Workgroup: tsvwg
Internet-Draft:
draft-duke-tsvwg-ecn-aggregating-tunnels-00
Published: 15 August 2022
Intended Status: Best Current Practice
Expires: 16 February 2023
Authors: M. Duke
Google LLC

ECN Over Aggregating Tunnels

Abstract

Explicit Congestion Notification (ECN) provides two bits in the IP header for routers to signal congestion to endpoints without resorting to packet loss. RFC6040 provided guidance for how IP-in-IP tunnels should transfer (ECN) markings between inner and outer IP headers. However, that document implicitly assumes that no more than one inner packet is present in an outer packet. As numerous tunneling technologies have since emerged that break this assumption, further guidance is needed.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Explicit Congestion Notification (ECN) [RFC3168] provides a means for routers to signal congestion to endpoints without dropping packets. This can achieve the goals of internet congestion control while not introducing a degraded quality of experience and/or delay due to packet retransmission. The internet community is also now experimenting with using unused ECN codepoints to provide extremely low-latency services [I-D.ietf-tsvwg-l4s-arch].

To take full advantage of ECN, [RFC6040] provides rules for encapsulating and decapsulating nodes for IP-in-IP tunnels to propagate ECN markings from inner headers to outer headers on tunnel ingress, and from outer to inner headers on tunnel egress.

RFC6040 implicitly assumes that no more than one inner IP header is present in a tunnel packet. (RFC3168 is clear that an IP packet reassembled from fragments takes the highest congestion indication from its fragments). Nevertheless, there are several IP-in-IP tunnel architectures that allow multiple inner IP datagrams in a single tunnel packet. For examples, see [I-D.ietf-ipsecme-rfc8229bis], [I-D.ietf-ipsecme-iptfs], and [I-D.ietf-masque-connect-ip]. Existing specifications do not provide recommendations when IP packets with different ECN marks are encapsulated in the same tunnel IP packet.

2. Conventions and Definitions

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.

3. Default Tunnel Ingress Behavior

An encapsulator **SHOULD NOT** aggregate packets marked Not-ECT, ECT(0), and ECT(1) in the same tunnel packet unless doing so prevents unacceptable delay, packet reordering, or other degradation in metrics.

The encapsulator checks the following conditions in order, until it finds an applicable marking instruction. In two cases, these rules offer an optional behavior because they might cause RFC6040-compliant egress to throw an alarm and/or log an error. If the ingress believes these conditions apply to the egress and the alarms or errors would produce an unacceptable operational burden, it uses the optional behavior.

1. If all inner packets have the same marking, the encapsulator follows the rules in Section 4.1 of [RFC6040].

2. If the tunnel packet contains both ECT(0) and ECT(1), mark the packet Not- ECT.

3. If any inner header is marked ECT(0), mark the outer header ECT(0). A tunnel ingress **MAY** mark it Not-ECT if there is also a Not-ECT header present, in order to avoid alarms or errors at the tunnel egress.

4. If any inner packet is marked Not-ECT, mark the outer header Not-ECT.

5. If no above rules apply, the inner headers are all marked ECT(1) or CE. Mark the outer header ECT(1). Encapsulators **MAY** instead mark the tunnel packet Not- ECT to avoid generating alerts or alarms at the egress.

The following table summarizes the possible outcomes for all 16 combinations of inner header packet markings:

Present in Tunnel Packet ?				Outer Header Marking	Applicable Rule Number(s)
Not-ECT	ECT(0)	ECT(1)	CE		
Y	N	N	any	Not-ECT	1,4
N	Y	N	any	ECT(0)	1,3
N	N	Y	N	ECT(1)	1
N	N	N	Y	CE	1
any	Y	Y	any	Not-ECT	2
Y	Y	N	any	ECT(0)*	3
Y	N	Y	any	Not-ECT	4
N	N	Y	Y	ECT(1)*	5
N	N	N	N	N/A	N/A

Table 1. Ingress Markings

* Ingress may mark outer packet Not-ECT to avoid errors and alarms at tunnel egress.

Encapsulators **MUST**, in the rules above, consider the marking of packet fragments where the inner IP header is not actually present in the tunnel packet being marked.

4. Default Tunnel Egress Behavior

Decapsulators follow the guidance in Section 4.2 of [RFC6040], except that they **SHOULD NOT** raise an alarm or log an error under the following conditions:

*the outer header is ECT(0) and an inner header is Not-ECT;

*the outer header is ECT(1) and an inner header is CE; or

*the outer header is CE and in the inner header is CE.

These are expected behaviors in this specification.

When reassembling an inner packet from fragments scattered over multiple outer packets, decapsulators apply the strictest outcome applied to any of the packets. If any outer packet is dropped, the inner packet is dropped. Otherwise, if any outer packet is marked

CE, the inner packet is dropped (if marked Not-ECT) or marked CE (if marked anything else). Other outer packet markings do not change the marking of the inner parking.

5. Rationale

The above rules minimize the changes necessary to tunnel egress. Marking the outer header Not-ECT always allows the egress to preserve the inner header markings, although it may result in a packet drop where a CE marking would have been a better outcome.

Unless an outer header containing ECT(0) and ECT(1) inner headers is marked Not-ECT, it risks being marked CE. As ECT(0) and ECT(1) flows react differently to CE markings, one will respond inappropriately. However, they will both respond correctly to a packet drop due to the Not-ECT setting.

A Not-ECT inner header cannot be in an ECT(1) outer header because the outer header will be marked CE more aggressively than an ECT(0) header, and does not correspond to a packet loss for Not-ECT. Thus, the egress's drop of the inner Not-ECT packet on CE is inappropriate.

CE inner header are always preserved on egress, so they can coexist with any outer header codepoint.

6. Security Considerations

The security considerations in [RFC6040] apply.

An attacker might attempt to degrade service by injecting packets into the ingress that force the outer header to be Not-ECT. They would inject ECT(1) if the legitimate traffic was mostly ECT(0), and Not-ECT otherwise. This is one reason tunnel encapsulators are encouraged to separate Not-ECT, ECT(0), and ECT(1) traffic.

7. IANA Considerations

This document has no IANA actions.

8. References

8.1. 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/info/rfc2119>.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <https://www.rfc-editor.org/info/rfc3168>.

[RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, DOI 10.17487/RFC6040, November 2010, <https://www.rfc-editor.org/info/rfc6040>.

[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/info/rfc8174>.

8.2. Informative References

[I-D.ietf-ipsecme-iptfs] Hopps, C., "IP-TFS: Aggregation and Fragmentation Mode for ESP and its Use for IP Traffic Flow Security", Work in Progress, Internet-Draft, draft-ietf-ipsecme-iptfs-13, 5 June 2022, <https://www.ietf.org/archive/id/draft-ietf-ipsecme-iptfs-13.txt>.

[I-D.ietf-ipsecme-rfc8229bis] Pauly, T. and V. Smyslov, "TCP Encapsulation of IKE and IPsec Packets", Work in Progress, Internet-Draft, draft-ietf-ipsecme-rfc8229bis-07, 3 June 2022, <https://www.ietf.org/archive/id/draft-ietf-ipsecme-rfc8229bis-07.txt>.

[I-D.ietf-masque-connect-ip] Pauly, T., Schinazi, D., Chernyakhovsky, A., Kuehlewind, M., and M. Westerlund, "IP Proxying Support for HTTP", Work in Progress, Internet-Draft, draft-ietf-masque-connect-ip-02, 11 July 2022, <https://www.ietf.org/archive/id/draft-ietf-masque-connect-ip-02.txt>.

[I-D.ietf-tsvwg-l4s-arch] Briscoe, B., Schepper, K. D., Bagnulo, M., and G. White, "Low Latency, Low Loss, Scalable Throughput (L4S) Internet Service: Architecture", Work in Progress, Internet-Draft, draft-ietf-tsvwg-l4s-arch-19, 27 July 2022, <https://www.ietf.org/archive/id/draft-ietf-tsvwg-l4s-arch-19.txt>.

Acknowledgments

TODO acknowledge.

Author's Address

Martin Duke
Google LLC

Email: martin.h.duke@gmail.com