Network Working Group
Internet-Draft
Updates: 8966 (if approved)
Intended status: Experimental
Expires: 11 October 2021

J. Chroboczek
IRIF, University of Paris
9 April 2021

IPv4 routes with an IPv6 next-hop in the Babel routing protocol
draft-ietf-babel-v4viav6-01

Abstract

This document defines an extension to the Babel routing protocol that allows annoncing routes to an IPv4 prefix with an IPv6 next-hop, which makes it possible for IPv4 traffic to flow through interfaces that have not been assigned an IPv4 address.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Traditionally, a routing table maps a network prefix of a given address family to a next-hop address in the same address family. The sole purpose of this next-hop address is to serve as an input to a protocol that will map it to a link-layer address, Neighbour Discovery (ND) [RFC4861] in the case of IPv6, Address Resolution (ARP) [RFC0826] in the case of IPv4. Therefore, there is no reason why the address family of the next hop address should match that of the prefix being announced: an IPv6 next-hop yields a link-layer address that is suitable for forwarding both IPv6 or IPv4 traffic.

We call a route towards an IPv4 prefix that uses an IPv6 next hop a "v4-via-v6" route. Since an IPv6 next-hop can use a link-local address that is autonomously configured, the use of v4-via-v6 routes enables a mode of operation where the network core has no statically assigned IP addresses of either family, thus significantly reducing the amount of manual configuration.

This document describes an extension that allows the Babel routing protocol [RFC8966] to announce routes towards IPv6 prefixes with IPv4 next hops. The extension is inspired by a previously defined extension to the BGP protocol [RFC5549].

1.1. Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Protocol operation

The Babel protocol fully supports double-stack operation: all data that represent a neighbour address or a network prefix are tagged by an Address Encoding (AE), a small integer that identifies the address family (IPv4 or IPv6) of the address of prefix, and describes how it is encoded. This extension defines a new AE, called v4-via-v6, which has the same format as the existing AE for IPv4 addresses. This new AE is only allowed in TLVs that carry network prefixes: TLVs that carry a neighbour address use the normal encodings for IPv6 addresses.

2.1. Announcing v4-via-v6 routes

A Babel node that needs to announce an IPv4 route over an interface that has no assigned IPv4 address MAY make a v4-via-v6 announcement. In order to do so, it first establishes an IPv6 next-hop address in the usual manner (either by sending the Babel packet over IPv6, or by including a Next Hop TLV containing an IPv6 address); it then sends an Update with AE equal to TBD containing the IPv4 prefix being announced.

If the outgoing interface has been assigned an IPv4 address, then, in the interest of maximising compatibility with existing routers, the sender SHOULD prefer an ordinary IPv4 announcement; even in that case, however, it MAY use a v4-via-v6 announcement. A node SHOULD NOT send both ordinary IPv4 and v4-via-v6 annoucements for the same prefix over a single interface (if the update is sent to a multicast address) or to a single neighbour (if sent to a unicast address), since doing that doubles the amount of routing traffic while providing no benefit.

2.2. Receiving v4-via-v6 routes

Upon reception of an Update TLV with a v4-via-v6 AE and finite metric, a Babel node computes the IPv6 next-hop, as described in Section 4.6.9 of [RFC8966]. If no IPv6 next-hop exists, then the Update MUST be silently ignored. If an IPv6 next-hop exists, then the node MAY acquire the route being announced, as described in Section 3.5.3 of [RFC8966]; the parameters of the route are as follows:

* the prefix, plen, router-id, seqno, metric MUST be computed as for an IPv4 route, as described in Section 4.6.9 of [RFC8966];

* the next-hop MUST be computed as for an IPv6 route, as described in Section 4.6.9 of [RFC8966]: it is taken from the last preceding Next-Hop TLV with an AE field equal to 2 or 3; if no such entry exists, and if the Update TLV has been sent in a Babel packet carried over IPv6, then the next-hop is the network-layer source address of the packet.

An Update TLV with a v4-via-v6 AE and metric equal to infinity is a retraction: it announces that a previously available route is being retracted. In that case, no next-hop is necessary, and the retraction is treated as described in Section 4.6.9 of [RFC8966].

As usual, a node MAY ignore the update, e.g., due to filtering (Appendix C of [RFC8966]). If a node cannot install v4-via-v6 routes, eg., due to hardware or software limitations, then routes to an IPv4 prefix with an IPv6 next-hop MUST NOT be selected, as described in Section 3.5.3 of [RFC8966].

2.3. Prefix and seqno requests

Prefix and seqno requests are used to request an update for a given prefix. Since they are not related to a specific Next-Hop, there is no semantic difference between IPv4 and v4-via-v6 requests. Therefore, a node SHOULD NOT send requests of either kind with the AE field being set to TBD (v4-via-v6); instead, it SHOULD request IPv4 updates using requests with the AE field being set to 1 (IPv4).

When receiving requests, AEs 1 (IPv4) and TBD (v4-via-v6) MUST be treated in the same manner: the receiver processes the request as described in Section 3.8 of [RFC8966]. If an Update is sent, then it MAY be sent with AE 1 or TBD, as described in Section 2.1 above, irrespective of which AE was used in the request.

When receiving a request with AE 0 (wildcard), the receiver SHOULD send a full route dump, as described in Section 3.8.1.1 of [RFC8966]. Any IPv4 routes contained in the route dump MAY use either AE 1 or AE TBD, as described in Section 2.1 above.

2.4. Other TLVs

The only other TLVs defined by [RFC8966] that carry an AE field are Next-Hop and TLV. Next-Hop and IHU TLVs MUST NOT carry the AE TBD (v4-via-v6).

3. ICMPv4 and PMTU discovery

The Internet Control Message Protocol (ICMPv4, or simply ICMP) [RFC792] is a protocol related to IPv4 that carries diagnostic and debugging information. ICMPv4 packets may be originated by end hosts (e.g., the "destination unreachable, port unreachable" ICMPv4 packet), but they may also be originated by intermediate routers (e.g., most other kinds of "destination unreachable" packets).

Path MTU Discovery (PMTUd) [RFC1191] is an algorithm executed by end hosts to discover the maximum packet size that a route is able to carry. While there exist variants of PMTUd that are purely end-to-end [RFC4821], the variant most commonly deployed in the Internet has a hard dependency on ICMPv4 packets originated by intermediate routers: if intermediate routers are unable to send ICMPv4 packets, PMTUd may lead to persistent blackholing of IPv4 traffic.

For that reason, every Babel router that is able to forward IPv4 traffic MUST be able originate ICMPv4 traffic. Since the extension described in this document enables routers to forward IPv4 traffic even when they have not been assigned an IPv4 address, a router implementing this extension MUST be able to originate ICMPv4 packets even when it has not been assigned an IPv4 address.

There are various ways to meet this requirement, and choosing between them is left to the implementation. For example, if a router has an interface that has been assigned an IPv4 address, or if it has an IPv4 address that has been assigned to the router itself (to the "loopback interface"), then that IPv4 address may be "borrowed" to serve as the source of originated ICMPv4 packets. If no IPv4 address is available, a router may use a dummy IPv4 address as the source of outgoing ICMPv4 packets, for example an address taken from a private address range [RFC1918] that is known to not be used in the local routing domain. Note however that using the same address on multiple routers may hamper debugging and fault isolation.

4. Backwards compatibility

This protocol extension adds no new TLVs or sub-TLVs.

This protocol extension uses a new AE. As discussed in Appendix D of [RFC8966] and specified in the same document, implementations that do not understand the present extension will silently ignore the various TLVs that use this new AE. As a result, incompatible versions will ignore v4-via-v6 routes. They will also ignore requests with AE TBD, which, as stated in Section 2.3, are NOT RECOMMENDED.

Using a new AE introduces a new compression state, used to parse the network prefixes. As this compression state is separate from other AEs' states, it will not interfere with the compression state of unextended nodes.

This extension reuses the next-hop state from AEs 2 and 3 (IPv6), but makes no changes to the way it is updated, and therefore causes no compatibility issues.

As mentioned in Section 2.1, ordinary IPv4 announcements are preferred to v4-via-v6 announcements when the outgoing interface has an assigned IPv4 address; doing otherwise would prevent routers that do not implement this extension from learning the route being announced.

5. Protocol encoding

This extension defines the v4-via-v6 AE, whose value is TBD. This AE is solely used to tag network prefixes, and MUST NOT be used to tag peers' addresses, eg. in Next-Hop or IHU TLVs.

This extension defines no new TLVs or sub-TLVs.

5.1. Prefix encoding

Network prefixes tagged with AE TBD MUST be encoded and decoded as prefixes tagged with AE 1 (IPv4), as described in Section 4.3.1 of [RFC8966].

A new compression state for AE TBD (v4-via-v6) distinct from that of AE 1 (IPv4) is introduced, and MUST be used for address compression of prefixes tagged with AE TBD, as described in Section 4.6.9 of [RFC8966]

5.2. Changes for existing TLVs

The following TLVs MAY be tagged with AE TBD:

* Update (Type = 8)

* Route Request (Type = 9)

* Seqno Request (Type = 10)

As AE TBD is suitable only to tag network prefixes, IHU (Type = 5) and Next-Hop (Type = 7) TLVs MUST NOT be tagged with AE TBD. Such (incorrect) TLVs MUST be silently ignored upon reception.

5.2.1. Update

An Update (Type = 8) TLV with AE = TBD is constructed as described in Section 4.6.9 of [RFC8966] for AE 1 (IPv4), with the following specificities:

* Prefix. The Prefix field is constructed according to the Section 5.1 above.

* Next hop. The next hop is determined as described in Section 2.2 above.

5.2.2. Other valid TLVs tagged with AE = TBD

Any other valid TLV tagged with AE = TBD MUST be constructed and decoded as described in Section 4.6 of [RFC8966]. Network prefixes within MUST be constructed and decoded as described in Section 5.1 above.

6. IANA Considerations

IANA is requested to allocate a value (4 suggested) in the "Babel Address Encodings" registry as follows:

AE	Name	Reference
TBD	v4-via-v6	(this document)

Table 1

7. Security Considerations

The extension defined in this document does not fundamentally change the security properties of the Babel protocol. However, by allowing IPv4 routes to be propagated across routers that have not been assigned IPv4 addresses, it might invalidate the assumptions made by some network administatoris, which could conceivably lead to security issues.

For example, if an island of IPv4-only hosts is separated from the IPv4 Internet by an area of routers that have not been assigned IPv4 addresses, a network administrator might reasonably assume that the IPv4-only hosts are unreachable from the IPv4 Internet. This assumption is broken if the intermediary routers implement the extension described in this document, which might expose the IPv4-only hosts to traffic from the IPv4 Internet. If this is undesirable, the flow of IPv4 traffic must be restricted by the use of suitable filtering rules (Appendix C of [RFC8966]) together with matching packet filters in the data plane.

8. Acknowledgments

This protocol extension was originally designed, described and implemented in collaboration with Theophile Bastian. The author is also indebted to Margaret Cullen, who pointed out the issues with ICMP and helped coin the expression "V4-via-V6".

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/rfc/rfc2119>.

[RFC792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, DOI 10.17487/RFC0792, September 1981, <https://www.rfc-editor.org/info/rfc792>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

[RFC8966] Chroboczek, J. and D. Schinazi, "The Babel Routing Protocol", RFC 8966, DOI 10.17487/RFC8966, January 2021, <https://www.rfc-editor.org/info/rfc8966>.

9.2. Informative References

[RFC0826] Plummer, D., "An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware", STD 37, RFC 826, DOI 10.17487/RFC0826, November 1982, <https://www.rfc-editor.org/rfc/rfc826>.

[RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <https://www.rfc-editor.org/info/rfc1191>.

[RFC1918] Rekhter, Y., Moskowitz, B., Karrenberg, D., de Groot, G. J., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, DOI 10.17487/RFC1918, February 1996, <https://www.rfc-editor.org/info/rfc1918>.

[RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, DOI 10.17487/RFC4821, March 2007, <https://www.rfc-editor.org/info/rfc4821>.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <https://www.rfc-editor.org/rfc/rfc4861>.

[RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <https://www.rfc-editor.org/rfc/rfc5549>.

Author's Address

Juliusz Chroboczek
IRIF, University of Paris
Case 7014
75205 Paris Cedex 13
France

Email: jch@irif.fr