Internet Engineering Task Force
INTERNET DRAFT
draft-floyd-ecn-tunnels-01.txt

S. Floyd
K. K. Ramakrishnan
D. Black

October 2000

# ECN Interactions with IP Tunnels

## Status of this Memo



## Abstract

The encapsulation of IP packet headers in tunnels is used in many places, including IPsec and IP in IP [RFC2003]. Explicit Congestion Notification (ECN) is an experimental addition to the IP architecture that uses the ECN field in the IP header to provide an indication of the onset of congestion to applications. ECN provides this congestion indication to enable end-node adaptation to network conditions without the use of dropped packets [RFC 2481]. Currently, the ECN specification does not accommodate the constraints imposed by some of these pre-existing specifications for tunnels. This document considers issues related to interactions between ECN and IP tunnels, and proposes two alternative solutions.

A different set of issues are raised, relative to ECN, when IP packets are encapsulated in tunnels with non-IP packet headers. This occurs with MPLS [MPLS], GRE [GRE], L2TP [L2TP], and PPTP [PPTP]. For these protocols, there is no conflict with ECN; it is just that ECN cannot be used within the tunnel unless an ECN codepoint can be specified for the header of the encapsulating protocol. [RFD99] presents a proposal for incorporating ECN into MPLS, and proposals for incorporating ECN into GRE, L2TP, or PPTP will be considered as the need arises.

## 1. Introduction.

Some IP tunnel modes are based on adding a new "outer" IP header that encapsulates the original, or "inner" IP header and its associated packet. In many cases, the new "outer" IP header may be added and removed at intermediate points along a connection, enabling the network to establish a tunnel without requiring endpoint participation. We denote tunnels that specify that the outer header be discarded at tunnel egress as ``simple tunnels''.

Explicit Congestion Notification (ECN) is an experimental addition to the IP architecture that provides congestion indication to end-nodes to enable them to adapt to network conditions without requiring the packet to be dropped [RFC 2481]. An ECN-capable router uses the ECN mechanism to signal congestion to connection endpoints by setting a bit in the IP header. These endpoints then react, in terms of congestion control, as if a packet had been dropped (e.g., TCP halves its congestion window). This ability to avoid dropping packets in response to congestion is supported by the use of active queue management mechanisms (e.g., RED) in routers; such mechanisms begin to mark or drop packets as a consequence of congestion before the congested router queue is completely full. ECN is defined to be used as an optimization -- routers are not required to support ECN, and even an ECN-capable router may drop packets from ECN-capable connections when necessary. The advantage to a router of not dropping such packets is that ECN can provide a more timely indication of congestion to the end nodes than indications based on packet drops being detected by duplicate ACKs or timeout. As a result, the queues at the router are better managed.

Currently, the ECN specification does not interact appropriately with simple IP tunnels. Current use of ECN over simple IP tunnels results in routers attempting to use the outer IP header to signal congestion to endpoints, but those congestion warnings never arrive because the outer header is discarded at the tunnel egress point. It is desirable for the tunnel egress point to recognize the use of ECN on the inner IP header. This problem was encountered with ECN and IPsec in tunnel mode, and RFC 2481 recommends that ECN not be used with the older simple IPsec tunnels in order to avoid this behavior and its consequences.

This document considers issues related to interactions between ECN and IP tunnels and proposes solutions. From a security point of view, the use of ECN in the outer header of an IP tunnel might raise security concerns because an adversary could tamper with the ECN information that propagates beyond the tunnel endpoint. Based on an analysis of these concerns and the resultant risks [IPsecECN], our overall approach is to make support for ECN an option for IP tunnels, so that an IP tunnel can be specified or configured either to use ECN or not to use ECN in the outer header of the tunnel. Thus, in environments or tunneling protocols where the risks of using ECN are judged to outweigh its benefits, the tunnel can simply not use ECN in the outer header. Then the only indication of congestion experienced at routers within the tunnel would be through packet loss.

The result is that there are two viable options for the behavior of ECN-capable connections over an IP tunnel, especially IPSec tunnels:

- A limited-functionality option in which ECN is preserved in the inner header, but disabled in the outer header. The only mechanism available for signaling congestion occurring within the tunnel in this case is dropped packets.
- A full functionality option that supports ECN in both the inner and outer headers, and propagates congestion warnings from nodes within the tunnel to endpoints.

Support for these options requires varying amounts of changes to IP header processing at tunnel ingress and egress. A small subset of these changes sufficient to support only the limited-functionality option would be sufficient to eliminate any incompatibility between ECN and IP tunnels.

One goal of this document is to give guidance about the tradeoffs between the limited-functionality and full-functionality options. A full discussion of the potential effects of an adversary's modifications of the CE and ECT bits is given in [IPsecECN]. This document draws heavily on [IPsecECN], both in terms of the approach and the text itself.

## 2. Architecture.

ECN uses two bits in the IP header, the ECT bit (ECN-Capable Transport) and the CE bit (Congestion Experienced), for signaling between routers and connection endpoints, and uses two flags in the TCP header, the ECN-Echo bit (to Echo the ECN bit in IP header) and the CWR bit (Congestion Window Reduced) for TCP-endpoint to TCP-endpoint signaling. For a TCP connection, a typical sequence of events in an ECN-based reaction to congestion is as follows:

- The ECT bit is set in packets transmitted by the sender to indicate that ECN is supported on this TCP connection.
- An ECN-capable router detects impending congestion and detects that the ECT bit is set in the packet it is about to drop. Instead of dropping the packet, the router sets the CE bit and forwards the packet.
- The receiver receives the packet with CE set, and sets the ECN-Echo flag in its next TCP ACK sent to the sender.
- The sender receives the TCP ACK with ECN-Echo set, and reacts to the congestion as if a packet had been dropped.
- The sender sets the CWR flag in the TCP header of the next packet sent to the receiver to acknowledge its receipt of and reaction to the ECN-Echo flag.

Further details on ECN functionality, including negotiation of ECN-capability as part of TCP connection setup as well as the responsibilities and requirements of ECN-capable routers and transports, can be found in [RFC2481].

ECN interacts with IP tunnels because the two ECN bits are in the DS field octet in the IP header [RFC2474] (also referred to as the IPv4 TOS octet or IPv6 Traffic Class octet).  The DS field octet is generally copied or mapped from the inner IP header to the outer IP header at IP tunnel ingress, and in simple IP tunnels the outer header's copy of this field is discarded at IP tunnel egress.  If an ECN-capable router were to set the CE (Congestion Experienced) bit within a packet in a simple IP tunnel, this indication would be discarded at tunnel egress, losing the indication of congestion.  As a consequence of this behavior, ECN usage within a simple IP tunnels (with no changes at the ingress and egress) is not recommended.

The limited-functionality option for ECN encapsulation in IP tunnels is for the ECT bit in the outside (encapsulating) header to be off (i.e., set to 0), regardless of the value of the ECT bit in the inside (encapsulated) header.  With this option, the ECN field in the inner header is not altered upon de-capsulation.  The disadvantage of this approach is that the flow does not have ECN support for that part of the path that is using IP tunneling, even if the encapsulated packet is ECN-Capable.  That is, if the encapsulated packet arrives at a congested router that is ECN-capable, and the router can decide to drop or mark the packet as an indication of congestion to the end nodes, the router will not be permitted to set the CE bit in the packet header, but instead will have to drop the packet.

The IP full-functionality option for ECN encapsulation follows the description in Section 10.1 of RFC 2481 of tunneling with ECN.  This option is to copy the ECT bit of the inside header to the outside header on encapsulation, and to OR the CE bit from the outer header with the CE bit of the inside header on decapsulation. With the full-functionality option, a flow can take advantage of ECN for those parts of the path that might use IP tunneling. The disadvantage of the full-functionality option from a security perspective is that the IP tunnel cannot protect the flow from certain modifications to the ECN bits in the IP header within the tunnel. The potential dangers from modifications to the ECN bits in the IP header are described in detail in [IPsecECN].

This document proposes either the limited-functionality or full-functionality option for IP tunnels in order to enable ECN experimentation over IP tunnels, and avoid losing congestion indications in the case that an ECN-capable router or routers are traversed by an IP tunnel carrying ECN-capable connections. In summary, two changes are proposed to IP tunnel functionality:

(1) Modify the handling of the DS field octet at IP tunnel endpoints by implementing either the limited-functionality or the full-functionality option.
(2) Optionally, enable the endpoints of an IP tunnel to negotiate the choice between the limited-functionality and the full-functionality option for ECN in the tunnel.

The minimum required to make ECN usable with IP tunnels is the limited-functionality option, which prevents ECN from being enabled in the outer header of an IPsec tunnel. Full support for ECN requires the use of the full-functionality option. Optional mechanisms to negotiate a choice between the tunnel endpoints of either the limited-functionality or full-functionality option are not discussed in this document. We assume that there is a pre-existing agreement between the tunnel endpoints about whether to support the limited-functionality or the full-functionality ECN option.

The two ECN bits in the IP header, ECT and CE, occupy bits 6 and 7 of the DS Field octet [RFC2481]. For full ECN support the encapsulation and decapsulation processing for the DS field octet involves the following: At tunnel ingress, the full-functionality option copies the value of ECT (bit 6) in the inner header to the outer header. CE (bit 7) is set to 0 in the outer header. At tunnel egress, the full-functionality option sets CE to 1 in the inner header if the value of ECT (bit 6) in the inner header is 1, and the value of CE (bit 7) in the outer header is 1. Otherwise, no change is made to this field of the inner header.

For the limited-functionality option, at tunnel ingress bits 6 and 7 (ECT and CE) of the DS field in the outer header are set to zero, and at tunnel egress no change is made to the DS field in the inner header.

In addition, it is RECOMMENDED that packets with ECN and CE both set to 1 in the outer header be dropped if they arrive on an tunnel egress for a tunnel that uses the limited-functionality option, or for a tunnel that uses the full-functionality option but for which the ECT bit in the inner header is set to zero.  This is motivated by backwards compatibility and to ensure that no unauthorized modifications of the ECN field takes place and is discussed further in Section 6.

## 4.  Possible Changes to the ECN Field

This section considers the issues when a router is operating, possibly maliciously, to modify either of the bits in the ECN field. In this section we represent the ECN field in the IP header by the tuple (ECT bit, CE bit).  The ECT bit, when set to 1, indicates an ECN-Capable Transport.  The CE bit, when set to 1, indicates that Congestion was Experienced in the path.

By tampering with the bits in the ECN field, an adversary (or a broken router) could do one or more of the following: falsely report congestion, disable ECN-Capability for an individual packet, erase the ECN congestion indication, or falsely indicate ECN-Capability. [IPsecECN] systematically examines the various cases by which the ECN field could be modified.  The important criterion considered in determining the consequences of such modifications is whether it is likely to lead to poorer behavior in any dimension (throughput, delay, fairness or functionality) than if a router were to drop a packet.

The first two possible changes, falsely report congestion or disabling ECN-Capability for an individual packet, are no worse than if the router were to simply drop the packet.  However, as discussed in Section 5 below, a router that erases the ECN congestion indication or falsely indicates ECN-Capability could potentially do more damage to the flow that if it has simply dropped the packet.

## 5.  Implications of Subverting End-to-End Congestion Control

This section considers the potential repercussions of subverting end-to-end congestion control by either falsely indicating ECN-Capability, or by erasing the congestion indication in ECN (the CE-bit).  Subverting end-to-end congestion control by either of these two methods can have consequences both for the application and for the network.

The first method to subvert end-to-end congestion control, falsely indicating ECN-Capability, effectively subverts end-to-end congestion control only if the packet would later encounter congestion that results in the setting of the CE bit. In this case, the transport protocol (which itself was not ECN capable) does not react appropriately to the indication of congestion from these downstream congested routers. It would have been better for these downstream congested routers to drop the packet instead.

The second method to subvert end-to-end congestion control, 'erasing' the (set) CE bit in a packet, effectively subverts end-to-end congestion control only when the CE bit in the packet was set earlier by a congested router. In this case, the transport protocol does not receive the indication of congestion from the upstream congested routers.

Either of these two methods of subverting end-to-end congestion control can potentially introduce more damage to the network (and possibly to the flow itself) than if the adversary had simply dropped packets from that flow. However, as we discuss in the subsequent sections, this potential damage is limited. This is also discussed extensively in [IPsecECN].

## 6. Changes to the ECN Field within an IP Tunnel.

The presence of a copy of the ECN field in the inner header of an IP tunnel mode packet provides an opportunity for detection of unauthorized modifications to the ECT bit in the outer header. Comparison of the ECT bits in the inner and outer headers falls into two categories for implementations that conform to this document:

(a) If the IP tunnel uses the full-functionality option, then the values of the ECT bits in the inner and outer headers should be identical.

(b) If the tunnel uses the limited-functionality option, then the ECT bit in the outer header should be 0.

Receipt of a packet not satisfying the appropriate condition could be a cause of concern.

Consider the case of an IP tunnel where the tunnel ingress point has not been updated to this document's requirements, while the tunnel egress point has been updated to support ECN. In this case, the IP tunnel is not explicitly configured to support the full-functionality ECN option. However, the tunnel ingress point is behaving identically to a tunnel ingress point that supports the full-functionality option. If packets from an ECN-capable connection use this tunnel, ECT will be set to 1 in the outer header at the tunnel ingress point. Congestion within the tunnel may then result in ECN-capable routers setting CE in the outer header. Because the tunnel has not been explicitly configured to support the full-functionality option, the tunnel egress point expects the ECT bit in the outer header to be 0.

When an ECN-capable tunnel egress point receives a packet with the ECT bit in the outer header set to 1, in a tunnel that has not been configured to support the full-functionality option, that packet should be processed, according to whether CE bit was set, as follows. It is RECOMMENDED that such packets, with the ECT bit set to 1 on a tunnel that has not been configured to support the full-functionality option, be dropped at the egress point if CE is set to 1 in the outer header but 0 in the inner header, and forwarded otherwise.

An IP tunnel cannot provide protection against erasure of congestion indications based on resetting the value of the CE bit in packets for which ECT is set in the outer header.  The erasure of congestion indications may impact the network and other flows in ways that would not be possible in the absence of ECN.  It is important to note that erasure of congestion indications can only be performed to congestion indications placed by nodes within the tunnel; the copy of the CE bit in the inner header preserves congestion notifications from nodes upstream of the tunnel ingress.  If erasure of congestion notifications is judged to be a security risk that exceeds the congestion management benefits of ECN, then tunnels could be specified or configured to use the limited-functionality option.

## 7.  Issues Raised by Monitoring and Policing Devices

One possibility is that monitoring and policing devices (or more informally, ``penalty boxes'') will be installed in the network to monitor whether best-effort flows are appropriately responding to congestion, and to preferentially drop packets from flows determined not to be using adequate end-to-end congestion control procedures. This is discussed in more detail in [IPsecECN]

For an ECN-capable flow, an `ideal' penalty box at a router would be a device that, when it detected that a flow was not responding to ECN indications, would switch to dropping, instead of marking, those packets of a flow that would otherwise have been chosen to carry indications of congestion.  In this way, these congestion indications could not be `erased' later in the network, and at the same time there would be no change in the router's treatment of packets of other flows.  If a router determines that a flow is still not responding to congestion indications when the congestion indications consist of packet drops, then the router could take whatever further action it deems appropriate for that flow.

We recommend that any ``penalty box'' that detects a flow or an aggregate of flows that is not responding to end-to-end congestion control first change from marking to dropping packets from that flow, before taking any additional action to restrict the bandwidth available to that flow.  Thus, initially, the router may drop packets in which the router would otherwise would have set the CE bit. This could include dropping those arriving packets for that flow that are ECN-Capable and that already have the CE bit set. In this way, any congestion indications seen by that router for that flow will be guaranteed to also be seen by the end nodes, even in the presence of malicious or broken routers elsewhere in the path. If we assume that the first action taken at any ``penalty box'' for an ECN-capable flow will be to drop packets instead of marking them, then there is no way that an adversary that subverts ECN-based end-to-end congestion control can cause a flow to be characterized as being non-cooperative and placed into a more severe action within the ``penalty box''.

If there were serious operational problems with routers inappropriately erasing the CE bit in packet headers, one potential fix would be to include a one-bit ECN nonce in packet headers, and for routers to erase the nonce when they set the CE bit [SCWA99]. Routers would be unable to consistently reconstruct the nonce when they erased the CE bit, and thus the repeated erasure of the CE bit would be detected by the end-nodes. (This could in fact be done without adding any extra bits for ECN in the IP header, by using the ECN codepoints (ECT=1, CE=0) and (ECT=0, CE=1) as the two values for the nonce, and by defining the codepoint (ECT=0, CE=1) to mean exactly the same as the codepoint (ECT=1, CE=0).) However, at this point the potential danger does not seem of sufficient concern to warrant this additional complication of adding an ECN nonce to protect against the erasure of the CE bit.

## 7.1. Complications Introduced by Split Paths

If a router or other network element has access to all of the packets of a flow, then that router could do no more damage to a flow by altering the ECN field than it could by simply dropping all of the packets from that flow. However, in some cases, a malicious or broken router might have access to only a subset of the packets from a flow. The question is as follows: can this router, by altering the ECN field in this subset of the packets, do more damage to that flow than if it has simply dropped that set of the packets?

This is also discussed in detail in [IPsecECN], which concludes as follows: It is true that the adversary that has access only to the A packets might, by subverting ECN-based congestion control, be able to deny the benefits of ECN to the other packets in the A&B aggregate. While this is undesireable, this is not a sufficient concern to result in disabling ECN within an IP tunnel.

## 8. Conclusions.

When ECN (Explicit Congestion Notification [RFC2481]) is used, it is desirable that congestion indications generated within an IP tunnel not be lost at the tunnel egress.  We propose a minor modification to the IP protocol's handling of the ECN field during encapsulation and de-capsulation to allow flows that will undergo IP tunneling to use ECN.

Two options were proposed:
1) A preferred alternative, which is the full-functionality option as described in RFC 2481. This copies the ECT bit of the inner header to the encapsulating header. At decapsulation, if the ECT bit is set in the inner header, the CE bit on the outer header is ORed with the CE bit of the inner header to update the CE bit of the packet.
2) A limited-functionality option that does not use ECN inside the IP tunnel, by turning the ECT bit in the outer header off, and not altering the inner header at the time of decapsulation.

In [IPsecECN] we examined the consequence of modifications of the ECN field within the tunnel, analyzing all the opportunities for an adversary to change the ECN field.  In many cases, the change to the ECN field is no worse than dropping a packet. However, we noted that some changes have the more serious consequence of subverting end-to-end congestion control.  However, we point out that even then the potential damage is limited, and is similar to the threat posed by an end-system intentionally failing to cooperate with end-to-end congestion control.  We therefore believe that with these changes it is reasonable to use ECN with IP tunnels, as described in RFC 2481.

## 9. Acknowledgements

We thank Tabassum Bint Haque from Dhaka, Bangladesh, for feedback on an earlier version of this draft.

## 10. References

[FF98] Floyd, S., and Fall, K., Promoting the Use of End-to-End Congestion Control in the Internet, IEEE/ACM Transactions on Networking, August 1999. URL "http://www-nrg.ee.lbl.gov/floyd/end2end-paper.html".

[GRE] S. Hanks, T. Li, D. Farinacci, and P. Traina, Generic Routing Encapsulation (GRE), RFC 1701, October 1994. URL "http://www.ietf.cnri.reston.va.us/rfc/rfc1701.txt".

[L2TP] W. Townsley, A. Valencia, A. Rubens, G. Pall, G. Zorn, and B. Palter Layer Two Tunneling Protocol "L2TP", RFC 2661, August 1999. URL "ftp://ftp.isi.edu/in-notes/rfc2661.txt".

[MPLS] D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, J. McManus, Requirements for Traffic Engineering Over MPLS, RFC 2702, September 1999. URL "ftp://ftp.isi.edu/in-notes/rfc2702.txt".

[PPTP] Hamzeh, K., Pall, G., Verthein, W., Taarud, J., Little, W. and G. Zorn, "Point-to-Point Tunneling Protocol (PPTP)", RFC 2637, July 1999. URL "ftp://ftp.isi.edu/in-notes/rfc2637.txt".

[RFD99] Ramakrishnan, Floyd, S., and Davie, B., A Proposal to Incorporate ECN in MPLS, work in progress, June 1999. URL "http://www.aciri.org/floyd/papers/draft-ietf-mpls-ecn-00.txt".

[RFC2003] Perkins, C., IP Encapsulation within IP, RFC 2003, October 1996. URL "http://www.ietf.cnri.reston.va.us/rfc/rfc2003.txt".

[RFC 2401] S. Kent, R. Atkinson, Security Architecture for the Internet Protocol, RFC 2401, November 1998.

[RFC2407] D. Piper, The Internet IP Security Domain of Interpretation for ISAKMP, RFC 2407, November 1998.

[RFC2474] K. Nichols, S. Blake, F. Baker, D. Black, Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers, RFC 2474, December 1998.

[RFC2481] K. Ramakrishnan, S. Floyd, A Proposal to add Explicit Congestion Notification (ECN) to IP, RFC 2481, January 1999.

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

[RFC1702] Hanks, S., Li, T., Farinacci, D., and P. Traina, Generic Routing Encapsulation over IPv4 networks, RFC 1702, October 1994.

[SCWA99] Stefan Savage, Neal Cardwell, David Wetherall, and Tom Anderson, TCP Congestion Control with a Misbehaving Receiver, ACM Computer Communications Review, October 1999.

## 11. Security Considerations

Security considerations have been addressed in the main body of the document.

## AUTHORS' ADDRESSES

Sally Floyd
AT&T Center for Internet Research at ICSI (ACIRI)
Phone: +1 (510) 666-2989
Email: floyd@aciri.org
URL: http://www-nrg.ee.lbl.gov/floyd/

K. K. Ramakrishnan
TeraOptic Networks
Phone: +1 (408) 666-8650
Email: kk@teraoptic.com

David L. Black
EMC Corporation
42 South St.
Hopkinton, MA  01748
Phone:  +1 (508) 435-1000 x75140
Email: black_david@emc.com

This draft was created in October 2000.
It expires April 2001.