BGP/IP VPNs: BGP and CE-Based Virtual Private Networks

draft-berger-l3vpn-ip-tunnels-01.txt

Status of this Memo

Copyright Notice

Abstract
   This memo describes a routing architecture that is most applicable to Customer Edge (CE)-based Virtual Private Networks (VPNs).

   In this architecture, customer devices use BGP to exchange VPN routes with one another. The BGP UPDATES include a new attribute that identifies the endpoint of a tunnel that can be used to reach a particular VPN prefix. The encapsulation strategy described in this memo is more flexible than that described in RFC 4364. In this architecture, the edge router can encapsulate the original datagram twice, as in RFC 4364. In this case, the inner header provides VPN context and the outer header identifies the tunnel between edge routers. Alternatively, the edge router can encapsulate the original

datagram only once, with the tunnel providing both VPN context and
identifying a tunnel to the remote edge router.

Contents

1. Introduction

[RFC4110] provides a taxonomy for Layer 3 Virtual Private Networks
(VPNs). All VPNs share the following characteristics:

  - Customer enclaves are connected to a Service Provider (SP)
    network
  - Customer enclaves are assigned to a Virtual Routing and
    Forwarding (VRF) Instance
  - Customers can communicate within the VRF
  - Customers cannot communicate across VRF boundaries
  - Addressing is unique only within the VRF
  - VPN routing environments are isolated from one another
  - VPN routing environments are isolated from the service provider
    routing environment
  - Tunnels (MPLS, GRE, IPSec) connect customer enclaves to one
    another across the SP network

[RFC4110] divides VPNs into two classes. These are:

  - Provider Edge (PE)-based VPNs
  - Customer Edge (CE)-based VPNs

In a PE-based VPN, SP tunnels connect PE routers to one another.
BGP/MPLS IP VPNs, see [RFC4364] and [RFC4659], leverages this
architecture, using BGP to exchange VPN routes among PE routers. The
BGP advertisements specify tunnel endpoint as the BGP next-hop for
the VPN route. Therefore, SP interior routers need not carry VPN
routes.

In a CE-based VPN, tunnels connect CE routers to one another.
Therefore, the routing architecture used in BGP/MPLS IP VPNs is not
applicable. BGP/MPLS IP VPNs are also not applicable when MPLS is not
supported. This memo describes a routing architecture similar to the
routing architecture used in BGP/MPLS IP VPNs that is applicable to
CE-based VPNs. We refer to the approach described in this document as
"BGP/IP VPNs".

In this architecture, customer devices use BGP to exchange VPN routes
with one another. The BGP UPDATES include a new attribute that
identifies the endpoint of a tunnel that can be used to reach a
particular VPN prefix.

The encapsulation strategy described in this memo is more flexible
than that used in BGP/MPLS IP VPNs. In this architecture, the CE
router can encapsulate the original datagram twice, as in BGP/MPLS IP
VPNs. In this case, the inner header provides VPN context and the
outer header identifies the tunnel between CE routers. Alternatively,

the CE router can encapsulate the original datagram only once, with
the tunnel providing both VPN context and identifying a tunnel to the
remote CE.

1.1. Conventions used in this document

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. VPN Route Distribution

BGP/IP VPNs use much the same mechanisms as [RFC4364] and [RFC4659]
to support VPN route distribution. The specific mechanism differs
only in that no MPLS labels are required. This Section is modeled
after Section 4 of [RFC4364] and Section 3 of [RFC4659] and borrows
from those sources.

2.1. Address Families

This document reuses the VPN-IPv4 and VPN-IPv6 address families
specified in [RFC4364] and [RFC4659]. Different BGP encoding is used
for these families as MPLS labels are not used. This document adds
the prefix "IP" to the names of the address families, i.e., "IP VPN-
IPv4" and "IP VPN-IPv6", to identify the different encoding.

2.2. IP VPN-IPv4 NLRI and IP VPN-IPv6 NLRI Encoding

As with BGP/MPLS IP VPNs, BGP/IP VPN information is carried in the
Multiprotocol Reachable and Unreachable Network Layer Reachability
Information (NLRIs) introduced in [RFC4760]. Unlike BGP/MPLS IP
VPNs, the BGP/IP VPN encoding of the VPN-IPv4 and VPN-IPv6 address
families provide tunnel related information (rather than MPLS
labels). This difference is reflected in the use of new formats in
the "Network Address of Next Hop" and "Network Layer Reachability
Information" NLRI fields. In all other respects, including
formatting and processing, BGP/IP VPN route distribution is identical
to BGP/MPLS VPN route distribution. This includes the encoding of
the Route Distinguishers (RD) in prefixes and Route Target (RT)
Attributes in extended communities.

The use of BGP/IP VPN related NLRI formats is indicated by the SAFI
value TBA (by IANA). The address family of a VPN route determines
the type of IP VPN NLRI which, as typical, is indicated by the AFI.

   IPv4 VPN routes MUST be encoded in an IP VPN-IPv4 NLRI with an AFI of
   1. IPv6 VPN routes MUST be encoded in an IP VPN-IPv6 NLRI with an AFI
   of 2. With the exception of the "Network Address of Next Hop" and
   "Network Layer Reachability Information" NLRI fields, both types of
   NLRIs MUST be encoded as defined in [RFC4364] (for the IP VPN-IPv4
   NLRI) and in [RFC4659] (for the IP VPN-IPv4 NLRI).

2.2.1. BGP/IP VPN - Network Address of Next Hop

   With IP VPN NLRIs, the Network Address of Next Hop NLRI field carries
   the IP address and related tunnel parameters that should be used to
   reach a particular VPN route. The IP address is the address to be
   used by receiving routers as the destination tunnel end-point
   address. Tunnel parameters indicate the type of tunnel supported by
   the advertising router for the particular route. Tunnel parameters
   may optionally indicate tunnel specific information, e.g, destination
   port or other tunnel identifying information. The format of the
   Network Address of Next Hop field used in BGP/IP VPN route
   distribution is:

```
       +---------------------------+
       | Tunnel Flags (1 octet)    |
       +---------------------------+
       | Tunnel Type (1 octet)     |
       +---------------------------+
       | Tunnel Address (variable) |
       +---------------------------+
       | Tunnel Params. (variable) |
       +---------------------------+
```

   The use and meaning of these fields are as follows:

      Tunnel Flags: (Bit Field)

         The following flags are defined:

```
             0 1 2 3 4 5 6 7
            +-+-+-+-+-+-+-+-+
            |V|   Reserved  |
            +-+-+-+-+-+-+-+-+
```

         Version (V):

            The Version bit is used to indicate the IP version of the
            Tunnel Address field. The flag MUST be cleared, i.e., zero
            (0), when the tunnel end-point address carried in the Tunnel
            Address field is an IPv4 address. The flag MUST be set,

i.e., one (1), when the Tunnel Address field carries an IPv6 address.

Tunnel Type: (Unsigned Integer)

This field indicates the type of tunnel that should be used to transport data associated with the advertised VPN route. Note that the mechanisms related to establishment and management of tunnels are outside the scope of this document. The following Type values are defined:

```
Value   Type
-----   ---------------------------------------
  0     Reserved
  1     Generic Route Encapsulation (GRE) [RFC1701]
  2     IP-in-IP [RFC2003]
  3     IP Authentication Header in the Tunnel-mode (AH)
        [RFC4301]
  4     IP Encapsulating Security Payload in the Tunnel-mode
        (ESP) [RFC4303]
  5     Reserved (contact authors)
```

Tunnel Address:

The Tunnel Address field contains the destination tunnel end-point IP address that SHOULD be used to reach the advertised VPN route. The type of address and the size of the field can be determined by examining the V-bit. When the V-bit is not set, i.e., zero (0), the Tunnel Address field MUST contain a 4-octet IPv4 address. When the V-bit is set, i.e., one (1), the Tunnel Address field MUST contain a 16-octet IPv6 address.

Tunnel Parameters:

The Tunnel Parameters field contains a series of variable-length data items called Tunnel Parameter, or TP, subobjects. Each TP subobject has the following format:

```
    +---------------------------+
    |   TP Type (1 octet)       |
    +---------------------------+
    |   TP Length (1 octet)     |
    +---------------------------+
    |   TP Contents (variable)  |
    +---------------------------+
```

The use and meaning of these fields are as follows:

Tunnel Parameter (TP) Type: (Unsigned Integer)

This field indicates the type of tunnel parameters
contained in the TP subobject. Two value ranges are
defined. The first range is used for parameters that are
independent of any particular tunneling technology and
are shared across all Tunnel Type field values. The
second range is specific to, and only has meaning in the
context of a particular Tunnel Type field values, i.e.,
identified by the tuple <Tunnel Type, TP Type>. The
value ranges are as followed:

```
  Range   Use
  -------  ----------------------------------------
    0-63  Common tunnel parameters (Applies to all
          Tunnel Types)
  64-127  Tunnel Type (technology) specific tunnel
          parameters
 128-255  Reserved
```

TP Type one (1) is defined below in Section 2.2.1.1.

Tunnel Parameter (TP) Length: (Unsigned Integer)

The TP Length field contains of the total length of the
tunnel parameters subobject in octets, including the TP
Type and TP Length fields. The TP Length field MUST be
equal to or greater than 2.

Tunnel Parameter (TP) Contents:

The actual information carried in the subobject.

Subobjects with TP subobject types that are not recognized by a
receiver SHOULD be silently ignored.

2.2.1.1. Alternate Address Tunnel Parameter (AA-TP) Subobject

The Alternate Address Tunnel Parameter (AA-TP) Subobject is used to
provide an additional destination tunnel end-point IP address. When
this subobject is present, the address provided in the subobject
along with the address provided in the next hop Tunnel Address field
and any other AA-TP Subobjects SHOULD be used as "equal cost" next
hops. Multiple AA-TP Subobjects MAY be included in a NLRI Next Hop.
The format of the AA-TP Subobject is:

```
+---------------------------+
|   TP Type (=1)            |
+---------------------------+
|   TP Length (=6 or =18)   |
+---------------------------+
|   AA (variable)           |
+---------------------------+
```

The use and meaning of these fields are as follows:

TP Type:

The AA-TP Subobject may be used with any type of tunnel and
MUST use the TP Type of 1.

TP Length:

Per the definition of TP Length, see above, this field is set
to the length of the AA field in octets plus 2. TP Length MUST
be set to 6 when the V bit is not set, i.e. zero (0), and MUST
be set to 18 when the V bit is set, i.e., one (1).

Alternate Address (AA):

The Alternate Address (AA) field contains an address of the
type identified by the V bits in the Tunnel Flags field. The
field MUST contain an IPv4 address when the V bit is not set
(0), and an IPv6 address when the V bit is set (1).

2.2.2. BGP/IP VPN Prefix Information

BGP/IP VPN Prefix Information parallels the label mapping information
used in BGP/MPLS VPNs, and the general definition and processing of
BGP/IP VPN Prefix Information follows [RFC3107]. As with label
mapping information tunnel, BGP/IP VPN Prefix Information is carried
as part of an NLRI in the Multiprotocol Extensions attributes.

2.2.2.1. BGP/IP VPN Prefix Route Information Encoding

The Network Layer Reachability information is encoded as one or more
triples of the form <length, next hop token, prefix>. The Next Hop
Token corresponds to the Network Address of Next Hop field of the
NRLI. Prefix contains the reachable VPN route. When a router
advertises the same Prefix with multiple NLRI Next Hop fields, the
advertiser uses different Next Hop Tokens for each next hop. This
allows the router to independently withdraw each advertised route.

The format of the BGP/IP VPN Prefix Route Information is:

```
+---------------------------+
|   Length (1 octet)        |
+---------------------------+
|   NH Token (1 octet)      |
+---------------------------+
|   Prefix (variable)       |
+---------------------------+
```

The use and meaning of these fields are as follows:

Length: (Unsigned Integer)

   The Length field indicates the length in bits of the address
   prefix. The size of the Length and NH Token fields MUST NOT be
   included.

Next Hop (NH) Token: (Unsigned Integer)

   An identifier that within the scope of the advertising router
   uniquely identifies the contents of the Network Address of Next
   Hop field associated with this route. All routes advertised
   with the same Next Hop field SHOULD have the same NH Token
   value. Routes advertised with different Next Hop field values
   MUST have different value NH Tokens. Note, zero (0) is a valid
   NH Token field value.

Prefix:

   This field contains the IP VPN prefix that is being advertised.
   Both unicast and multicast prefixes MAY be carried in this
   field.

   The format of this field is the same as defined in [RFC4364].
   The [RFC4364] definition is based on the following definition
   from [RFC3107]: "The Prefix field contains address prefixes
   followed by enough trailing bits to make the end of the field
   fall on an octet boundary. Note that the value of trailing
   bits is irrelevant." Additionally, per [RFC4364], the Prefix
   field includes an 8 octet RD.

   The length of the Prefix field can be determined by examining
   the V-bit. When the V-bit is not set, i.e., zero (0), the
   Prefix field is a 12 octet quantity which MUST contain an
   8-octet RD followed by a 4-octet IPv4 address. When the V-bit
   is set, i.e., one (1), the Prefix field is a 24-octet quantity
   which MUST contain an 8-octet RD followed by a 16-octet IPv6

address.

[2.2.2.2](#). Advertising Multiple Routes to a Destination

A BGP speaker may maintain (and advertise to its peers) more than one
route to a given destination. Each of these routes can be advertised
using separate NLRIs with different Network Address of Next Hops, or
via the AA-TP Subobject defined above in [Section 2.2.1.1](#). When
routes are advertised with different NLRI Next Hops, the routes will
have different NH Tokens. Routes, with independent NH Tokens, may be
independently withdrawn. When the AA-TP Subobject is used, all next
hops included in the same advertisement will share the same NLRI Next
Hop field and will be covered under the same NH Token and, therefore,
can only be withdrawn as a group.

[2.3](#). BGP Capability Negotiation

In order for two edge routers to exchange IP VPN-IPv4 and VPN-IPv6
NLRIs, they MUST use BGP Capabilities Negotiation to ensure that they
both are capable of properly processing such NLRIs. This is done as
specified in [[RFC4760](#)] and [[RFC3392](#)], by using capability code 1
(multiprotocol BGP), with AFI and SAFI values as specified above,
[Section 2.2.1](#) and 2.2.2.

[3](#). Forwarding

This Section is modeled after [Section 5 of [RFC4364]](#) and [Section 4 of](#)
[[RFC4659]](#) and borrows from those sources.

When an edge router receives an IP packet from a CE device, the edge
router MUST choose a particular VRF in which to look up the packet's
destination address. This choice is typically based on the packet's
ingress attachment circuit. The router MUST then look for a best
match for the packet's destination IP address within the VRF.

If the packet's next hop is reached directly over a VRF attachment
circuit (see definition in [[RFC4364](#)]) from the processing edge router
(i.e., the packet's egress attachment circuit is on the same edge
router as its ingress attachment circuit), then the packet MUST be
sent on the egress attachment circuit.

If the ingress and egress attachment circuits are on the same edge
router, but are associated with different VRFs, and if the route that
best matches the destination address in the ingress attachment
circuit's VRF is an aggregate of several routes in the egress

attachment circuit's VRF, it may be necessary to look up the packet's destination address in the egress VRF as well.

If the packet's next hop is NOT reached through a VRF attachment circuit, then the packet must travel at least one hop through the backbone. The packet thus has a "BGP VPN Next Hop", which will have been advertised per Section 2.2. The packet must then be tunneled to the BGP VPN Next Hop.

The packet MUST be encapsulated in a tunnel according to the type specified in the NLRI Next Hop of the advertised route. The encapsulated packet MUST then be forwarded as a standard IP packet. As previously mentioned, the specifics of tunnel establishment are outside the scope of this document.

When the packet arrives at the destination tunnel end-point, it will be at the BGP VPN Next Hop. The BGP VPN Next Hop MUST strip the tunnel encapsulation, and MUST identify how the received packet is to be forwarded. The tunnel destination address will typically indicate the outgoing VRF. In this case, the packet's original IP destination address MUST be looked up in a particular VRF before being forwarded to a CE device. In other cases, the tunnel's destination address will determine the packet's egress attachment circuit. In this case, a lookup (e.g., ARP) may still need to be done in order to determine the packet's data link header on that attachment circuit.

4. Security Considerations

This section borrows from Section 11 of [RFC4659]. The extensions defined in this document allow [RFC4271] to propagate reachability information about IPv4 and IPv6 VPN routes. Propagation of VPN routes within BGP is already defined in [RFC4364] and [RFC4659].

Security considerations for the transport of IPv4 and IPv6 reachability information using BGP are discussed in [RFC4271] and [RFC2545], respectively, and are equally applicable for the extensions described in this document.

The extensions described in this document for offering VPNs over IP tunnels use the same BGP based route distribution approach as the approach described in [RFC4364] and [RFC4659]. Therefore, the same security considerations apply with regards to Control Plane security, and edge router and P device security as described in [RFC4364], Section 13.

This document uses IP based tunnel technologies to support data plane transport. Consequently, the security considerations of those tunnel

technologies apply. This document defines support for GRE [RFC1701], IP-in-IP [RFC2003] and IPsec AH [RFC4301] and ESP [RFC4303]. The security considerations from those documents apply to the data plane aspects of this document.

5. IANA Considerations

IANA is requested to administer assignment of new namespaces and new values for namespaces defined in this document and reviewed in this section.

5.1. BGP/IP VPN SAFI

Upon approval of this document, the IANA will make the assignment in the Subsequence Address Family Identifiers (SAFI) registry located at http://www.iana.org/assignments/safi-namespace:

```
Value    Description                                  Reference
-----    -----------                                  ---------
141*     BGP/IP VPN address                       [This document]
```

(*) Suggested value.

5.2. BGP/IP VPN Tunnel Types

Upon approval of this document, the IANA will establish a new registry called the "BGP/IP VPN Tunnel Types registry". This registry should be established with the following initial values.

```
   Value   Type
   -----   ----------------------------------------
       0   Reserved
       1   Generic Route Encapsulation (GRE) [RFC1701]
       2   IP-in-IP [RFC2003]
       3   IP Authentication Header in the Tunnel-mode (AH)
           [RFC4301]
       4   IP Encapsulating Security Payload in the Tunnel-mode
           (ESP) [RFC4303]
       5   Reserved
```

Future assignments are to be made using either the IETF Consensus process defined in [RFC2434], or the Early IANA Allocation process defined in [RFC4020]. Reserved values MUST NOT be reassigned without permission of the authors of this document.

5.3. BGP/IP VPN Tunnel Parameter Subobject Types

   Upon approval of this document, the IANA will establish a new
   registry called the "BGP/IP VPN Tunnel Parameter Subobject Types
   registry". The assignable values are broken down into the following
   ranges:

```
     Range   Use
     -------  --------------------------------------------------------
       0-63   Common tunnel parameters (Applies to all Tunnel Types)
      64-127  Tunnel Type (technology) specific tunnel parameters
     128-255  Reserved
```

   This registry should be established with the following initial
   values:

```
     Value   Tunnel Parameter (TP) Type
     -----   -------------------------------------------
         0   Reserved
         1   Alternate Address Tunnel Parameter Subobject
```

   Assignments in the range of 64-127 MUST be made in the context of a
   particular BGP/IP VPN Tunnel Type, see Section 5.3, i.e., assignments
   take the form of <Tunnel Type, TP Type>.

   Future assignments in the range of 0-127 are to be made using either
   the IETF Consensus process defined in [RFC2434], or the Early IANA
   Allocation process defined in [RFC4020]. Assignments in the range of
   128-255 require Standards Action, which may impact how subsequent
   allocations within this range are to be made.

6. References

6.1. Normative References

   [RFC3392]   Chandra, R. and J. Scudder, "Capabilities
               Advertisement with BGP-4", RFC 3392, November 2002.

   [RFC4271]   Rekhter, Y., Ed. et al, "A Border Gateway Protocol 4
               (BGP-4)", RFC 4271, January 2006.

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

   [RFC4659]   De Clercq, J., et al, "BGP-MPLS IP Virtual Private
               Network (VPN) Extension for IPv6 VPN", RFC 4659,
               September 2006.

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

6.2. Informative References

   [BGP-VPN]   Ould-Brahim, H., et al, "BGP/VPN: VPN Information
               Discovery for Network-based VPNs", Work in progress,
               July 2000.

   [RFC1701]   Hanks, S., Li, T., Traina, P., "Generic Routing
               Encapsulation (GRE)", RFC 1701, October 1994.

   [RFC2003]   Perkins, C., "IP Encapsulation within IP", RFC 2003,
               October 1996.

   [RFC2434]   Narten, T. and H. Alvestrand, "Guidelines for Writing an
               IANA Considerations Section in RFCs", BCP 26, RFC 2434,
               October 1998.

   [RFC2545]   Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol
               Extensions for IPv6 Inter-Domain Routing", RFC 2545,
               March 1999.

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

   [RFC3107]   Rekhter, Y. and E. Rosen, "Carrying Label Information
               in BGP-4", RFC 3107, May 2001.

   [RFC4110]   Callon, R., Suzuki, M., "A Framework for Layer 3
               Provider-Provisioned Virtual Private Networks
               (PPVPNs)", RFC 4110, July 2005.

   [RFC4301]   Kent, S., Seo, K., "Security Architecture for
               the Internet Protocol", RFC 4301, December 2005.

   [RFC4303]   Kent, S., "IP Encapsulating Security Payload (ESP)"
               RFC 4303, December 2005.

   [RFC4020]   Kompella, K. and A. Zinin, "Early IANA Allocation of
               Standards Track Code Points", BCP 100, RFC 4020,
               February 2005.

7. Acknowledgments

   This work has similarities to [BGP-VPN], but does not draw from that
   work. Several sections of this document are modeled after and use
   text from [RFC4364] and [RFC4659]. The very useful ASCII drawing
   tool JavE (www.jave.de) was used to create Figures 1 and 2.

8. Authors' Addresses

   Lou Berger
   LabN Consulting, L.L.C.
   Phone: +1-301-468-9228
   Email: lberger@labn.net

   Ronald P. Bonica
   Juniper Networks
   2251 Corporate Park Drive
   Herndon, VA 20171
   USA
   Email: rbonica@juniper.net

   Russ White
   Cisco Systems
   Email: riw@cisco.com

Intellectual Property

   The IETF takes no position regarding the validity or scope of any
   Intellectual Property Rights or other rights that might be claimed to
   pertain to the implementation or use of the technology described in
   this document or the extent to which any license under such rights
   might or might not be available; nor does it represent that it has
   made any independent effort to identify any such rights. Information
   on the procedures with respect to rights in RFC documents can be
   found in BCP 78 and BCP 79.

   Copies of IPR disclosures made to the IETF Secretariat and any
   assurances of licenses to be made available, or the result of an
   attempt made to obtain a general license or permission for the use of
   such proprietary rights by implementers or users of this
   specification can be obtained from the IETF on-line IPR repository at
   http://www.ietf.org/ipr.

   The IETF invites any interested party to bring to its attention any
   copyrights, patents or patent applications, or other proprietary
   rights that may cover technology that may be required to implement
   this standard. Please address the information to the IETF at ietf-
   ipr@ietf.org.

Acknowledgement

Generated on: Wed Oct 24 15:59:02 EDT 2007