INTERNET DRAFT J. De Clercq, G. Gastaud, D. Ooms
<draft-ooms-v6ops-bgp-tunnel-00.txt> Alcatel
S. Prevost
BTexact
F. Le Faucheur
Cisco
October, 2002
Expires April, 2003

Connecting IPv6 Islands across IPv4 Clouds with BGP

Status of this Memo

Abstract

This document explains how to interconnect IPv6 islands over an IPv4 cloud, including the exchange of IPv6 reachability information using BGP. Two approaches will be explained, both requiring a Dual Stack MP-BGP-speaking edge router per IPv6 island. The hosts in the IPv6 islands can use native IPv6 addresses.

The first approach uses MP-BGP over IPv4, relies on identification of the MP-BGP-speaking edge routers by their IPv4 address and uses a trivial tunneling mechanism without any explicit tunnel configuration. The second approach uses MP-BGP over IPv6 and relies on existing ngtrans tunneling mechanisms to tunnel packets.

Table of Contents

Changes
ngtrans history (draft-ietf-ngtrans-bgp-tunnel-0x.txt)
00->01: editorial changes
extended section 4
01->02: editorial changes
added tunnel-specific considerations
added case of multiple IPv4 domains between IPv6 islands
added discussion on v6[v4]addresses in appendix A
02->03: complete rewrite: it turned out that two interpretations of the previous drafts existed, the two different interpretations are described explicitly in this version
03->04: renaming of the two approaches
editorial changes
clearly indicate which part requires standards track
04->05: added 5.1.3 to clarify how DS-BGP routers agree on tunnel type

v6ops history (draft-ooms-v6ops-bgp-tunnel-0x.txt)
05->00 individual submission: no changes. The document passed ngtrans last call early 2002, but the transfer to the IESG was postponed because of the reorg and closing down of ngtrans.

1. Introduction

This document explains how to interconnect IPv6 islands over an IPv4 cloud, including the exchange of IPv6 reachability information using

BGP. Two approaches will be explained, both requiring a Dual Stack MP-BGP-speaking edge router per IPv6 island. The hosts in the IPv6 islands can use native IPv6 addresses.

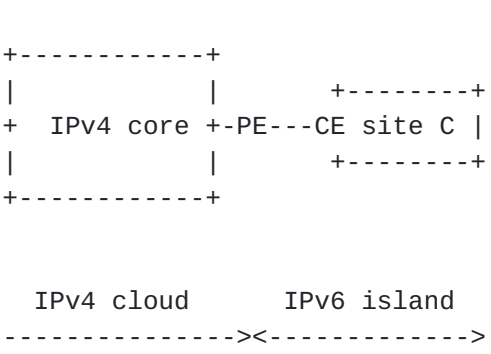
The first approach uses MP-BGP over IPv4, relies on identification of the MP-BGP-speaking edge routers by their IPv4 address and uses a trivial tunneling mechanism without any explicit tunnel configuration. The second approach uses MP-BGP over IPv6 and relies on existing ngtrans tunneling mechanisms to tunnel packets.

Most of the document is informational. Only section 4 and section 5 contain a parts that require standardization. Those are clearly identified through the use of the keywords "MUST", "MAY", etc. in accordance with [KEYWRD].

2. Terminology

The terminology of [IPV6] and [TRANS] applies to this document. We also use some of the terminology of [VPN].

In this document an 'IPv6 island' is an IPv6-upgraded network (which can be cross-AS). A typical example of one island would be one or more Customer IPv6 sites connected via their Customer Edge (CE) router to one (or more) Dual Stack Provider Edge (PE) router(s) of a Service Provider.



3. Applicability

The transition methods described in this document typically applies to an ISP that is familiar with BGP (possibly already offering BGP/MPLS VPN services) and that wants to offer IPv6 services to some of its customers. However, the ISP does not (yet) want to upgrade its network core to IPv6. With the mechanisms described here, the provider only has to upgrade some PE routers in some POPs to Dual Stack MP-BGP routers. The Dual Stack MP-BGP routers provide access to IPv6 customers and may provide access to IPv4 customers in addition.

The ISP may also have access to the global IPv6 Internet. The ISP provides global IPv6 connectivity through its peering relationship with an upstream ISP, or by peering relationships with other IPv6 ISPs in the default free routing zone (DFZ).

A Dual Stack MP-BGP router in the provider's network is connected to an upstream IPv6 ISP or forms part of the IPv6 backbone network, such as the 6bone. The ISP advertises IPv6 reachability of its IPv6 allocated prefix using MP-BGP to its IPv6 upstream provider or into the IPv6 DFZ. The IPv6 prefixes received from the upstream provider or from the DFZ can be redistributed within the ISP using MP-BGP.

The interface between the CE router and the PE router is a native IPv6 interface which can be physical or logical. A routing protocol (IGP or EGP) may run between the CE router and the PE router for a customer IPv6 site to exchange its reachability. Alternatively, static routes and/or a default route may be used on PE and CE to control reachability. A customer site may connect to the provider network over more than one interface.

The methods in this document can be used for customers that already have an IPv4 service from the network provider and additionally require an IPv6 service, as well as for customers that require only IPv6 connectivity. In both cases the network provider allocates IPv6 addresses to the site.

4. Description

Each IPv6 site is connected to at least one Dual Stack MP-BGP-speaking edge router that is located on the border with the IPv4 cloud. We refer to such a router as a DS-BGP router. The DS-BGP router MUST have at least one IPv4 address on the IPv4 side and one IPv6 address on the IPv6 side. The IPv4 address MUST be routable in the IPv4 cloud.

We refer to the DS-BGP router receiving IPv6 packets from an IPv6 site as an Ingress DS-BGP router (relative to these IPv6 packets). We refer to a DS-BGP router sending IPv6 packets to an IPv6 site as an Egress DS-BGP router (relative to these IPv6 packets).

No extra routes will be injected in the IPv4 cloud.

Interconnecting IPv6 islands over an IPv4 cloud requires following steps:

(1) Exchange IPv6 reachability information among DS-BGP Routers:

(1.a) The DS-BGP routers exchange, via MP-BGP [MP-BGP], IPv6 reachability information over the IPv4 cloud with their peers.

(1.b) In doing so, the Egress DS-BGP routers announce themselves as the BGP Next Hop.

(2) Tunnel IPv6 packets from Ingress DS-BGP Router to Egress DS-BGP Router: the Ingress DS-BGP router tunnels an IPv6 packet over the IPv4 cloud towards the Egress DS-BGP router identified as the BGP Next Hop in step (1.b) for the packet's destination IPv6 address.

We distinguish two approaches for connecting IPv6 islands across IPv4 clouds via BGP, which are respectively referred to as the "MP-BGP over IPv4" approach and the "MP-BGP over IPv6" approach.

Step (1.a) is identical for both approaches.

Steps (1.b) and (2) differ between the two approaches.

4.1. "MP-BGP over IPv4" approach

With this approach, the DS-BGP routers MUST run MP-BGP over an IPv4 stack (MP-BGP/TCP/IPv4). The DS-BGP router conveys to its peer its IPv4 address as the BGP Next Hop.

Since MP-BGP requires that the BGP Next Hop is of the same address family as the NLRI, this IPv4 address needs to be embedded in an IPv6 format. The IPv4-mapped IPv6 address is defined in [V6ADDR] as an "address type used to represent the addresses of IPv4 nodes as IPv6 addresses", thus this precisely fits for the above purpose. Encoding the routable IPv4 address into a IPv4-mapped IPv6 address allows the remote DS-BGP router to automatically tunnel data over the IPv4 cloud to the destination IPv6 island. Any type of encapsulation can be used (IPv4, MPLS, GRE).

In the "MP-BGP over IPv4" approach the IPv4 address of the MP-BGP next hop MUST be encoded as an IPv4-mapped IPv6 address.

The ingress DS-BGP Router MUST tunnel IPv6 data over the IPv4 cloud towards the Egress DS-BGP router identified by the IPv4 address advertised in the IPv4-mapped IPv6 address of the BGP Next Hop for the corresponding IPv6 prefix.

4.2. "MP-BGP over IPv6" approach

With this approach, the DS-BGP routers MUST run MP-BGP over an IPv6 stack (MP-BGP/TCP/IPv6). The DS-BGP router conveys to its peer its IPv6 address as the BGP Next Hop. The transport of MP-BGP messages as

well as IPv6 packets over the IPv4 cloud relies on any existing ngtrans tunneling technique ([6TO4], [ISATAP], [TRANS], ...). Thus, the IPv6 address of the BGP Next Hop MUST match the actual ngtrans tunneling technique used. For example, if ISATAP is used as the IPv6 over IPv4 tunneling technique, then the IPv6 address of the BGP Next Hop MUST be an ISATAP address.

The ingress DS-BGP Router MUST tunnel IPv6 data over the IPv4 cloud towards the Egress DS-BGP Router using the relevant ngtrans tunnelling technique applied to the IPv6 address advertised as the BGP Next Hop for the corresponding IPv6 prefix.

4.3. Characteristics Common To Both Approaches

For both approaches, the MP-BGP AFI MUST be IPv6 (value 2). The SAFI depends on the details of the tunneling technique used. This is discussed below in the tunnel technique specific section.

When the number of PEs is not too high, PEs MAY peer in a meshed fashion. Otherwise Route Reflectors MAY be used.

The hosts in the IPv6 island MAY have native IPv6 addresses. This is different from e.g. 6to4 [6TO4], which requires that special addresses (6to4 addresses) are allocated to the IPv6 hosts.

5. Tunneling

5.1. "MP-BGP over IPv4" approach

In the "MP-BGP over IPv4" approach, the IPv4-mapped IPv6 addresses allow a DS-BGP router that has to forward a packet to automatically determine the IPv4 endpoint of the tunnel by looking at the MP-BGP routing info.

If this approach is used to connect to the public IPv6 Internet, tunnels without special security mechanisms MAY be used (e.g. IPv4 tunnels [TUNNEL], GRE tunnels [GRE] or MPLS LSPs without IPSec).

Note that even when the number of peers is high, the number of tunnels is not a scalability concern from an operational viewpoint since those are automatic tunnels and thus require no configuration.

Considerations on 'common tunneling techniques' in [TRANS] are valid for this approach.

5.1.1. Tunneling over IPv4/GRE tunnels

When tunneling is done using IPv4 or GRE Tunnels, the SAFI used in MP-BGP MUST be one of the basic values: unicast, multicast or both (1, 2 or 3).

The Ingress DS-BGP Router MUST simply use the IPv4 address of the BGP next hop as the destination address of the prepended tunneling header. It uses one of its IPv4 addresses as the source address of the prepended tunneling header.

5.1.2. Tunneling over MPLS LSPs

When the IPv4 backbone supports MPLS, MPLS LSPs MAY be used as the tunneling technique. These LSPs can be established using any existing technique (LDP, RSVP, ...).

When MPLS LSPs are used with the "MP-BGP over IPv4" approach, rather than successively prepend an IPv4 header and then perform label imposition based on the IPv4 header, the ingress DS-BGP Router MAY directly perform label imposition of the IPv6 header without prepending any IPv4 header. The (outer) label imposed corresponds to an LSP starting on the ingress DS-BGP Router and ending on the egress DS-BGP Router.

While the "MP-BGP over IPv4" approach can operate using a single level of labels, there are advantages in using a second level of labels which are bound to IPv6 prefixes via MP-BGP advertisements in accordance with [LABEL]. For instance, use of a second level label allows Penultimate Hop Popping (PHP) on the Label Switch Router (LSR) upstream of the egress DS-BGP router without any IPv6 capabilities/upgrade on the penultimate router even when the IPv6 packet is directly encapsulated in MPLS (without an IPv4 header); since it still transmits MPLS packets even after the PHP (instead of having to transmit IPv6 packets and encapsulate them appropriately). Thus, the "MP-BGP over IPv4" approach MAY be used with a single label and MAY also be used with a second label.

Where a single level of labels is used and no label is advertised by MP-BGP, the SAFI used in MP-BGP MUST be one of the basic values: unicast, multicast or both (1, 2 or 3).

Where two levels of labels are used and labels are advertised by MP-BGP, the SAFI used in MP-BGP MUST be the "label" SAFI (4) or the "VPN" SAFI (128) depending on the procedures for allocating these labels.

5.1.3. Tunnel Type determination

There is work underway on a new BGP extended community attribute:

Tunnel Type [TTYPE]. This extended community attribute permits a BGP router to specify which tunnel type MUST be used to tunnel data to itself. The "MP-BGP over IPv4" approach allows the use of this new BGP extended community attribute, but it also wants to define a default tunnel type that must be used in the absence of this Tunnel Type extended community attribute. The default behavior is:

1. If a labeled route is advertised (an advertisement with SAFI 4 or 128), the default tunnel must be MPLS (yielding two levels of labels).

2. If an unlabeled route is advertised (SAFI is not 4 or 128), the default tunnel must be IP-in-IP.

5.2. "MP-BGP over IPv6" approach

As said before, the "MP-BGP over IPv6" approach relies on any existing ngtrans tunneling mechanism to carry the IPv6 packets over the IPv4 cloud.

To determine the IPv4 endpoint of the tunnel, the DS-BGP Router applies the relevant ngtrans tunneling mechanism over the IPv6 address of the Egress DS-BGP Router. Thus, as said before, the IPv6 address of the Egress DS-BGP Router advertised in MP-BGP as the BGP Next Hop MUST be compatible with the ngtrans mechanism used.

The SAFI used in MP-BGP MUST be one of the basic values: unicast, multicast or both (1,2 or 3).

6. Crossing multiple IPv4 domains

When the IPv6 islands are separated by multiple IPv4 domains, two cases can be distinguished:

1. The border routers between the IPv4 domains are not DS-BGP routers, i.e they are IPv4-only BGP routers. The DS-BGP routers of the IPv6 islands from the different IPv4 domains will be configured as MP-BGP peers for the exchange of IPv6 reachability. Alternatively, where the total number of such DS-BGP routers is high, IPv6 reachability across domains can be achieved via MP-BGP connection of Route Reflectors in different domains. One direct inter-domain tunnel per pair of such DS-BGP routers will effectively be created. Note that the exchange of IPv6 routes can only start after BGP has created IPv4 connectivity between the domains.

2. The border routers between the IPv4 domains are DS-BGP routers.

Each of these border DS-BGP routers will peer with the DS-BGP routers in its domain and regular IPv6 routing will take place between the two domains. No inter-domain tunnels are used. There is effectively a separate mesh of tunnels across the DS-BGP Routers of each domain.

7. Comparison

7.1. "MP-BGP over IPv4" approach versus "MP-BGP over IPv6" approach

The "MP-BGP over IPv6" approach requires that an ngtrans tunneling mechanism (eg. ISATAP, 6to4, automatic tunneling, ...) be supported and activated on the DS-BGP Router ahead of time and that IPv6 addresses compatible with this tunneling mechanism be allocated to the DS-BGP Routers.

In contrast, the "MP-BGP over IPv4" approach requires that no other ngtrans mechanism be used.

Because it allows direct label imposition of IPv6 packets (i.e. without prepending an IPv4 header), the "MP-BGP over IPv4" approach can result in less overhead if applied in an MPLS backbone. However it must be noted that, in that case, if for some reason, the LSP fails, forwarding of IPv6 packets towards the corresponding Egress DS-BGP Router will be interrupted. Forwarding of IPv6 packets is not interrupted in case of LSP failure with the "MP-BGP over IPv6" approach or with the "MP-BGP over IPv4" approach when an IPv4 header is prepended before label imposition, since forwarding can fall back to IPv4 forwarding.

7.2. "MP-BGP over IPvX" approaches versus other ngtrans mechanisms

[TRANS] specifies a method to create automatic tunnels by using IPv4-compatible IPv6 addresses. This method is restricted to the case in which the destination coincides with the endpoint of the tunnel (host-to-host or router-to-host tunnels). It has the disadvantage that it requires an IPv4 address per host. "MP-BGP over IPv4" and "MP-BGP over IPv6" approaches require only one IPv4 address per island and enables automatic tunnels for the router-to-router case in contrast to the automatic tunneling described in [TRANS] where the tunnel end-point is the final destination.

With "MP-BGP over IPv4" and "MP-BGP over IPv6" approaches , the hosts in the IPv6 island can have native IPv6 addresses. This is different from e.g. 6to4 [6TO4], which requires that special addresses (6to4 addresses) are assigned to the IPv6 hosts.

7.3. "MP-BGP over IPv4" approach versus MPLS/BGP VPNs

"MP-BGP over IPv4" approach can also be viewed as an instantiation of the solution proposed for IPv6 VPNs over an IPv4 backbone [V6VPN] (the IPv6 Internet is considered as one large 'public' VPN) which is:

(i) generalized since it can also operate with other tunneling techniques than MPLS.

(ii) simplified since it omits the VPN specific parts:

- No need for a Route Distinguisher (RD).

- VPN Routing and Forwarding (VRF) tables are not required.

- No need for a Route Target.

- Except when two (or more) levels of label are used, the basic SAFI values (1, 2, 3) suffice.

- Except when two (or more) levels of label are used, there is no need to carry labels in MP-BGP.

8. Security considerations

This proposal can use the security features of BGP and any policy defined in the ISP domain.

Acknowledgement

We like to thank Tri T. Nguyen, who was the first to come up with the idea described in this document, but who unfortunately passed away much too soon.

Normative References

[GRE] Farinacci D., T. Li, S. Hanks, D. Meyer, P. Traina, "Generic Routing Encapsulation (GRE)", RFC2784.

[IPV6] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC2460.

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

[LABEL] Rekhter Y., E. Rosen, "Carrying Label Information in BGP-4",

RFC 3107, May 2001.

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

[TUNNEL]W. Simpson, "IP in IP Tunneling", RFC1853.

[V6ADDR]Deering, S., and R. Hinden, "IP Version 6 Addressing Architecture", draft-ietf-ipngwg-addr-arch-v3-07.txt (work in progress).

Informative References

[6TO4] B. Carpenter, K. Moore, "Connection of IPv6 domains via IPv4 Clouds", RFC3056, February 2001.

[ISATAP]F. Templin, "Intra-Site Automatic Tunnel Addressing Protocol (ISATAP), draft-ietf-ngtrans-isatap-02.txt (work in progress).

[TRANS] R. Gilligan & E. Nordmark, "Transition Mechanisms for IPv6 Hosts and Routers", RFC2893.

[TTYPE] E. Rosen et al, "Use of PE-PE IPsec in RFC2547 VPNs", draft-rosen-ppvpn-ipsec-2547-00.txt (work in progress).

[V6VPN] Nguyen T., Gastaud G., De Clercq J., Ooms D.,"BGP-MPLS VPN extension for IPv6 VPN over an IPv4 infrastructure", draft-ietf-ppvpn-bgp-ipv6-vpn-01.txt> (work in progress).

[VPN] Rosen E., Rekhter Y., Brannon S., Chase C., De Clercq J., Hitchin P., Marshall , Srinivasan V., "BGP/MPLS VPNs", draft-ietf-ppvpn-rfc2547bis-00.txt (work in progress).

Authors' Addresses

Dirk Ooms
Alcatel
Fr. Wellesplein 1, 2018 Antwerp, Belgium
E-mail: dirk.ooms@alcatel.be

Gerard Gastaud
Alcatel
10 rue Latecoere, BP 57, 78141 Velizy Cedex, France
E-mail: gerard.gastaud@alcatel.fr

Jeremy De Clercq
Alcatel
Fr. Wellesplein 1, 2018 Antwerp, Belgium
E-mail: jeremy.de_clercq@alcatel.be

Stuart Prevost
BTexact Technologies
Room 136 Polaris House, Adastral Park,
Martlesham Heath, Ipswich, Suffolk IP5 3RE, England
E-mail: stuart.prevost@bt.com

Francois Le Faucheur
Cisco Systems
Domaine Green Side, 400, Avenue de Roumanille, Batiment T3
06 410 BIOT, SOPHIA ANTIPOLIS, FRANCE
E-mail: flefauch@cisco.com