follows:

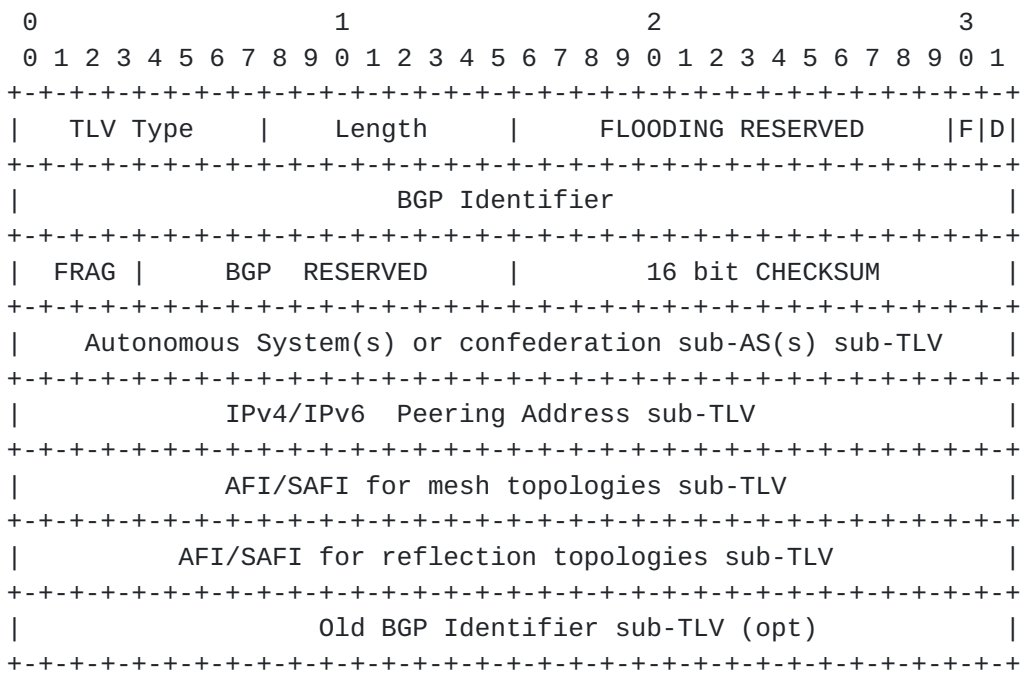

Figure 1. BGP Auto Discovery TLV

3.1. TLV Type

BGP Auto Discovery TLV is proposed to be carried in OSPF Router Information LSA [LINDEM1] with area scope or domain wide scope depending on the configuration. Details describing OSPF specifics and encoding will be described in [OSPF-BGP].

In ISIS BGP Auto Discovery TLV is proposed to be carried as a new TLV with flooding scope local to the intra area or domain wide. Details describing ISIS specific encoding will be described in [ISIS-BGP].

The selection of ISIS and OSPF for flooding is mostly based on the fact that those protocols already have a flooding mechanism which can be reused for the purpose of required in this proposal information distribution.

It is a strong design goal that flooding of BGP Auto Discovery TLV can be realized over any other protocol when such is deployed and when it can provide further benefits. For example: selective groups of destinations or disjoined information distribution trees per

AFI/SAFI.

In cases of multiple bgp processes running on a router each BGP process should send it's own BGP Auto Discovery TLV with a different BGP Identifier.

3.2. TLV Length

Length - Total length in octets of this TLV

The minimum TLV length can be 10 octets.

When a size of the TLV reaches 255 octets TLV fragmentation needs to occur. A special "FRAG" 4 bit counter has been allocated in the BGP Control Information's first octet for unlikely cases where BGP Auto Discovery TLV needs to be splitted across multiple TLVs for a given BGP speaker.

It is estimated that the average size of BGP Auto Discovery TLV in today's production environments will be anywhere from 30-50 octets.

3.3. Global flooding flags

This comprises of one or more global for given BGP Auto Discovery TLV flags related to flooding. Currently defined are the following flags:

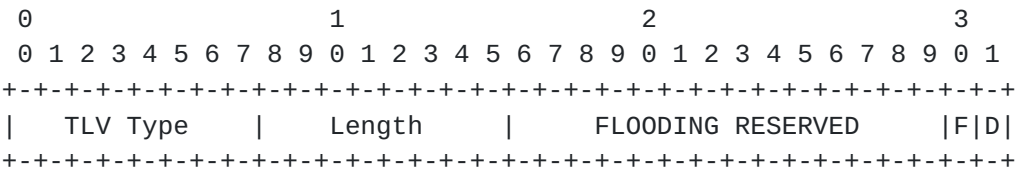

Figure 2. Value field of global flags

Symbol	Definition
F	Flooding scope
D	Down Bit

"F" {flooding} flooding scope of this TLV. When set domain wide flooding scope is required, when not set TLV should not be flooded into other areas or levels. Default not set indicating area/level wide flooding only.

"D" {down} down bit set by ISIS when leaked to other areas/levels.

When advertised BGP Auto Discovery TLV for a given BGP Identifier does not contain any sub-TLVs it should be interpreted as an implicit

withdraw of any previously advertised BGP Auto Discovery TLV for a given BGP Identifier.

This is intentional not to duplicate BGP capabilities in this message. BGP operation in session establishment to auto discovered peers should remain without any changes as compared to today's operation.

All reserved flags and not defined yet should be set to zeros.

3.4. BGP Identifier

4 octet value set to BGP ID [IDR-BGP4]. When BGP ID is being changed on a BGP speaker next BGP Auto Discovery TLV announced should contain "Old BGP ID sub-TLV". Presence of this sub-TLV is an indication that the received BGP Auto Discovery TLV is being related to previously flooded BGP information with the contained in this sub-TLV BGP Identifier.

3.5. BGP Control Information

The below 4 octets with BGP specific control information has been defined:

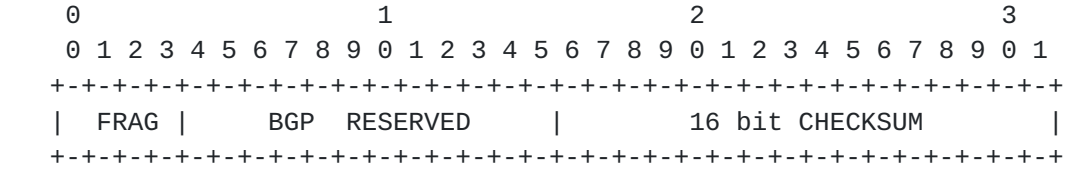

Figure 3 BGP Control Information

3.6. Fragmentation counter

When a size of the BGP Auto Discovery TLV reaches 255 octets TLV fragmentation needs to occur. A special "FRAG" 4 bit counter has been allocated for unlikely cases where BGP Auto Discovery TLV needs to be splitted across multiple TLVs for a given BGP speaker.

3.7. Checksum

Depending on the flooding protocol being utilized BGP may be receiving identical information periodically. To allow easy determination if the content of the TLV has changed for a given BGP Identifier 16 bit checksum has been defined. Checksum should be computed from the content of the entire BGP Auto Discovery TLV (before potential fragmentation) excluding first 4 octets of the TLV.

4. The BGP Auto Discovery sub-TLVs

In following subsections we will focus on describing each of the sub-TLVs directly related to BGP operation. The format of each sub-TLV will be following:

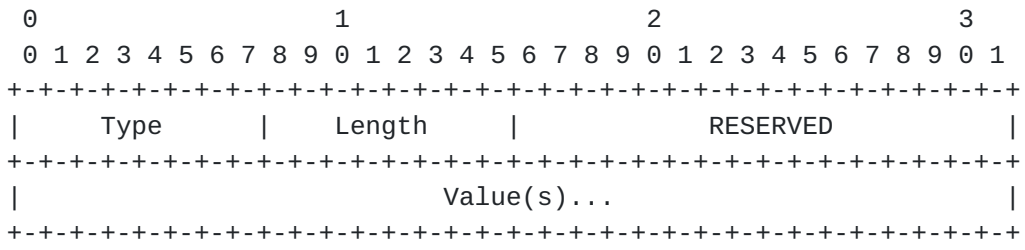

Figure 4. BGP Sub-TLV

4.1. BGP Autonomous System(s) sub-TLV

Type	One octet field set to value of 1.
Length	One octet field that indicates the length of the value portion in octets.
Reserved	Two octet field reserved for flags and sub-TLV control
Value	4 octet BGP Autonomous System Number(s) [IDR-BGP4] [IDR-AS4] or confederation sub-AS [RFC1965].

Advertising multiple autonomous system numbers may be required during AS renumbering and merges with other ASes. Therefore this proposal does not limit advertisement to a single AS value per BGP speaker.

The peering attempt should be made only to those peers which match locally configured AS number or numbers (multi-as migration case).

When confederation is used sub-AS will limit the scope of full mesh peering only to a given sub-AS even if flooding scope is common to all sub-ASes. Usage of route reflectors within each confederation sub-AS is also supported.

4.2. IPv4 Peering Address sub-TLV

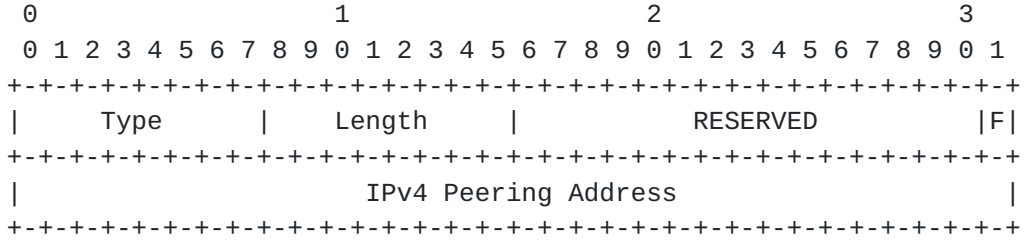

Figure 5. BGP IPv4 Peering Sub-TLV

Type	One octet field set to value of 2.
Length	One octet field that indicates the length of the value portion in octets.
Reserved	Set to all zeros
Flag	"F" - Force new peering. Default not set.
Value	4 octet ipv4 peering address

This address will be used by BGP speakers as a destination in BGP Open message. Sending a BGP Auto Discovery TLV with new peering address is an explicit withdraw of the previously advertised one.

When such a messages is received old peering should remain intact when "F" flag is not set (default). When session is cleared manually or IGP reachability to the old peering address disappears 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.

4.3. IPv6 Peering Address sub-TLV

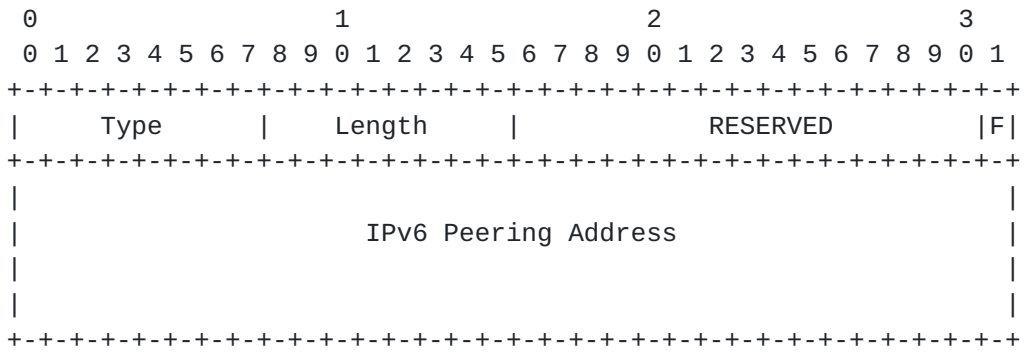

Figure 6. BGP IPv6 Peering Sub-TLV

Type	One octet field set to value of 3.
Length	One octet field that indicates the length of the value portion in octets.
Reserved	Set to all zeros
Flag	"F" - Force new peering. Default not set.
Value	16 octet ipv6 peering address

This address will be used by BGP speakers as the destination in BGP Open message. Sending a BGP Auto Discovery TLV with new peering address is an explicit withdraw of the previously advertised one.

When such a messages is received old peering should remain intact when "F" flag is not set (default). When session is cleared manually or IGP reachability to the old peering address disappears 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.

When both IPv4 & IPv6 peering addresses are present it is up to the implementation to decide on the peering address selection.

4.4. AFI/SAFI for mesh topologies sub-TLV

Type	One octet field set to value of 4.
Length	One octet field that indicates the length of the value portion in octets.
Reserved	Set to zero
Value	2 octet Address Family Identifier(s) [RFC2858]
	1 octet Subsequent Address Family Identifier [RFC2858]
	1 octet Per AFI/SAFI Flags

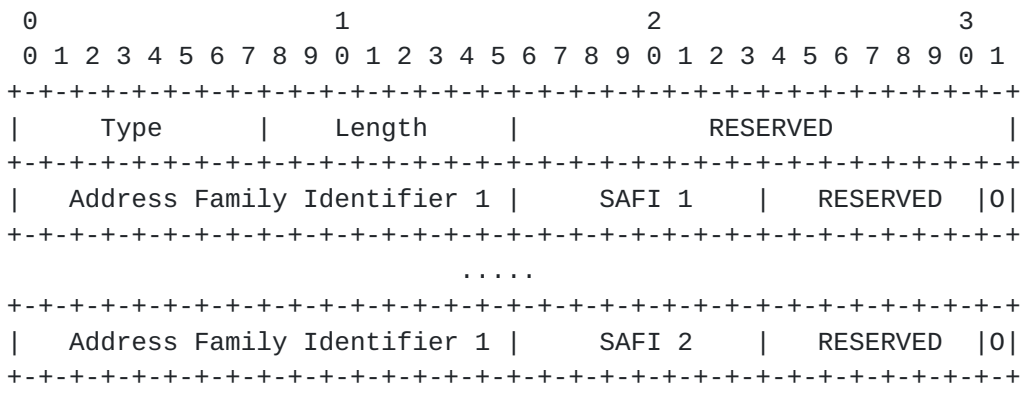

Figure 7. Value field of BGP AFI/SAFI mesh topologies sub-TLV

Sub-TLV Value Flags:

"O" - Originator or EBGP speaker

Originator flag:

IBGP sessions in a full or partial mesh topology are required to directly peer with those BGP speakers which originate routes or which maintain EBGP sessions. This flag should be used to mark such a bgp speakers when advertising BGP Auto Discovery TLV. On reception this flag should be used for selection or required IBGP peering candidates.

It is important to note that the actual state of EBGP session or present or not in the routing table of redistribuited prefix is not relevant and this bit should be set always when EBGP session or local

route origination is configured.

4.5. AFI/SAFI for reflection topologies sub-TLV

Type One octet field set to value of 5.
Length One octet field that indicates the length of the value portion in octets.
Reserved Set to zero
Value 2 octet Address Family Identifier(s) [RFC2858]
1 octet Subsequent Address Family Identifier [RFC2858]
1 octet Per AFI/SAFI Flags

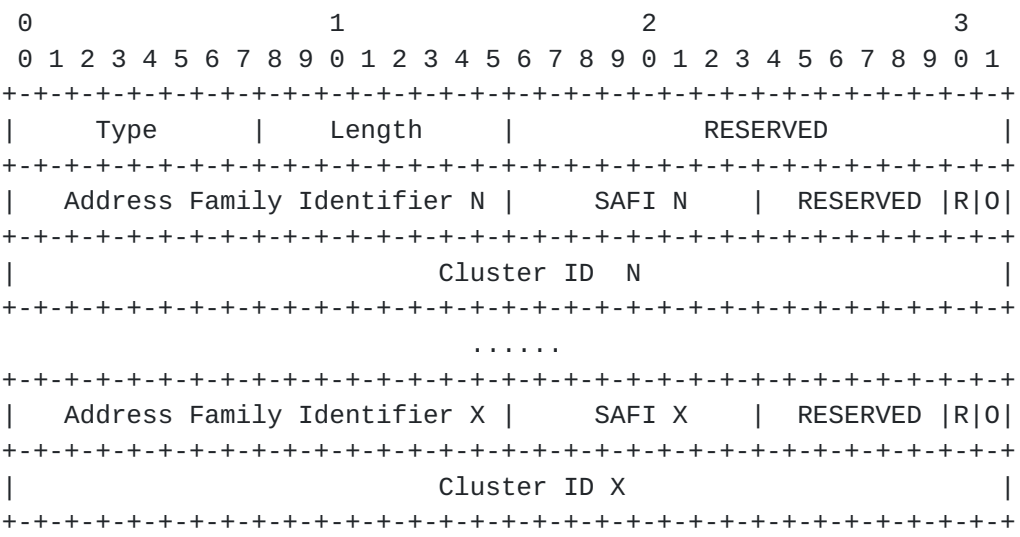

Figure 8. Value field of AFI/SAFI reflection sub-TLV

Sub-TLV Value Flags:

"O" - Originator or EBGP speaker
"R" - Route Reflector for given AFI-SAFI/Cluster_ID combination

See section "AFI/SAFI for mesh topologies sub-TLV" for explanation of "O" flag.

"R" - Route reflector flag.

When "R" flag is set BGP speaker announcing this TLV is configured for route reflection function for a given AFI/SAFI combination. In addition when "R" bit is set the following cluster ID 4 octet value [RFC2796] indicates cluster id assigned for a given reflection function.

Clients of route reflection will send their cluster ID lists assigned to each AFI/SAFI without "R" bit set. When client wishes to indicate

the request to become a member of all possible cluster_ids for given AFI/SAFI combination within a flooding scope of his BGP Auto Discovery TLV the "R" bit should not be set and the value of cluster_id associated with the AFI/SAFI should be set to all zeros.

To allow chain of route reflection (hierarchy) it is perfectly valid for a BGP speaker to have for a given AFI/SAFI a "R" bit set for one cluster IDs (ie to perform a route reflection function) and in the same time for other cluster ID values be a client of other route reflectors (R bit not set).

Network designs of reflection within confederations are also supported.

At this time of publication authors will also leave the implementor's freedom to allow single sided signaling only from route reflectors to the clients. When client receives the BGP Auto Discovery TLV which contains the interesting cluster ID and has R bit set it can initiate BGP Open without injecting any information about his own BGP configuration in the reflection topologies into the network.

4.6. Old BGP Identifier sub-TLV

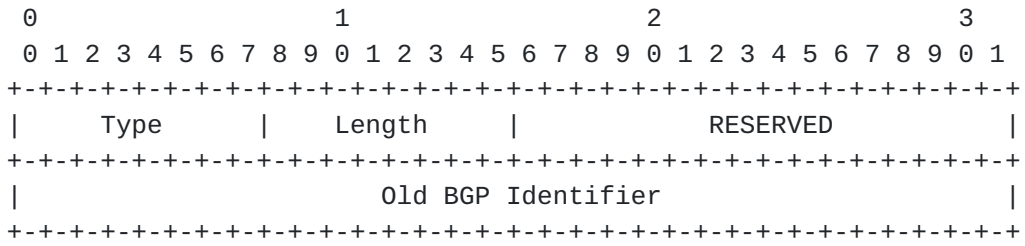

Figure 9. Old BGP Identifier Sub-TLV

Type	One octet field set to value of 6.
Length	One octet field that indicates the length of the value portion in octets.
Reserved	Set to all zeros
Value	4 octet BGP Identifier [IDR-BGP4].

When BGP Identifier is being replaced with a new value the "Old BGP Identifier" sub-TLV must be present and contain a previously advertised BGP ID for this given BGP speaker.

5. Operation

Each BGP speaker configured to participate in an IBGP auto mesh should pass to flooding component BGP Auto Discovery TLV with it's own local configuration dependent information.

On the receive side, a cache should be maintained by BGP with all received information from flooding component about other BGP speakers announcing their BGP Auto Discovery TLVs in a given area or domain.

IBGP auto mesh configuration should allow for per address family and subsequent address family disjoint topologies granularity.

When multiple AFI/SAFI pairs match on any two BGP speakers only one IBGP session should be attempted. Regular BGP capabilities will be used to negotiate given AFI/SAFI mutual set. Never less AFI/SAFI granularity is required to allow for very fine grade disjoint topologies for different types of distributed by BGP information.

5.1. Full mesh topologies

When operator finds required to fully or partially mesh BGP speakers "AFI/SAFI for mesh topologies" sub-TLV should be utilized.

BGP speakers may be eligible for origination of routes or may be configured for EBGP peering. We will call them "O" flag eligible (see section "AFI/SAFI for mesh topologies sub-TLV").

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.

One possible example of such a configuration could be vpnv4 AF connecting all PEs in a domain into a full IBGP mesh.

After reception of peers BGP Auto Discovery TLV BGP speaker should check for autonomous system numbers match as well as AFI/SAFI identifiers match. Positive results from the above actions should trigger a standard process of connection establishment attempt with the peer.

It is also highly recommened that a local range of allowed peering addressed be also maintained and consulted at each attempt to establish a new IBGP peering.

BGP Auto Discovery TLV may be area/level or domain wide in full mesh topologies. The default should be area/level wide flooding.

5.2. Confederations

To automate iBGP full mesh establishment in each confederation sub-AS each confederation member should advertise it's confederation sub-AS instead of main AS (confederation_id) it is a member of in BGP Autonomous System(s) sub-TLV.

There could be two cases here:

A) Confederation sub-ASes strictly contained within flooding scope
B) Confederation sub-ASes unrelated to flooding topology

Case (A) BGP auto discovery TLV flooding scope should be limited to one area/level.

Case (B) BGP auto discovery TLV flooding scope should be domain wide and use of Auto Peering Range(s) sub-TLV is highly recommended.

In the cases where reflectors are used within each confederation rather then direct peering "AFI/SAFI for reflection topologies sub-TLV" should be used instead of "AFI/SAFI for mesh topologies sub-TLV".

5.3. Route Reflectors

When operator wishes to automate establishment of BGP sessions to route reflectors the only additional information required is configuration of at least one identical cluster id on both route reflector as well as on route reflector client. As mentioned earlier even this requirement could be relaxed by implementation supporting single sided signaling of reflector capabilities. The drawback in such a case is that route reflector injecting his BGP Auto Discovery TLV would also need to be configured with an additional information allowing to distinguish BGP Open requests coming from clients as well as those coming from non clients based on the peering address range and mark such a peering accordingly.

Routers or devices designated to serve route reflector function shall advertise their "AFI/SAFI for reflection topologies" sub-TLV with "R" flag set as well as with their cluster id(s) attached.

If IBGP session will be established between route reflector ("R" flag set) and non route reflector BGP speaker ("R" flag not set) who's specific AFI/SAFI cluster ID matches on at least one entry with given route reflector cluster id it should be marked as route reflector

client.

BGP speakers which are not to act as route reflectors ("R" flag not set) and do not have configured cluster id value(s) indicating their designation as route reflector clients would attempt to establish regular IBGP peering to other BGP speakers in the domain (per rules described in section "Full mesh topologies").

An implementation may also allow the additional route reflection client to client full mesh. This is left for the implementor's choice to be enabled with a configuration option.

Route reflection chaining (reflector hierarchy) is fully supported. Route reflector server may advertise for a given AFI/SAFI his ability to reflect routes for one set of cluster ID(s) ("R" bit set) and in the same time for the same AFI/SAFI value submit a list of cluster IDs without "R" bit set indicating the willingness to become a regular client on servers eligible to reflect those cluster ID(s).

When client wishes to indicate the request to become a member of all possible cluster_ids for given AFI/SAFI combination within a flooding scope of his BGP Auto Discovery TLV the "R" bit should not be set and the value of cluster_id associated with the AFI/SAFI should be set to all zeros.

6. Local peering verification

It is highly recommended for an implementation to support local configuration of all possible remote peering address ranges expected to be received via BGP Auto Discovery TLV messages.

In particular this can protect from configuration mistakes when peering in a full or partial mesh and setting flooding scope accidentally to domain wide.

In this version of the draft the decision has been made not to flood this local peering range list to the remote peers. Such a flooding could further protect from even sending BGP Open message when given bgp speaker own peering address does not match received list from a peer.

That decision can be revisited in the future versions of this work and new sub-TLV for flooding this information can be added.

7. Deployment Considerations

The idea described in this document should not present any deployment challenges. It is expected that all implementations would still allow manual neighbor establishments which in fact could be complimentary and co-existing to the IBGP auto mesh.

In addition BGP Auto Discovery TLV exchange could be enabled just for informational purposes while provisioning would remain manual before operational teams get familiar with new functionality.

Incremental deployment with enabling just a few routers to advertise BGP Auto Discovery TLV while maintaining manual configuration based peering with the rest of the network is supported. An implementation may also allow for mixed peering for example reflector client being configured automatically while reflector's clusters itself being interconnected manually.

8. IANA Considerations

There are no IANA considerations required in this document. Extensions to ISIS [ISIS-BGP] & OSPF [OSPF-BGP] will have their own IANA consideration sections.

9. Security Considerations

This document fully relies on authentication mechanism that an implementation of BGP MUST support as specified in [RFC2385]. The authentication provided by this mechanism could be done on a per peer basis.

It also relies on security of flooding mechanism being used for information distribution.

10. Acknowledgments

I would like to express our thanks to Alvaro Retana, Keyur Patel, Barry Friedman & Gargi Nalawade for their valuable comments.

11. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119, March 1997.

[RFC2434] Narten, T., Alvestrand, H., "Guidelines for Writing an IANA Considerations Section in RFCs", RFC2434, October 1998.

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

12. Informative References

[RFC1771] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[IDR-BGP4] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4)", Work in Progress (draft-ietf-idr-bgp4-21.txt), April 2003.

[RFC1965] Traina, P., "Autonomous System Confederations for BGP", RFC 1965, June 1996.

[RFC2796] Bates, T., Chandra, R., and Chen, E., "BGP Route Reflection An Alternative to Full Mesh IBGP", RFC 2796, April 2000.

[RFC2858] Bates, T., Rekhter, Y., Chandra, R., Katz, D., "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000

[IDR-ORF] Chen, E., Rekhter, Y., "Cooperative Route Filtering Capability for BGP-4", draft-ietf-idr-route-filter-08.txt

[IDR-AS4] Vohra, Q., Chen, E., "BGP support for four-octet AS number space" draft-ietf-idr-as4octets-06.txt, January 2003

[LINDEM1] Lindem, A. at all, "Extensions to OSPF for Advertising Optional Router Capabilities", draft-lindem-ospf-cap-00.txt, May 2003

[ADD-PATH] Walton, D.,at all, "Advertisement of Multiple Paths in

BGP" draft-walton-bgp-add-paths-01.txt

[ISIS-BGP] Raszuk, R., "ISIS Extensions for BGP Peer Discovery" , draft-raszuk-isis-bgp-peer-discovery-00.txt, June 2003 Work in progress

[OSPF-BGP] Raszuk, R., "OSPF Extensions for BGP peer discovery", draft-raszuk-ospf-bgp-peer-discovery.txt, Work in Progress

13. Authors' Addresses

Robert Raszuk
Cisco Systems, Inc.
Al. Jerozolimskie 146C
02-305 Warsaw, Poland
E-mail: rraszuk@cisco.com

Pedro Marques
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
E-mail: roque@juniper.net

Chandrashekhar Appanna
Cisco Systems, Inc
170 West Tasman Dr
San Jose, CA 95134
E-mail: achandra@cisco.com

14. IPR Notices

be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

15. Terms of Use

Cisco has a pending patent which relates to the subject matter of this Internet Draft. If a standard relating to this subject matter is adopted by IETF and any claims of any issued Cisco patents are necessary for practicing this standard, any party will be able to obtain a license from Cisco to use any such patent claims under openly specified, reasonable, non-discriminatory terms to implement and fully comply with the standard.

16. Full Copyright Notice