Network Working Group
Internet-Draft
Intended status: Experimental
Expires: September 4, 2010

Marques
Raszuk
Patel
Cisco Systems
Kumaki
Yamagata
KDDI Corporation

March 3, 2010

Internal BGP as PE-CE protocol
draft-marques-l3vpn-ibgp-02

Abstract

This document defines protocol extensions and procedures for BGP PE-CE router iteration in BGP/MPLS IP VPN [RFC4364] 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

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

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 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."

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.

This Internet-Draft will expire on September 4, 2010.

Table of Contents

1. Introduction

In current deployments, when BGP is used as the PE-CE routing protocol, these peering sessions are typically configured as an external peering between the VPN provider AS and the customer network AS. At each External BGP boundary, Path Attributes [RFC4271] are modified as per standard BGP rules. This includes prepending the AS_PATH attribute with the autonomous system of the originating customer CE and the automomous system(s) of the provider edge 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 which use an hierarchical design are typically modeled such that the top-level core or backbone participates in a full iBGP mesh which distributes routing information between sites via BGP route reflection [RFC4456] or confederations [RFC5065]. This will be our service model definition.

2. 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 network 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 bellow 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.

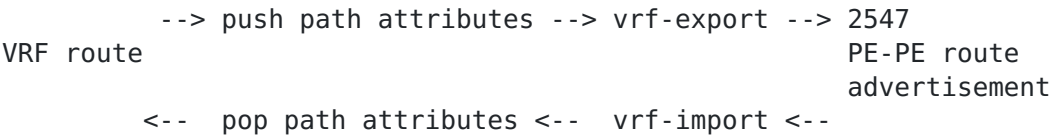
On a BGP/MPLS VPN PE, a CE-received route MUST be advertised to other VPN PEs that import the Route Targets which are associated with the route. This is independent of whether the CE route has been received as an external or internal route. However, a CE received route is not readvertised back to other CEs unless Route Reflection 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.

3. Path attributes



The diagram above shows the BGP path attribute stack processing in relation to existing 2547 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 poped after the Import Target processing step that identifies the VRF table in which a PE received route is accepted.

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, VPN 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 network prevents the formation of routing loops due to mutual redistribution between the two networks.

4. Carrying internal BGP routes

In order to carry the original BGP attributes of a route received from a CE, 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.

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 network 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 VPN 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 VPN network route reflectors and vice-versa.

The autonomous system number present in the ATTR_SET attribute 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 of the customer network that originates the given set of attributes.

The NEXT_HOP attribute SHOULD NOT be included in an ATTR_SET.

5. 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.

An important component of BGP route selection is the IGP distance to the NEXT_HOP of the route.

When a BGP/MPLS VPN service is used to provide interconnection between different sites, since the VPN network runs a different IGP domain, metrics between the VPN and customer networks are not comparable.

However, the most important component of a metric is the inter-area metric, which is known to the VPN 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.

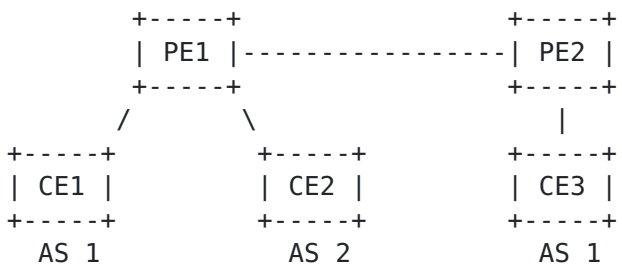
The approach taken by this document is to rewrite the NEXT_HOP attribute at the PE-CE boundary. PE routers take into account the PE-PE IGP distance calculated by the VPN network IGP, when selecting between routes advertised from different PEs.

An advantage of the proposed method is that the customer network can run independent IGPs at each site.

6. Exchanging routes between different VPN customer networks

A given VPN customer network SHOULD use internal or external BGP sessions consistently for peering sessions where the same autonomous system is used.

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 RFC2547 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 a external BGP session between them.

It order to achieve this goal the following set of rules apply:

When advertising an iBGP originated route to iBGP, a PE router MUST check that the autonomous-system contained in the ATTR_SET attribute matches the autonomous system of the CE to which the route is being advertised.

In case the autonomous-systems do match, the route is advertised with the attributes contained in the ATTR_SET attribute. Otherwise, in the case of an autonomous-system mismatch, the set of attributes to be advertised to the CE in question 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 autonomous-system 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. Internal BGP specific attributes corresponding to the configuration of destination AS (LOCAL_PREF) are added.

When advertising an iBGP originated route to eBGP, a PE router shall apply steps 1 to 3 defined above and subsequently prepend its own autonomous-system number to the AS_PATH attribute (i.e. both the originator and VPN network as numbers are prepended).

When advertising an eBGP originated route to iBGP, a PE router MUST prepend its own as number before adding iBGP only as-path attributes (LOCAL_PREF).

In all cases where an iBGP originating route is processed, 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.

7. Contributors

8. 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 a 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.

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.

9. 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 registry with the value specified above (128) for the ATTR_SET path attribute.

10. Normative References

[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.

[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.

Authors' Addresses

Pedro Marques
Cisco Systems
170 W. Tasman Dr
San Jose, CA 95134
US

Email: roque@cisco.com

Robert Raszuk
Cisco Systems
170 W. Tasman Dr
San Jose, CA 95134
US

Email: raszuk@cisco.com

Keyur Patel
Cisco Systems
170 W. Tasman Dr
San Jose, CA 95134
US

Email: keyupate@cisco.com

Kenji Kumaki
KDDI Corporation
Garden Air Tower
Iidabashi
Chiyoda-ku, Tokyo 102-8460
JAPAN

Email: ke-kumaki@kddi.com

Tomohiro Yamagata
KDDI Corporation
Garden Air Tower
Iidabashi
Chiyoda-ku, Tokyo 102-8460
JAPAN

Email: to-yamagata@kddi.com