| | |
|---|---|
| Network Working Group | P. Marques |
| Internet-Draft | R. Raszuk |
| Intended status: Standards Track | K. Patel |
| Expires: December 26, 2011 | Cisco Systems |
| | K. Kumaki |
| | T. Yamagata |
| | KDDI Corporation |
| | June 24, 2011 |

Internal BGP as Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)
draft-ietf-l3vpn-ibgp-08

## Abstract

This document defines protocol extensions and procedures for BGP Provider/Customer edge router iteration in BGP/MPLS IP VPN networks. These have the objective of making the usage of the BGP/MPLS IP VPN transparent to the customer network, as far as routing information is concerned.

## Status of this Memo



## Copyright Notice

## Table of Contents


*1. Introduction

*2. Requirements Language

*3. IP VPN network as a Route Server

*4. Path attributes

*5. BGP customer route attributes

*6. Next-hop handling

*7. Exchanging routes between different VPN customer networks

*8. Deployment considerations

*9. Security considerations

*10. IANA considerations

*11. Acknowledgments

*12. References

*12.1. Normative References

*12.2. Informative References

*Authors' Addresses


## 1. Introduction

In current deployments, when BGP is used as the Provider/Customer Edge routing protocol, these peering sessions are typically configured as an external peering between the VPN provider autonomous-system (AS) and the customer network autonomous-system. At each External BGP boundary, BGP Path Attributes *[RFC4271]* are modified as per standard BGP rules. This includes prepending the AS_PATH attribute with the autonomous-system number of the originating customer edge (CE) router and the autonomous-system number(s) of the provider edge (PE) router(s).
In order for such routes not to be rejected by AS_PATH loop detection, a PE router advertising a route received from a remote PE, often remaps the customer network autonomous-system number to its own. Otherwise the customer network can use different autonomous-system numbers at different sites or configure their CE routers to accept routes containing their own AS number.
While this technique works well in situations where there are no BGP routing exchanges between the client network and other networks, it

does have drawbacks for customer networks that use BGP internally for purposes other than interaction between CE and PE routers.
In order to make the usage of BGP/MPLS VPN services as transparent as possible to any external interaction, it is desirable to define a mechanism by which PE-CE routers can exchange BGP routes by means other than external BGP.
One can consider a BGP/MPLS VPN as a provider-managed backbone service interconnecting several customer-managed sites. While this model is not universal it does constitute a good starting point.
Independently of the presence of VPN service, networks often use an hierarchical design utilizing either BGP route reflection *[RFC4456]* or confederations *[RFC5065]*. This document assumes that the IP VPN service interacts with the customer network following a similar model.

## 2. Requirements Language

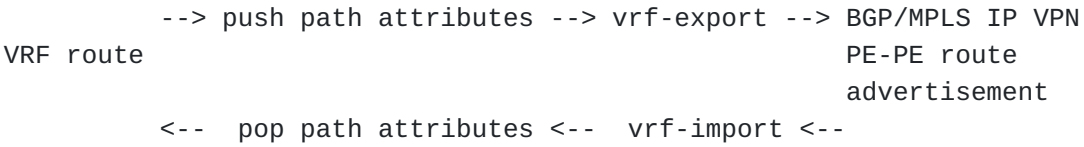
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 *[RFC2119]*.

## 3. IP VPN network as a Route Server

In a typical backbone/area hierarchical design, routers that attach an area (or site) to the core, use BGP route reflection (or confederations) to distribute routes between the top-level core iBGP mesh and the local area iBGP cluster.
To provide equivalent functionality in a network using a provider provisioned backbone, one can consider the VPN as the equivalent of an Internal BGP Route Server which multiplexes information from _N_ VPN attachment points.
A route learned by any of the PEs in the IP VPN network, is available to all other PEs that import the Route Target used to identify the customer network. This is conceptually equivalent to a centralized route server.
In a PE router, PE received routes are not advertised back to other PEs. It is this split horizon technique that prevents routing loops in an IP VPN environment. This is also consistent with the behavior of a top level mesh of RRs.
In order to complete the Route Server model, is necessary to be able to transparently carry the Internal BGP PATH attributes of customer network routes through the BGP/MPLS VPN core. This is achieved by using a new BGP path attribute described below that allows the customer network attributes to be saved and restored at the BGP/MPLS VPN boundaries.
When a route is advertised from PE to CE, if it is advertised as an iBGP route, the CE will not advertise it further unless it is itself configured as a Route Reflector (or has an external BGP session). This is a consequence of the default BGP behavior of not advertising iBGP routes back to iBGP peers. This behavior is not modified.

On a BGP/MPLS VPN PE, a CE-received route MUST be advertised to other VPN PEs that import the Route Targets which are associated with the route. This is independent of whether the CE route has been received as an external or internal route. However, a CE received route is not readvertised back to other CEs unless Route Reflection (RR) is explicitly configured. This is the equivalent of disabling client to client reflection in BGP RR implementations.
When reflection is configured on the PE router, with local CE routers as clients, there is no need to internally mesh multiple CEs that may exist in the site.
This Route Server model can also be used to support a confederation style abstraction to CE devices. We choose not to describe in detail the procedures for that mode of operation, at this point.
Confederations are considered to be less common than route reflection in enterprise environments.

## 4. Path attributes




The diagram above shows the BGP path attribute stack processing in relation to existing BGP/MPLS IP VPN *[RFC4364]* route processing procedures. BGP path attributes received from a customer network are pushed into the stack, before adding the Export Route Targets to the BGP path attributes. Conversely, the stack is popped after the Import Target processing step that identifies the VPN Routing and Forwarding (VRF) table in which a PE received route is accepted.
When the advertising PE performs a "push" operation at the "vrf-export" processing stage it SHOULD initialize the attributes of the BGP IP VPN route advertisement as if for a locally originated route from the respective VRF context.
When a PE received route is imported into a VRF, its IGP metric, as far as BGP path selection is concerned, SHOULD be the metric to the remote PE address, expressed in terms of the service provider metric domain.
For the purposes of VRF route selection performed at the PE, between routes received from local CEs and remote PEs, customer network IGP metrics SHOULD always be considered higher (thus least preferred) than local site metrics.
When backdoor links are present, this would tend to direct the traffic between two sites through the backdoor link for BGP routes originated by a remote site. However BGP already has policy mechanisms to address this type of situations such as the LOCAL_PREF attribute.
When a given CE is connected to more than one PE, it will not advertise the route that it receives from a PE to another PE unless configured as a route reflector, due to the standard BGP route advertisement rules.

When a CE reflects a PE received route to another PE, the fact that the original attributes of a route are preserved across the VPN prevents the formation of routing loops due to mutual redistribution between the two networks.

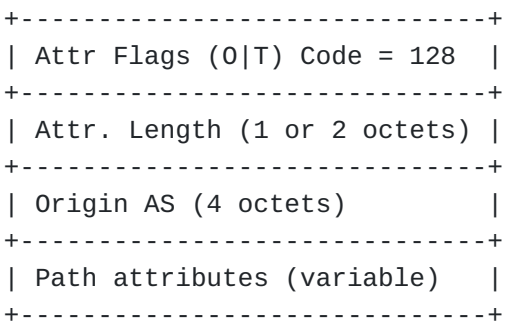
## 5. BGP customer route attributes

In order to transparently carry the BGP Path Attributes of customer routes, this document defines a new BGP Path Attribute:

- *ATTR_SET (type code 128)

- *ATTR_SET is an optional transitive attribute that carries a set of BGP path attributes. An attribute set (ATTR_SET) can include any BGP attribute that can occur in a BGP UPDATE message, except the MP_REACH and MP_UNREACH attributes.

The ATTR_SET attribute is encoded as follows:




The Attribute Flags are encoded according to RFC4271 *[RFC4271]*. The Extended Length bit determines whether the Attribute Length is one or two octets.
The attribute value consists on a 4 octet "Origin AS" value followed by a variable length field which conforms to the BGP UPDATE message path attribute encoding rules. The attribute length is 4 plus the total length of the encoded attributes.
This attribute is used by a PE router to store the original set of BGP attributes it receives from a CE. When a PE router advertises a PE-received route to a CE, it will use the path attributes carried in the ATTR_SET attribute.
In other words, the BGP Path Attributes are "pushed" into this stack like attribute when the route is received by the VPN and "popped" when the route is advertised in the PE to CE direction.
Using this mechanism isolates the customer network from the attributes used in the customer network and vice versa. Attributes as the route reflection cluster list attribute are segregated such that customer network cluster identifiers won't be considered by the customer network route reflectors and vice-versa.

The Origin autonomous-system number is designed to prevent a route originating in a given autonomous system iBGP to be leaked into a different autonomous system, without proper AS_PATH manipulation. It SHOULD contain the autonomous-system number of the customer network that originates the given set of attributes. The value is encoded as a 32-bit unsigned integer in network byte order, regardless of whether or not the originating PE supports Four-octet AS Numbers *[RFC4893]*.
The AS_PATH and AGGREGATOR attributes contained within an ATTR_SET attribute MUST be encoded using Four-octet AS Numbers *[RFC4893]*, regardless of the capabilities advertised by the BGP speaker to which the ATTR_SET attribute is transmitted. BGP speakers that support the extensions defined in this document MUST also support RFC4893 *[RFC4893]*. The reason for this requirement is to remove ambiguity between two-octet and four-octet AS_PATH attribute encoding.
The NEXT_HOP attribute SHOULD NOT be included in an ATTR_SET. When present it SHOULD be ignored by the receiving PE. Future applications of the ATTR_SET attribute MAY define meaninful semantics for an included NEXT_HOP attribute.
The ATTR_SET attribute SHALL be considered malformed if any of the following applies:

- Its length is less than 4 octets.
- The original path attributes carried in the variable length attribute data include the MP_REACH or MP_UNREACH attribute.
- The included attributes are malformed themselves.

An UPDATE message with a malformed ATTR_SET attribute SHALL be handled as follows. If its Partial flag is set and its Neighbor-Complete flag is clear, the UPDATE is treated as a route withdraw as discussed in [I-D.ietf-idr-optional-transitive]. Otherwise (i.e. Partial flag is clear or Neighbor-Complete is set), the procedures of the BGP-4 base specification *[RFC4271]* MUST be followed with respect to an Optional Attribute Error.

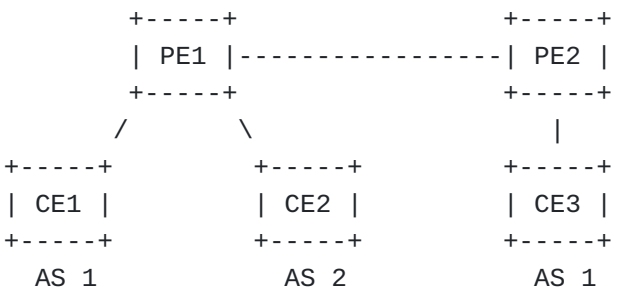
## 6. Next-hop handling

When BGP/MPLS VPNs are not in use, the NEXT_HOP attribute in iBGP routes carries the address of the border router advertising the route into the domain. The IGP distance to the NEXT_HOP of the route is an important component of BGP route selection.
When a BGP/MPLS VPN service is used to provide interconnection between different sites, since the customer network runs a different IGP domain, metrics between the provider and customer networks are not comparable.
However, the most important component of a metric is the inter-area metric, which is known to the customer network. The intra-area metric is typically negligible.

The use of route reflection, for instance, requires metrics to be configured so that inter-cluster/area metrics are always greater than intra-cluster metrics.
The approach taken by this document is to rewrite the NEXT_HOP attribute at the VRF import/export boundary. PE routers take into account the PE-PE IGP distance calculated by the customer network IGP, when selecting between routes advertised from different PEs.
An advantage of the proposed method is that the customer network can run independent IGPs at each site.

## 7. Exchanging routes between different VPN customer networks

In the traditional model, where External BGP sessions are used between the BGP/MPLS VPN PE and CE, the PE router identifies itself as belonging to the customer network autonomous-system.
In order to use Internal BGP sessions the PE router has to identify itself as belonging to the Customer AS. More specifically, the VRF that is used to interconnect to that customer site is assigned to the Customer AS rather than the VPN provider AS.
The Origin AS element in the ATTR_SET path attribute conveys the AS number of the originating VRF. This AS number is used in a receiving PE in order to identify route exchanges between VRFs in different ASes.
In scenarios such as what is commonly referred to an "extranet" VPN, routes MAY be advertised to both internal and external VPN attachments, belonging to different autonomous systems.




Consider the example given above where (PE1, CE1) and (PE2, CE3) sessions are iBGP. In BGP/MPLS VPNs, a route received from CE1 above may be distributed to the VRFs corresponding to the attachment points for CEs 2 and 3.
The desired result, in such a scenario is to present the internal peer (CE3) with a BGP advertisement that contains the same BGP Path Attributes received from CE1 and to the external peer (CE 2) a BGP advertisement that would correspond to a situation where AS 1 and 2 have an external BGP session between them.
It order to achieve this goal the following set of rules apply:

*When importing a VPN route that contains the ATTR_SET attribute into a destination VRF, a PE router MUST check that the "Origin AS" number contained in the ATTR_SET attribute matches the autonomous-system associated with the VRF.

*In case the autonomous-system numbers do match, the route is imported into the VRF with the attributes contained in the ATTR_SET attribute. Otherwise, in the case of an autonomous-system number mismatch, the set of attributes to be associated with the route SHALL be constructed as follows:

*  1. The path attributes are set to the attributes contained in the ATTR_SET attribute.

   2. Internal BGP specific attributes are discarded (LOCAL_PREF, ORIGINATOR, CLUSTER_LIST, etc).

   3. The "Origin AS" number contained in the ATTR_SET attribute is prepended to the AS_PATH following the rules that would apply to an external BGP peering between the source and destination ASes.

   4. If the autonomous-system associated with the VRF is the same as the VPN provider autonomous-system and the AS_PATH attribute of the VPN route is not empty, it SHALL be prepended to the AS_PATH attribute of the VRF route.

*When advertising the VRF route to an Exterior BGP peer, a PE router SHALL apply steps 1 to 4 defined above and subsequently prepend its own autonomous-system number to the AS_PATH attribute. For example, if the route originated in a VRF that supports Internal BGP peering and the ATTR_SET attribute and is advertised to a CE that is configured in the traditional Exterior BGP mode then both the originator AS, the VPN AS_PATH segment and the customer network AS are prepended to the AS_PATH.

*When importing a route without the ATTR_SET attribute to a VRF that is configured in a different autonomous-system, a PE router MUST prepend the VPN provider AS number to the AS_PATH.

In all cases where a route containing the ATTR_SET attribute is imported, attributes present on the VPN route other than the NEXT_HOP attribute are ignored, both from the point of view of route selection in the VRF Adj-RIB-in and route advertisement to a CE router. In other words, the information contained in ATTR_SET attribute overrides the VPN route attributes on "vrf-import".

## 8. Deployment considerations

It is RECOMMENDED that different VRFs of the same VPN (i.e. in different PE routers) which are configured with iBGP PE-CE peering sessions use different Route Distinguisher (RD) values. Otherwise (in the case where the same RD is used) the BGP IP VPN infrastructure may select a single BGP customer path for a given IP Network Layer Reachability Information (NLRI); without access to the detailed path information that is contained in the ATTR_SET attribute.

As mentioned previously, the model for this service is a "Route Server" where the IP VPN provides the customer network with all the BGP paths known by the CEs. This effectively implies the use of unique RDs per VRF.

The stated goal of this extension is to isolate the customer network from the BGP path attribute operations performed by the IP VPN and conversely isolate the service provider network from any attributes injected by the customer. For instance, BGP communities can be used to influence the behavior of the IP VPN infrastructure. Using this extension, the service provider network can transparently carry these attributes without interference with its operations.

Another example of unwanted interaction between customer and IP VPN BGP attributes is a scenario where the same Service Provider autonomous-system number is used both to provide Internet service as well as the IP VPN service. In this case, it is not uncommon to have a VPN customer route contain the AS Number of the Service Provider. The IP VPN network should work transparently in this case as in all others.

This protocol extension is designed to behave such that each PE VRF operates as a router in the configured AS. Previously VRFs operate in the provider network AS only. The VPN backbone provides interconnection between VRFs of the same AS, as well as interconnection between different ASes (subject to the appropriate policies). When interconnecting VRFs in the same AS, the VPN backbone operates as a top level Route Reflection mesh. When interconnecting VRFs in different ASes, the provider network provides an implicit peering relationship between the ASes that originate and import a specific route.

This extension is also applicable to scenarios where the VPN backbone spans multiple ASes. When the VPN backbone Inter-AS operation follows option b) or c) as defined in Section 10 of [RFC4364], the Provider networks are able to influence the route attributes and route selection of the VPN routes while providing a transparent service to the customer AS. Both internal BGP connectivity or extranets can be provided to the customer AS.

When VPN Provider networks interconnect via option a), there is no possibility of providing a fully transparent service. By definition option a) implies that each autonomous-system border router (ASBR) has a VRF associated with the customer VPN that is configured to operate in the respective Provider AS. These ASBR VRFs then communicate via eBGP with their peer Provider ASes.

In this case it is still possible to have all the customer VRFs with one Provider network to be configured in the same customer AS. This customer AS will then peer with the Provider AS implicitly at the ABSR. Which will in turn peer explicitly with a second Provider AS. This is not however a scenario in which transparency to the customer AS is possible.

## 9. Security considerations

It is worthwhile to consider the security implications of this proposal from two independent perspectives: the IP VPN provider and the IP VPN customer.
From an IP VPN provider perspective, this mechanism will assure separation between the BGP path attributes advertised by the customer CE router and the BGP attributes used within the provider network, thus potentially improving security.
Although this behavior is largely implementation dependent, currently it is possible for a CE device to inject BGP attributes (extended communities, for example) that have semantics on the IP VPN provider network, unless explicitly disabled by configuration in the PE.
With the rules specified for the ATTR_SET path attribute, any attribute that has been received from a CE is pushed into the stack before the route is advertised out to other PEs.
As with any other field based on values received from an external system, an implementation must consider the issues of input validation and resource management.
From the perspective of the VPN customer network, it is our opinion that there is no change to the security profile of PE-CE interaction. While having an iBGP session allows the PE to specify additional attributes not allowed on an eBGP session (e.g. local-pref), this does not significantly change the fact that the VPN customer must trust its service provider to provide it correct routing information.

## 10. IANA considerations

This document defines a new BGP path attribute which is part of a registry space managed by IANA. We request that IANA update its BGP Path Attributes registry with the value specified above (128) for the ATTR_SET path attribute.

## 11. Acknowledgments

The authors would like to thank Stephane Litkowski and Bruno Decraene for their comments.

## 12. References

### 12.1. 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. |
| **[RFC4364]** | Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006. |
| **[RFC4893]** | Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007. |
| **[RFC4456]** | Bates, T., Chen, E. and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006. |
| **[RFC5065]** | Traina, P., McPherson, D. and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007. |


### 12.2. Informative References


| | |
|---|---|
| **[I-D.ietf-idr-optional-transitive]** | Scudder, J, Chen, E, Mohapatra, P and K Patel, "Revised Error Handling for BGP UPDATE Messages", Internet-Draft draft-ietf-idr-optional-transitive-04, October 2011. |


## Authors' Addresses


Pedro Marques Marques EMail: pedro.r.marques@gmail.com

Robert Raszuk Raszuk Cisco Systems 170 W. Tasman Dr. San Jose, CA 95134 US EMail: raszuk@cisco.com

Keyur Patel Patel Cisco Systems 170 W. Tasman Dr.
San Jose, CA 95134 US EMail: keyupate@cisco.com

Kenji Kumaki Kumaki KDDI Corporation Garden Air Tower Iidabashi Chiyoda-ku, Tokyo 102-8460 Japan EMail: ke-kumaki@kddi.com

Tomohiro Yamagata Yamagata KDDI Corporation Garden Air Tower Iidabashi Chiyoda-ku, Tokyo 102-8460 Japan EMail: to-yamagata@kddi.com