                     More Accurate ECN Feedback in TCP
                    draft-ietf-tcpm-accurate-ecn-17

Abstract

   Explicit Congestion Notification (ECN) is a mechanism where network
   nodes can mark IP packets instead of dropping them to indicate
   incipient congestion to the end-points. Receivers with an ECN-
   capable transport protocol feed back this information to the sender.
   ECN was originally specified for TCP in such a way that only one
   feedback signal can be transmitted per Round-Trip Time (RTT). Recent
   new TCP mechanisms like Congestion Exposure (ConEx), Data Center TCP
   (DCTCP) or Low Latency Low Loss Scalable Throughput (L4S) need more
   accurate ECN feedback information whenever more than one marking is
   received in one RTT. This document updates the original ECN
   specification to specify a scheme to provide more than one feedback
   signal per RTT in the TCP header. Given TCP header space is scarce,
   it allocates a reserved header bit previously assigned to the ECN-
   Nonce. It also overloads the two existing ECN flags in the TCP
   header. The resulting extra space is exploited to feed back the IP-
   ECN field received during the 3-way handshake as well. Supplementary
   feedback information can optionally be provided in a new TCP option,
   which is never used on the TCP SYN. The document also specifies the
   treatment of this updated TCP wire protocol by middleboxes, updating
   BCP 69 with respect to ACK filtering.

Status of This Memo

   This Internet-Draft is submitted in full conformance with the
   provisions of BCP 78 and BCP 79.

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF). Note that other groups may also distribute
   working documents as Internet-Drafts. The list of current Internet-
   Drafts is at https://datatracker.ietf.org/drafts/current/.

   Internet-Drafts are draft documents valid for a maximum of six months
   and may be updated, replaced, or obsoleted by other documents at any

   time. It is inappropriate to use Internet-Drafts as reference
   material or to cite them other than as "work in progress."

   This Internet-Draft will expire on September 8, 2022.

Copyright Notice

Table of Contents

1. Introduction

     Explicit Congestion Notification (ECN) [RFC3168] is a mechanism where
     network nodes can mark IP packets instead of dropping them to
     indicate incipient congestion to the end-points. Receivers with an
     ECN-capable transport protocol feed back this information to the
     sender. In RFC 3168, ECN was specified for TCP in such a way that
     only one feedback signal could be transmitted per Round-Trip Time

     (RTT). Recently, proposed mechanisms like Congestion Exposure (ConEx
     [RFC7713]), DCTCP [RFC8257] or L4S [I-D.ietf-tsvwg-l4s-arch] need to
     know when more than one marking is received in one RTT which is
     information that cannot be provided by the feedback scheme as
     specified in [RFC3168]. This document specifies an update to the ECN
     feedback scheme of RFC 3168 that provides more accurate information
     and could be used by these and potentially other future TCP
     extensions. A fuller treatment of the motivation for this
     specification is given in the associated requirements document
     [RFC7560].

     This documents specifies a standards track scheme for ECN feedback in
     the TCP header to provide more than one feedback signal per RTT. It
     will be called the more accurate ECN feedback scheme, or AccECN for
     short. This document updates RFC 3168 with respect to negotiation
     and use of the feedback scheme for TCP. All aspects of RFC 3168
     other than the TCP feedback scheme, in particular the definition of
     ECN at the IP layer, remain unchanged by this specification.
     Section 4 gives a more detailed specification of exactly which
     aspects of RFC 3168 this document updates.

     AccECN is intended to be a complete replacement for classic TCP/ECN
     feedback, not a fork in the design of TCP. AccECN feedback
     complements TCP's loss feedback and it can coexist alongside
     'classic' [RFC3168] TCP/ECN feedback. So its applicability is
     intended to include all public and private IP networks (and even any
     non-IP networks over which TCP is used today), whether or not any
     nodes on the path support ECN, of whatever flavour. This document
     uses the term Classic ECN when it needs to distinguish the RFC 3168

ECN TCP feedback scheme from the AccECN TCP feedback scheme.

AccECN feedback overloads the two existing ECN flags in the TCP
header and allocates the currently reserved flag (previously called
NS) in the TCP header, to be used as one three-bit counter field
indicating the number of congestion experienced marked packets.
Given the new definitions of these three bits, both ends have to
support the new wire protocol before it can be used. Therefore
during the TCP handshake the two ends use these three bits in the TCP
header to negotiate the most advanced feedback protocol that they can
both support, in a way that is backward compatible with [RFC3168].

AccECN is solely a change to the TCP wire protocol; it covers the
negotiation and signaling of more accurate ECN feedback from a TCP
Data Receiver to a Data Sender. It is completely independent of how
TCP might respond to congestion feedback, which is out of scope, but
ultimately the motivation for accurate ECN feedback. Like Classic
ECN feedback, AccECN can be used by standard Reno congestion control
[RFC5681] to respond to the existence of at least one congestion

notification within a round trip. Or, unlike Reno, AccECN can be
used to respond to the extent of congestion notification over a round
trip, as for example DCTCP does in controlled environments [RFC8257].
For congestion response, this specification refers to RFC 3168, or
ECN experiments such as those referred to in [RFC8311], namely: a
TCP-based Low Latency Low Loss Scalable (L4S) congestion control
[I-D.ietf-tsvwg-l4s-arch]; or Alternative Backoff with ECN (ABE)
[RFC8511].

It is recommended that the AccECN protocol is implemented alongside
SACK [RFC2018] and the experimental ECN++ protocol
[I-D.ietf-tcpm-generalized-ecn], which allows the ECN capability to
be used on TCP control packets. Therefore, this specification does
not discuss implementing AccECN alongside [RFC5562], which was an
earlier experimental protocol with narrower scope than ECN++.

1.1. Document Roadmap

The following introductory section outlines the goals of AccECN
(Section 1.2). Then terminology is defined (Section 1.3) and a recap
of existing prerequisite technology is given (Section 1.4).

Section 2 gives an informative overview of the AccECN protocol. Then
Section 3 gives the normative protocol specification, and Section 4
clarifies which aspects of RFC 3168 are updated by this
specification. Section 5 assesses the interaction of AccECN with
commonly used variants of TCP, whether standardized or not.
Section 6 summarizes the features and properties of AccECN.

Section 7 summarizes the protocol fields and numbers that IANA will
need to assign and Section 8 points to the aspects of the protocol
that will be of interest to the security community.

Appendix A gives pseudocode examples for the various algorithms that
AccECN uses and Appendix B explains why AccECN uses flags in the main
TCP header and quantifies the space left for future use.

1.2. Goals

[RFC7560] enumerates requirements that a candidate feedback scheme
will need to satisfy, under the headings: resilience, timeliness,
integrity, accuracy (including ordering and lack of bias),
complexity, overhead and compatibility (both backward and forward).
It recognizes that a perfect scheme that fully satisfies all the
requirements is unlikely and trade-offs between requirements are
likely. Section 6 presents the properties of AccECN against these
requirements and discusses the trade-offs made.

The requirements document recognizes that a protocol as ubiquitous as
TCP needs to be able to serve as-yet-unspecified requirements.
Therefore an AccECN receiver aims to act as a generic (dumb)
reflector of congestion information so that in future new sender
behaviours can be deployed unilaterally.

1.3. Terminology

AccECN: The more accurate ECN feedback scheme will be called AccECN
   for short.

Classic ECN: the ECN protocol specified in [RFC3168].

Classic ECN feedback: the feedback aspect of the ECN protocol
   specified in [RFC3168], including generation, encoding,

transmission and decoding of feedback, but not the Data Sender's
subsequent response to that feedback.

ACK: A TCP acknowledgement, with or without a data payload (ACK=1).

Pure ACK: A TCP acknowledgement without a data payload.

Acceptable packet / segment: A packet or segment that passes the
acceptability tests in [RFC0793] and [RFC5961].

TCP client: The TCP stack that originates a connection.

TCP server: The TCP stack that responds to a connection request.

Data Receiver: The endpoint of a TCP half-connection that receives
data and sends AccECN feedback.

Data Sender: The endpoint of a TCP half-connection that sends data
and receives AccECN feedback.

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 BCP 14 [RFC2119]
[RFC8174] when, and only when, they appear in all capitals, as shown
here.

1.4. Recap of Existing ECN feedback in IP/TCP

ECN [RFC3168] uses two bits in the IP header. Once ECN has been
negotiated with the receiver at the transport layer, an ECN sender
can set two possible codepoints (ECT(0) or ECT(1)) in the IP header
to indicate an ECN-capable transport (ECT). If both ECN bits are
zero, the packet is considered to have been sent by a Not-ECN-capable

Transport (Not-ECT). When a network node experiences congestion, it
will occasionally either drop or mark a packet, with the choice
depending on the packet's ECN codepoint. If the codepoint is Not-
ECT, only drop is appropriate. If the codepoint is ECT(0) or ECT(1),
the node can mark the packet by setting both ECN bits, which is
termed 'Congestion Experienced' (CE), or loosely a 'congestion mark'.
Table 1 summarises these codepoints.

```
+------------------+----------------+--------------------------+
| IP-ECN codepoint | Codepoint name | Description              |
+------------------+----------------+--------------------------+
| 0b00             | Not-ECT        | Not ECN-Capable Transport|
| 0b01             | ECT(1)         | ECN-Capable Transport (1)|
| 0b10             | ECT(0)         | ECN-Capable Transport (0)|
| 0b11             | CE             | Congestion Experienced   |
+------------------+----------------+--------------------------+
```

                 Table 1: The ECN Field in the IP Header

   In the TCP header the first two bits in byte 14 are defined as flags
   for the use of ECN (CWR and ECE in Figure 1 [RFC3168]). A TCP client
   indicates it supports ECN by setting ECE=CWR=1 in the SYN, and an
   ECN-enabled server confirms ECN support by setting ECE=1 and CWR=0 in
   the SYN/ACK. On reception of a CE-marked packet at the IP layer, the
   Data Receiver starts to set the Echo Congestion Experienced (ECE)
   flag continuously in the TCP header of ACKs, which ensures the signal
   is received reliably even if ACKs are lost. The TCP sender confirms
   that it has received at least one ECE signal by responding with the
   congestion window reduced (CWR) flag, which allows the TCP receiver
   to stop repeating the ECN-Echo flag. This always leads to a full RTT
   of ACKs with ECE set. Thus any additional CE markings arriving
   within this RTT cannot be fed back.

   The last bit in byte 13 of the TCP header was defined as the Nonce
   Sum (NS) for the ECN Nonce [RFC3540]. In the absence of widespread
   deployment RFC 3540 has been reclassified as historic [RFC8311] and
   the respective flag has been marked as "reserved", making this TCP
   flag available for use by the AccECN experiment instead.

```
     0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
   +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
   |                   |           | N | C | E | U | A | P | R | S | F |
   | Header Length     | Reserved  | S | W | C | R | C | S | S | Y | I |
   |                   |           |   | R | E | G | K | H | T | N | N |
   +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
```
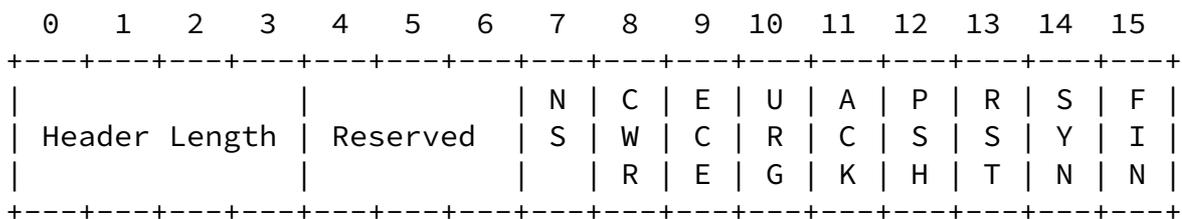
   Figure 1: The (post-ECN Nonce) definition of the TCP header flags

2. AccECN Protocol Overview and Rationale

This section provides an informative overview of the AccECN protocol that will be normatively specified in [Section 3](#)

Like the original TCP approach, the Data Receiver of each TCP half-connection sends AccECN feedback to the Data Sender on TCP acknowledgements, reusing data packets of the other half-connection whenever possible.

The AccECN protocol has had to be designed in two parts:

o   an essential part that re-uses ECN TCP header bits for the Data Receiver to feed back the number of packets arriving with CE in the IP-ECN field. This provides more accuracy than classic ECN feedback, but limited resilience against ACK loss;

o   a supplementary part using a new AccECN TCP Option that provides additional feedback on the number of bytes that arrive marked with each of the three ECN codepoints in the IP-ECN field (not just CE marks). This provides greater resilience against ACK loss than the essential feedback, but it is more likely to suffer from middlebox interference.

The two part design was necessary, given limitations on the space available for TCP options and given the possibility that certain incorrectly designed middleboxes prevent TCP using any new options.

The essential part overloads the previous definition of the three flags in the TCP header that had been assigned for use by ECN. This design choice deliberately replaces the classic ECN feedback protocol, rather than leaving classic ECN feedback intact and adding more accurate feedback separately because:

o   this efficiently reuses scarce TCP header space, given TCP option space is approaching saturation;

o   a single upgrade path for the TCP protocol is preferable to a fork in the design;

o   otherwise classic and accurate ECN feedback could give conflicting feedback on the same segment, which could open up new security concerns and make implementations unnecessarily complex;

o   middleboxes are more likely to faithfully forward the TCP ECN flags than newly defined areas of the TCP header.

AccECN is designed to work even if the supplementary part is removed or zeroed out, as long as the essential part gets through.

2.1. Capability Negotiation

AccECN is a change to the wire protocol of the main TCP header, therefore it can only be used if both endpoints have been upgraded to understand it. The TCP client signals support for AccECN on the initial SYN of a connection and the TCP server signals whether it supports AccECN on the SYN/ACK. The TCP flags on the SYN that the client uses to signal AccECN support have been carefully chosen so that a TCP server will interpret them as a request to support the most recent variant of ECN feedback that it supports. Then the client falls back to the same variant of ECN feedback.

An AccECN TCP client does not send the new AccECN Option on the SYN as SYN option space is limited. The TCP server sends the AccECN Option on the SYN/ACK and the client sends it on the first ACK to test whether the network path forwards the option correctly.

2.2. Feedback Mechanism

A Data Receiver maintains four counters initialized at the start of the half-connection. Three count the number of arriving payload bytes respectively marked CE, ECT(1) and ECT(0) in the IP-ECN field. The fourth counts the number of packets arriving marked with a CE codepoint (including control packets without payload if they are CE-marked).

The Data Sender maintains four equivalent counters for the half connection, and the AccECN protocol is designed to ensure they will match the values in the Data Receiver's counters, albeit after a little delay.

Each ACK carries the three least significant bits (LSBs) of the packet-based CE counter using the ECN bits in the TCP header, now renamed the Accurate ECN (ACE) field (see Figure 3 later). The 24 LSBs of each byte counter are carried in the AccECN Option.

2.3. Delayed ACKs and Resilience Against ACK Loss

With both the ACE and the AccECN Option mechanisms, the Data Receiver continually repeats the current LSBs of each of its respective counters. There is no need to acknowledge these continually repeated counters, so the congestion window reduced (CWR) mechanism is no longer used. Even if some ACKs are lost, the Data Sender should be

able to infer how much to increment its own counters, even if the protocol field has wrapped.

The 3-bit ACE field can wrap fairly frequently. Therefore, even if it appears to have incremented by one (say), the field might have actually cycled completely then incremented by one. The Data Receiver is not allowed to delay sending an ACK to such an extent that the ACE field would cycle. However cycling is still a possibility at the Data Sender because a whole sequence of ACKs carrying intervening values of the field might all be lost or delayed in transit.

The fields in the AccECN Option are larger, but they will increment in larger steps because they count bytes not packets. Nonetheless, their size has been chosen such that a whole cycle of the field would never occur between ACKs unless there had been an infeasibly long sequence of ACK losses. Therefore, as long as the AccECN Option is available, it can be treated as a dependable feedback channel.

If the AccECN Option is not available, e.g. it is being stripped by a middlebox, the AccECN protocol will only feed back information on CE markings (using the ACE field). Although not ideal, this will be sufficient, because it is envisaged that neither ECT(0) nor ECT(1) will ever indicate more severe congestion than CE, even though future uses for ECT(0) or ECT(1) are still unclear [RFC8311]. Because the 3-bit ACE field is so small, when it is the only field available, the Data Sender has to interpret it assuming the most likely wrap, but with a degree of conservatism.

Certain specified events trigger the Data Receiver to include an AccECN Option on an ACK. The rules are designed to ensure that the order in which different markings arrive at the receiver is communicated to the sender (as long as options are reaching the sender and as long as there is no ACK loss). Implementations are encouraged to send an AccECN Option more frequently, but this is left up to the implementer.

2.4. Feedback Metrics

The CE packet counter in the ACE field and the CE byte counter in the AccECN Option both provide feedback on received CE-marks. The CE packet counter includes control packets that do not have payload

data, while the CE byte counter solely includes marked payload bytes.
If both are present, the byte counter in the option will provide the
more accurate information needed for modern congestion control and
policing schemes, such as L4S, DCTCP or ConEx. If the option is
stripped, a simple algorithm to estimate the number of marked bytes
from the ACE field is given in [Appendix A.3](#).

Feedback in bytes is recommended in order to protect against the
receiver using attacks similar to 'ACK-Division' to artificially

inflate the congestion window, which is why [RFC5681] now recommends
that TCP counts acknowledged bytes not packets.

2.5. Generic (Dumb) Reflector

The ACE field provides feedback about CE markings in the IP-ECN field
of both data and control packets. According to [RFC3168] the Data
Sender is meant to set the IP-ECN field of control packets to Not-
ECT. However, mechanisms in certain private networks (e.g. data
centres) set control packets to be ECN capable because they are
precisely the packets that performance depends on most.

For this reason, AccECN is designed to be a generic reflector of
whatever ECN markings it sees, whether or not they are compliant with
a current standard. Then as standards evolve, Data Senders can
upgrade unilaterally without any need for receivers to upgrade too.
It is also useful to be able to rely on generic reflection behaviour
when senders need to test for unexpected interference with markings
(for instance Section 3.2.2.3, Section 3.2.2.4 and Section 3.2.3.2 of
the present document and para 2 of Section 20.2 of [RFC3168]).

The initial SYN is the most critical control packet, so AccECN
provides feedback on its IP-ECN field. Although RFC 3168 prohibits
an ECN-capable SYN, providing feedback of ECN marking on the SYN
supports future scenarios in which SYNs might be ECN-enabled (without
prejudging whether they ought to be). For instance, [RFC8311]
updates this aspect of RFC 3168 to allow experimentation with ECN-
capable TCP control packets.

Even if the TCP client (or server) has set the SYN (or SYN/ACK) to
not-ECT in compliance with RFC 3168, feedback on the state of the IP-
ECN field when it arrives at the receiver could still be useful,

because middleboxes have been known to overwrite the IP-ECN field as
if it is still part of the old Type of Service (ToS) field
[Mandalari18]. If a TCP client has set the SYN to Not-ECT, but
receives feedback that the IP-ECN field on the SYN arrived with a
different codepoint, it can detect such middlebox interference and
send Not-ECT for the rest of the connection. Previously, if a TCP
server received ECT or CE on a SYN, it could not know whether it was
invalid (or valid) because only the TCP client knew whether it
originally marked the SYN as Not-ECT (or ECT). Therefore, prior to
AccECN, the server's only safe course of action was to disable ECN
for the connection. Instead, the AccECN protocol allows the server
to feed back the received ECN field to the client, which then has all
the information to decide whether the connection has to fall-back
from supporting ECN (or not).

3. AccECN Protocol Specification

3.1. Negotiating to use AccECN

3.1.1. Negotiation during the TCP handshake

   Given the ECN Nonce [RFC3540] has been reclassified as historic
   [RFC8311], the present specification re-allocates the TCP flag at bit
   7 of the TCP header, which was previously called NS (Nonce Sum), as
   the AE (Accurate ECN) flag (see IANA Considerations in Section 7) as
   shown below.

```
      0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
    +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
    |                       |   | A | C | E | U | A | P | R | S | F |
    | Header Length | Reserved  | E | W | C | R | C | S | S | Y | I |
    |                       |   |   | R | E | G | K | H | T | N | N |
    +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
```
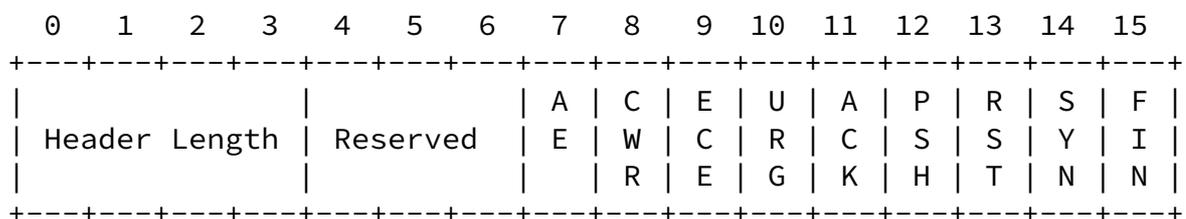
                Figure 2: The (post-AccECN) definition of the TCP header flags during
                          the TCP handshake

   During the TCP handshake at the start of a connection, to request
   more accurate ECN feedback the TCP client (host A) MUST set the TCP
   flags AE=1, CWR=1 and ECE=1 in the initial SYN segment.

If a TCP server (B) that is AccECN-enabled receives a SYN with the
above three flags set, it MUST set both its half connections into
AccECN mode. Then it MUST set the AE, CWR and ECE TCP flags on the
SYN/ACK to the combination in the top block of Table 2 that feeds
back the IP-ECN field that arrived on the SYN. This applies whether
or not the server itself supports setting the IP-ECN field on a SYN
or SYN/ACK (see Section 2.5 for rationale).

When the TCP server returns any of the 4 combinations in the top
block of Table 2, it confirms that it supports AccECN. The TCP
server MUST NOT set one of these 4 combination of flags on the SYN/
ACK unless the preceding SYN requested support for AccECN as above.

Once a TCP client (A) has sent the above SYN to declare that it
supports AccECN, and once it has received the above SYN/ACK segment
that confirms that the TCP server supports AccECN, the TCP client
MUST set both its half connections into AccECN mode.

Once in AccECN mode, a TCP client or server has the rights and
obligations to participate in the ECN protocol defined in
Section 3.1.5.

The procedure for the client to follow if a SYN/ACK does not arrive
before its retransmission timer expires is given in Section 3.1.4.

3.1.2. Backward Compatibility

The three flags set to 1 to indicate AccECN support on the SYN have
been carefully chosen to enable natural fall-back to prior stages in
the evolution of ECN, as above. Table 2 tabulates all the
negotiation possibilities for ECN-related capabilities that involve
at least one AccECN-capable host. The entries in the first two
columns have been abbreviated, as follows:

AccECN: More Accurate ECN Feedback (the present specification)

Nonce: ECN Nonce feedback [RFC3540]

ECN: 'Classic' ECN feedback [RFC3168]

No ECN: Not-ECN-capable. Implicit congestion notification using
     packet drop.

```
+--------+--------+-----------+-----------+----------------------+
| A      | B      | SYN A->B  |  SYN/ACK  | Feedback Mode        |
|        |        |           |   B->A    |                      |
+--------+--------+-----------+-----------+----------------------+
|        |        | AE CWR ECE| AE CWR ECE|                      |
| AccECN | AccECN | 1   1   1 | 0   1   0 | AccECN(no ECT on SYN)|
| AccECN | AccECN | 1   1   1 | 0   1   1 | AccECN (ECT1 on SYN) |
| AccECN | AccECN | 1   1   1 | 1   0   0 | AccECN (ECT0 on SYN) |
| AccECN | AccECN | 1   1   1 | 1   1   0 | AccECN (CE on SYN)   |
|        |        |           |           |                      |
| AccECN | Nonce  | 1   1   1 | 1   0   1 | (Reserved)           |
| AccECN | ECN    | 1   1   1 | 0   0   1 | classic ECN          |
| AccECN | No ECN | 1   1   1 | 0   0   0 | Not ECN              |
|        |        |           |           |                      |
| Nonce  | AccECN | 0   1   1 | 0   0   1 | classic ECN          |
| ECN    | AccECN | 0   1   1 | 0   0   1 | classic ECN          |
| No ECN | AccECN | 0   0   0 | 0   0   0 | Not ECN              |
|        |        |           |           |                      |
| AccECN | Broken | 1   1   1 | 1   1   1 | Not ECN              |
+--------+--------+-----------+-----------+----------------------+
```

     Table 2: ECN capability negotiation between Client (A) and Server (B)

     Table 2 is divided into blocks each separated by an empty row.

   1. The top block shows the case already described in Section 3.1
       where both endpoints support AccECN and how the TCP server (B)
       indicates congestion feedback.

   2. The second block shows the cases where the TCP client (A)
       supports AccECN but the TCP server (B) supports some earlier
       variant of TCP feedback, indicated in its SYN/ACK. Therefore, as
       soon as an AccECN-capable TCP client (A) receives the SYN/ACK
       shown it MUST set both its half connections into the feedback
       mode shown in the rightmost column. If it has set itself into
       classic ECN feedback mode it MUST then comply with [RFC3168].

The server response called 'Nonce' in the table is now historic.
For an AccECN implementation, there is no need to recognize or
support ECN Nonce feedback [RFC3540], which has been reclassified
as historic [RFC8311]. AccECN is compatible with alternative ECN
feedback integrity approaches (see Section 5.3).

3. The third block shows the cases where the TCP server (B) supports
    AccECN but the TCP client (A) supports some earlier variant of
    TCP feedback, indicated in its SYN.

    When an AccECN-enabled TCP server (B) receives a SYN with
    AE,CWR,ECE = 0,1,1 it MUST do one of the following:

    *   set both its half connections into the classic ECN feedback
        mode and return a SYN/ACK with AE, CWR, ECE = 0,0,1 as shown.
        Then it MUST comply with [RFC3168].

    *   set both its half-connections into No ECN mode and return a
        SYN/ACK with AE,CWR,ECE = 0,0,0, then continue with ECN
        disabled. This latter case is unlikely to be desirable, but
        it is allowed as a possibility, e.g. for minimal TCP
        implementations.

    When an AccECN-enabled TCP server (B) receives a SYN with
    AE,CWR,ECE = 0,0,0 it MUST set both its half connections into the
    Not ECN feedback mode, return a SYN/ACK with AE,CWR,ECE = 0,0,0
    as shown and continue with ECN disabled.

4. The fourth block displays a combination labelled `Broken'. Some
    older TCP server implementations incorrectly set the reserved
    flags in the SYN/ACK by reflecting those in the SYN. Such broken
    TCP servers (B) cannot support ECN, so as soon as an AccECN-
    capable TCP client (A) receives such a broken SYN/ACK it MUST
    fall back to Not ECN mode for both its half connections and
    continue with ECN disabled.

The following additional rules do not fit the structure of the table,
but they complement it:

Simultaneous Open: An originating AccECN Host (A), having sent a SYN

with AE=1, CWR=1 and ECE=1, might receive another SYN from host B.
   Host A MUST then enter the same feedback mode as it would have
   entered had it been a responding host and received the same SYN.
   Then host A MUST send the same SYN/ACK as it would have sent had
   it been a responding host.

   In-window SYN during TIME-WAIT: Many TCP implementations create a
      new TCP connection if they receive an in-window SYN packet during
      TIME-WAIT state. When a TCP host enters TIME-WAIT or CLOSED
      state, it should ignore any previous state about the negotiation
      of AccECN for that connection and renegotiate the feedback mode
      according to Table 2.

3.1.3. Forward Compatibility

   If a TCP server that implements AccECN receives a SYN with the three
   TCP header flags (AE, CWR and ECE) set to any combination other than
   000, 011 or 111, it MUST negotiate the use of AccECN as if they had
   been set to 111. This ensures that future uses of the other
   combinations on a SYN can rely on consistent behaviour from the
   installed base of AccECN servers.

   For the avoidance of doubt, the behaviour described in the present
   specification applies whether or not the three remaining reserved TCP
   header flags are zero.

3.1.4. Retransmission of the SYN

   If the sender of an AccECN SYN times out before receiving the SYN/
   ACK, the sender SHOULD attempt to negotiate the use of AccECN at
   least one more time by continuing to set all three TCP ECN flags on
   the first retransmitted SYN (using the usual retransmission time-
   outs). If this first retransmission also fails to be acknowledged,
   the sender SHOULD send subsequent retransmissions of the SYN with the
   three TCP-ECN flags cleared (AE=CWR=ECE=0). A retransmitted SYN MUST
   use the same ISN as the original SYN.

   Retrying once before fall-back adds delay in the case where a
   middlebox drops an AccECN (or ECN) SYN deliberately. However,
   current measurements imply that a drop is less likely to be due to
   middlebox interference than other intermittent causes of loss,
   e.g. congestion, wireless interference, etc.

Implementers MAY use other fall-back strategies if they are found to be more effective (e.g. attempting to negotiate AccECN on the SYN only once or more than twice (most appropriate during high levels of congestion). However, other fall-back strategies will need to follow all the rules in Section 3.1.5, which concern behaviour when SYNs or SYN/ACKs negotiating different types of feedback have been sent within the same connection.

Further it may make sense to also remove any other new or experimental fields or options on the SYN in case a middlebox might be blocking them, although the required behaviour will depend on the specification of the other option(s) and any attempt to co-ordinate fall-back between different modules of the stack.

Whichever fall-back strategy is used, the TCP initiator SHOULD cache failed connection attempts. If it does, it SHOULD NOT give up attempting to negotiate AccECN on the SYN of subsequent connection attempts until it is clear that the blockage is persistently and specifically due to AccECN. The cache should be arranged to expire so that the initiator will infrequently attempt to check whether the problem has been resolved.

The fall-back procedure if the TCP server receives no ACK to acknowledge a SYN/ACK that tried to negotiate AccECN is specified in Section 3.2.3.2.

3.1.5. Implications of AccECN Mode

Section 3.1.1 describes the only ways that a host can enter AccECN mode, whether as a client or as a server.

As a Data Sender, a host in AccECN mode has the rights and obligations concerning the use of ECN defined below, which build on those in [RFC3168] as updated by [RFC8311]:

o Using ECT:

   * It can set an ECT codepoint in the IP header of packets to indicate to the network that the transport is capable and willing to participate in ECN for this packet.

   * It does not have to set ECT on any packet (for instance if it has reason to believe such a packet would be blocked).

o Switching feedback negotiation (e.g. fall-back):

   * It SHOULD NOT set ECT on any packet if it has received at least

one valid SYN or Acceptable SYN/ACK with AE=CWR=ECE=0. A

"valid SYN" has the same port numbers and the same ISN as the
SYN that caused the server to enter AccECN mode.

*   It MUST NOT send an ECN-setup SYN [RFC3168] within the same
    connection as it has sent a SYN requesting AccECN feedback.

*   It MUST NOT send an ECN-setup SYN/ACK [RFC3168] within the same
    connection as it has sent a SYN/ACK agreeing to use AccECN
    feedback.

The above rules are necessary because, if one peer were to
negotiate the feedback mode in two different types of handshake,
it would not be possible for the other peer to know for certain
which handshake packet(s) the other end had eventually received or
in which order it received them. So, in the absence of these
rules, the two peers could end up using different feedback modes
without knowing it.

o Congestion response:

*   It is still obliged to respond appropriately to AccECN feedback
    that indicates there were ECN marks on packets it had
    previously sent, as defined in Section 6.1 of [RFC3168] and
    updated by Sections 2.1 and 4.1 of [RFC8311].

*   The commitment to respond appropriately to incoming indications
    of congestion remains even if it sends a SYN packet with
    AE=CWR=ECE=0, in a later transmission within the same TCP
    connection.

*   Unlike an RFC 3168 data sender, it MUST NOT set CWR to indicate
    it has received and responded to indications of congestion (for
    the avoidance of doubt, this does not preclude it from setting
    the bits of the ACE counter field, which includes an overloaded
    use of the same bit).

As a Data Receiver:

o a host in AccECN mode MUST feed back the information in the IP-ECN
   field of incoming packets using Accurate ECN feedback, as

specified in [Section 3.2](#) below.

   o if it receives an ECN-setup SYN or ECN-setup SYN/ACK [[RFC3168](#)]
      during the same connection as it receives a SYN requesting AccECN
      feedback or a SYN/ACK agreeing to use AccECN feedback, it MUST
      reset the connection with a RST packet.

   o If for any reason it is not willing to provide ECN feedback on a
      particular TCP connection, to indicate this unwillingness it
      SHOULD clear the AE, CWR and ECE flags in all SYN and/or SYN/ACK
      packets that it sends.

   o it MUST NOT use reception of packets with ECT set in the IP-ECN
      field as an implicit signal that the peer is ECN-capable. Reason:
      ECT at the IP layer does not explicitly confirm the peer has the
      correct ECN feedback logic, as the packets could have been mangled
      at the IP layer.

[3.2](#). AccECN Feedback

   Each Data Receiver of each half connection maintains four counters,
   r.cep, r.ceb, r.e0b and r.e1b:

   o The Data Receiver MUST increment the CE packet counter (r.cep),
      for every Acceptable packet that it receives with the CE code
      point in the IP ECN field, including CE marked control packets but
      excluding CE on SYN packets (SYN=1; ACK=0).

   o A Data Receiver that supports sending of the AccECN TCP Option
      MUST increment the r.ceb, r.e0b or r.e1b byte counters by the
      number of TCP payload octets in Acceptable packets marked
      respectively with the CE, ECT(0) and ECT(1) codepoint in their IP-
      ECN field, including any payload octets on control packets, but
      not including any payload octets on SYN packets (SYN=1; ACK=0).

   Each Data Sender of each half connection maintains four counters,
   s.cep, s.ceb, s.e0b and s.e1b intended to track the equivalent
   counters at the Data Receiver.

   A Data Receiver feeds back the CE packet counter using the Accurate

ECN (ACE) field, as explained in Section 3.2.2. And it optionally
feeds back all the byte counters using the AccECN TCP Option, as
specified in Section 3.2.3.

Whenever a host feeds back the value of any counter, it MUST report
the most recent value, no matter whether it is in a pure ACK, an ACK
with new payload data or a retransmission. Therefore the feedback
carried on a retransmitted packet is unlikely to be the same as the
feedback on the original packet.

3.2.1. Initialization of Feedback Counters

When a host first enters AccECN mode, in its role as a Data Receiver
it initializes its counters to r.cep = 5, r.e0b = r.e1b = 1 and r.ceb
= 0,

Non-zero initial values are used to support a stateless handshake
(see Section 5.1) and to be distinct from cases where the fields are
incorrectly zeroed (e.g. by middleboxes - see Section 3.2.3.2.4).

When a host enters AccECN mode, in its role as a Data Sender it
initializes its counters to s.cep = 5, s.e0b = s.e1b = 1 and s.ceb =
0.

3.2.2. The ACE Field

After AccECN has been negotiated on the SYN and SYN/ACK, both hosts
overload the three TCP flags (AE, CWR and ECE) in the main TCP header
as one 3-bit field. Then the field is given a new name, ACE, as
shown in Figure 3.

```
   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
 +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
 |               |           |           | U | A | P | R | S | F |
 | Header Length | Reserved  |    ACE    | R | C | S | S | Y | I |
 |               |           |           | G | K | H | T | N | N |
 +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
```
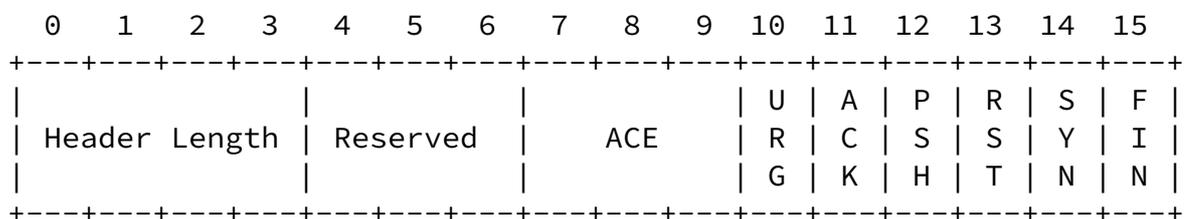
   Figure 3: Definition of the ACE field within bytes 13 and 14 of the
        TCP Header (when AccECN has been negotiated and SYN=0).

The original definition of these three flags in the TCP header,

including the addition of support for the ECN Nonce, is shown for
comparison in Figure 1. This specification does not rename these
three TCP flags to ACE unconditionally; it merely overloads them with
another name and definition once an AccECN connection has been
established.

With one exception ([Section 3.2.2.1](#)), a host with both of its half-
connections in AccECN mode MUST interpret the AE, CWR and ECE flags
as the 3-bit ACE counter on a segment with the SYN flag cleared
(SYN=0). On such a packet, a Data Receiver MUST encode the three
least significant bits of its r.cep counter into the ACE field that
it feeds back to the Data Sender. A host MUST NOT interpret the 3
flags as a 3-bit ACE field on any segment with SYN=1 (whether ACK is
0 or 1), or if AccECN negotiation is incomplete or has not succeeded.

Both parts of each of these conditions are equally important. For
instance, even if AccECN negotiation has been successful, the ACE
field is not defined on any segments with SYN=1 (e.g. a
retransmission of an unacknowledged SYN/ACK, or when both ends send
SYN/ACKs after AccECN support has been successfully negotiated during
a simultaneous open).

3.2.2.1. ACE Field on the ACK of the SYN/ACK

A TCP client (A) in AccECN mode MUST feed back which of the 4
possible values of the IP-ECN field was on the SYN/ACK by writing it
into the ACE field of a pure ACK with no SACK blocks using the binary
encoding in Table 3 (which is the same as that used on the SYN/ACK in
Table 2). This shall be called the handshake encoding of the ACE
field, and it is the only exception to the rule that the ACE field
carries the 3 least significant bits of the r.cep counter on packets
with SYN=0.

Normally, a TCP client acknowledges a SYN/ACK with an ACK that
satisfies the above conditions anyway (SYN=0, no data, no SACK
blocks). If an AccECN TCP client intends to acknowledge the SYN/ACK
with a packet that does not satisfy these conditions (e.g. it has
data to include on the ACK), it SHOULD first send a pure ACK that
does satisfy these conditions (see [Section 5.2](#)), so that it can feed
back which of the four values of the IP-ECN field arrived on the SYN/
ACK. A valid exception to this "SHOULD" would be where the

implementation will only be used in an environment where mangling of
the ECN field is unlikely.

IP-ECN codepoint on SYN/ACK	ACE on pure ACK of SYN/ACK	r.cep of client in AccECN mode
Not-ECT	0b010	5
ECT(1)	0b011	5
ECT(0)	0b100	5
CE	0b110	6

Table 3: The encoding of the ACE field in the ACK of the SYN-ACK to
reflect the SYN-ACK's IP-ECN field

When an AccECN server in SYN-RCVD state receives a pure ACK with
SYN=0 and no SACK blocks, instead of treating the ACE field as a
counter, it MUST infer the meaning of each possible value of the ACE
field from Table 4, which also shows the value that an AccECN server
MUST set s.cep to as a result.

Given this encoding of the ACE field on the ACK of a SYN/ACK is
exceptional, an AccECN server using large receive offload (LRO) might
prefer to disable LRO until such an ACK has transitioned it out of
SYN-RCVD state.

ACE on ACK of SYN/ACK	IP-ECN codepoint on SYN/ACK inferred by server	s.cep of server in AccECN mode
0b000	{Notes 1, 3}	Disable ECN
0b001	{Notes 2, 3}	5
0b010	Not-ECT	5
0b011	ECT(1)	5
0b100	ECT(0)	5
0b101	Currently Unused {Note 2}	5
0b110	CE	6
0b111	Currently Unused {Note 2}	5

```
    +--------------+---------------------------+--------------------+
```

         Table 4: Meaning of the ACE field on the ACK of the SYN/ACK

   {Note 1}: If the server is in AccECN mode, the value of zero raises
   suspicion of zeroing of the ACE field on the path (see
   Section 3.2.2.4).

   {Note 2}: If the server is in AccECN mode, these values are Currently
   Unused but the AccECN server's behaviour is still defined for forward
   compatibility. Then the designer of a future protocol can know for
   certain what AccECN servers will do with these codepoints.

   {Note 3}: In the case where a server that implements AccECN is also
   using a stateless handshake (termed a SYN cookie) it will not
   remember whether it entered AccECN mode. The values 0b000 or 0b001
   will remind it that it did not enter AccECN mode, because AccECN does
   not use them (see Section 5.1 for details). If a stateless server
   that implements AccECN receives either of these two values in the
   ACK, its action is implementation-dependent and outside the scope of
   this spec, It will certainly not take the action in the third column
   because, after it receives either of these values, it is not in
   AccECN mode. I.e., it will not disable ECN (at least not just
   because ACE is 0b000) and it will not set s.cep.

3.2.2.2. Encoding and Decoding Feedback in the ACE Field

   Whenever the Data Receiver sends an ACK with SYN=0 (with or without
   data), unless the handshake encoding in Section 3.2.2.1 applies, the
   Data Receiver MUST encode the least significant 3 bits of its r.cep
   counter into the ACE field (see Appendix A.2).

   Whenever the Data Sender receives an ACK with SYN=0 (with or without
   data), it first checks whether it has already been superseded by
   another ACK in which case it ignores the ECN feedback. If the ACK
   has not been superseded, and if the special handshake encoding in

   Section 3.2.2.1 does not apply, the Data Sender decodes the ACE field
   as follows (see Appendix A.2 for examples).

   o It takes the least significant 3 bits of its local s.cep counter
      and subtracts them from the incoming ACE counter to work out the

minimum positive increment it could apply to s.cep (assuming the
ACE field only wrapped at most once).

o It then follows the safety procedures in [Section 3.2.2.5.2](#) to
   calculate or estimate how many packets the ACK could have
   acknowledged under the prevailing conditions to determine whether
   the ACE field might have wrapped more than once.

The encode/decode procedures during the three-way handshake are
exceptions to the general rules given so far, so they are spelled out
step by step below for clarity:

o If a TCP server in AccECN mode receives a CE mark in the IP-ECN
   field of a SYN (SYN=1, ACK=0), it MUST NOT increment r.cep (it
   remains at its initial value of 5).

   Reason: It would be redundant for the server to include CE-marked
   SYNs in its r.cep counter, because it already reliably delivers
   feedback of any CE marking using the encoding in Table 2 in the
   SYN/ACK. This also ensures that, when the server starts using the
   ACE field, it has not unnecessarily consumed more than one initial
   value, given they can be used to negotiate variants of the AccECN
   protocol (see [Appendix B.3](#)).

o If a TCP client in AccECN mode receives CE feedback in the TCP
   flags of a SYN/ACK, it MUST NOT increment s.cep (it remains at its
   initial value of 5), so that it stays in step with r.cep on the
   server. Nonetheless, the TCP client still triggers the congestion
   control actions necessary to respond to the CE feedback.

o If a TCP client in AccECN mode receives a CE mark in the IP-ECN
   field of a SYN/ACK, it MUST increment r.cep, but no more than once
   no matter how many CE-marked SYN/ACKs it receives
   (i.e. incremented from 5 to 6, but no further).

   Reason: Incrementing r.cep ensures the client will eventually
   deliver any CE marking to the server reliably when it starts using
   the ACE field. Even though the client also feeds back any CE
   marking on the ACK of the SYN/ACK using the encoding in Table 3,
   this ACK is not delivered reliably, so it can be considered as a
   timely notification that is redundant but unreliable. The client
   does not increment r.cep more than once, because the server can

only increment s.cep once (see next bullet). Also, this limits
the unnecessarily consumed initial values of the ACE field to two.

o If a TCP server in AccECN mode and in SYN-RCVD state receives CE
   feedback in the TCP flags of a pure ACK with no SACK blocks, it
   MUST increment s.cep (from 5 to 6). The TCP server then triggers
   the congestion control actions necessary to respond to the CE
   feedback.

   Reasoning: The TCP server can only increment s.cep once, because
   the first ACK it receives will cause it to transition out of SYN-
   RCVD state. The server's congestion response would be no
   different even if it could receive feedback of more than one CE-
   marked SYN/ACK.

   Once the TCP server transitions to ESTABLISHED state, it might
   later receive other pure ACK(s) with the handshake encoding in the
   ACE field. A server MAY implement a test for such a case, but it
   is not required. Therefore, once in the ESTABLISHED state, it
   will be sufficient for the server to consider the ACE field to be
   encoded as the normal ACE counter on all packets with SYN=0.

   Reasoning: Such ACKs will be quite unusual, e.g. a SYN/ACK (or ACK
   of the SYN/ACK) that is delayed for longer than the server's
   retransmission timeout; or packet duplication by the network. And
   the impact of any error in the feedback on such ACKs will only be
   temporary.

3.2.2.3. Testing for Mangling of the IP/ECN Field

   The value of the ACE field on the SYN/ACK indicates the value of the
   IP/ECN field when the SYN arrived at the server. The client can
   compare this with how it originally set the IP/ECN field on the SYN.
   If this comparison implies an unsafe transition (see below) of the
   IP/ECN field, for the remainder of the connection the client MUST NOT
   send ECN-capable packets, but it MUST continue to feed back any ECN
   markings on arriving packets.

   The value of the ACE field on the last ACK of the 3WHS indicates the
   value of the IP/ECN field when the SYN/ACK arrived at the client.
   The server can compare this with how it originally set the IP/ECN
   field on the SYN/ACK. If this comparison implies an unsafe
   transition of the IP/ECN field, for the remainder of the connection
   the server MUST NOT send ECN-capable packets, but it MUST continue to
   feed back any ECN markings on arriving packets.

   The ACK of the SYN/ACK is not reliably delivered (nonetheless, the
   count of CE marks is still eventually delivered reliably). If this

    ACK does not arrive, the server can continue to send ECN-capable
    packets without having tested for mangling of the IP/ECN field on the
    SYN/ACK.

    Invalid transitions of the IP/ECN field are defined in [RFC3168] and
    repeated here for convenience:

    o the not-ECT codepoint changes;

    o either ECT codepoint transitions to not-ECT;

    o the CE codepoint changes.

    RFC 3168 says that a router that changes ECT to not-ECT is invalid
    but safe. However, from a host's viewpoint, this transition is
    unsafe because it could be the result of two transitions at different
    routers on the path: ECT to CE (safe) then CE to not-ECT (unsafe).
    This scenario could well happen where an ECN-enabled home router
    congests its upstream mobile broadband bottleneck link, then the
    ingress to the mobile network clears the ECN field [Mandalari18].

    Once a Data Sender has entered AccECN mode it SHOULD check whether it
    is receiving continuous CE marking. Specifying exactly how to do
    this is beyond the scope of the present specification, but the sender
    might check whether the feedback for every packet it sends for the
    first three or four rounds indicates CE-marking. If continuous CE-
    marking is detected, for the remainder of the connection the Data
    Sender SHOULD NOT send ECN-capable packets and consequently it SHOULD
    NOT respond to any ECN feedback. The phrase 'MUST NOT' has been
    avoided to allow the sender to test whether it can resume sending
    ECN-capable packets. Throughout, it MUST remain in the AccECN
    feedback mode and it MUST continue to feed back any ECN markings on
    arriving packets (in its role as Data Receiver).

    All the fall-back behaviours in this section are necessary in case
    mangling of the IP/ECN field is asymmetric, which is currently common
    over some mobile networks [Mandalari18]. Then one end might see no
    unsafe transition and continue sending ECN-capable packets, while the
    other end sees an unsafe transition and stops sending ECN-capable
    packets.

3.2.2.4. Testing for Zeroing of the ACE Field

required the Data Receiver to initialize the r.cep
counter to a non-zero value. Therefore, in either direction the
initial value of the ACE counter ought to be non-zero.

   If AccECN has been successfully negotiated, the Data Sender SHOULD
   check the value of the ACE counter in the first packet (with or
   without data) that arrives with SYN=0. If the value of this ACE
   field is zero (0b000), the Data Sender disables sending ECN-capable
   packets for the remainder of the half-connection by setting the IP/
   ECN field in all subsequent packets to Not-ECT.

   Usually, the server checks the ACK of the SYN/ACK from the client,
   while the client checks the first data segment from the server.
   However, if reordering occurs, "the first packet ... that arrives"
   will not necessarily be the same as the first packet in sequence
   order. The test has been specified loosely like this to simplify
   implementation, and because it would not have been any more precise
   to have specified the first packet in sequence order, which would not
   necessarily be the first ACE counter that the Data Receiver fed back
   anyway, given it might have been a retransmission.

   The possibility of re-ordering means that there is a small chance
   that the ACE field on the first packet to arrive is genuinely zero
   (without middlebox interference). This would cause a host to
   unnecessarily disable ECN for a half connection. Therefore, in
   environments where there is no evidence of the ACE field being
   zeroed, implementations can skip this test.

   Note that the Data Sender MUST NOT test whether the arriving counter
   in the initial ACE field has been initialized to a specific valid
   value - the above check solely tests whether the ACE fields have been
   incorrectly zeroed. This allows hosts to use different initial
   values as an additional signalling channel in future.

3.2.2.5. Safety against Ambiguity of the ACE Field

   If too many CE-marked segments are acknowledged at once, or if a long
   run of ACKs is lost or thinned out, the 3-bit counter in the ACE
   field might have cycled between two ACKs arriving at the Data Sender.

The following safety procedures minimize this ambiguity.

3.2.2.5.1. Data Receiver Safety Procedures

The following rules define when a Data Receiver in AccECN mode emits an ACK:

Change-Triggered ACKs: An AccECN Data Receiver SHOULD emit an ACK whenever a data packet marked CE arrives after the previous packet was not CE.

Even though this rule is stated as a "SHOULD", it is important for a transition to trigger an ACK if at all possible, The only valid exception to this rule is given below these bullets.

For the avoidance of doubt, this rule is deliberately worded to apply solely when _data_ packets arrive, but the comparison with the previous packet includes any packet, not just data packets.

Increment-Triggered ACKs: An AccECN Data Receiver MUST emit an ACK if 'n' CE marks have arrived since the previous ACK. If there is newly delivered data to acknowledge, 'n' SHOULD be 2. If there is no newly delivered data to acknowledge, 'n' SHOULD be 3 and MUST be no less than 3. In either case, 'n' MUST be no greater than 7.

The above rules for when to send an ACK are designed to be complemented by those in Section 3.2.3.3, which concern whether the AccECN TCP Option ought to be included on ACKs.

If the arrivals of a number of data packets are all processed as one event, e.g. using large receive offload (LRO) or generic receive offload (GRO), both the above rules SHOULD be interpreted as requiring multiple ACKs to be emitted back-to-back (for each transition and for each repetition by 'n' CE marks). If this is problematic for high performance, either rule can be interpreted as requiring just a single ACK at the end of the whole receive event.

Even if a number of data packets do not arrive as one event, the 'Change-Triggered ACKs' rule could sometimes cause the ACK rate to be

problematic for high performance (although high performance protocols
such as DCTCP already successfully use change-triggered ACKs). The
rationale for change-triggered ACKs is so that the Data Sender can
rely on them to detect queue growth as soon as possible, particularly
at the start of a flow. The approach can lead to some additional
ACKs but it feeds back the timing and the order in which ECN marks
are received with minimal additional complexity. If CE marks are
infrequent, as is the case for most AQMs at the time of writing, or
there are multiple marks in a row, the additional load will be low.
However, marking patterns with numerous non-contiguous CE marks could
increase the load significantly. One possible compromise would be
for the receiver to heuristically detect whether the sender is in
slow-start, then to implement change-triggered ACKs while the sender
is in slow-start, and offload otherwise.

With ECN-capable pure ACKs [I-D.ietf-tcpm-generalized-ecn], the
'Increment-Triggered ACKs' rule could cause ECN-marked pure ACKs to
trigger further ACKs. Although TCP normally only ACKs newly
delivered data, in this case the ACKs of ACKs would feed back new
congestion state. The minimum of 3 for 'n' in this case ensures

that, even if there is pathological congestion in both directions,
any resulting ping-pong of ACKs will be rapidly damped.

These ACKs of ACKs could be misidentified as duplicate ACKs in
certain circumstances described below. Therefore, a host in AccECN
mode that is sending ECN-capable pure ACKs SHOULD add one of the
following additional checks when it tests whether an incoming pure
ACK is a duplicate:

o If SACK has been negotatiated for the connection, but there is no
   SACK option on the incoming pure ACK, it is not a duplicate;

o If timestamps are in use, and the incoming pure ACK echoes a
   timestamp older than the oldest unacknowledged data, it is not a
   duplicate.

In the unlikely event that neither SACK nor timestamps are in use, or
if the implementation has opted not to include either of the above
two checks, it SHOULD NOT send ECN-capable pure ACKs. If it does, it
could lead to false detection of duplicate ACKs, causing spurious
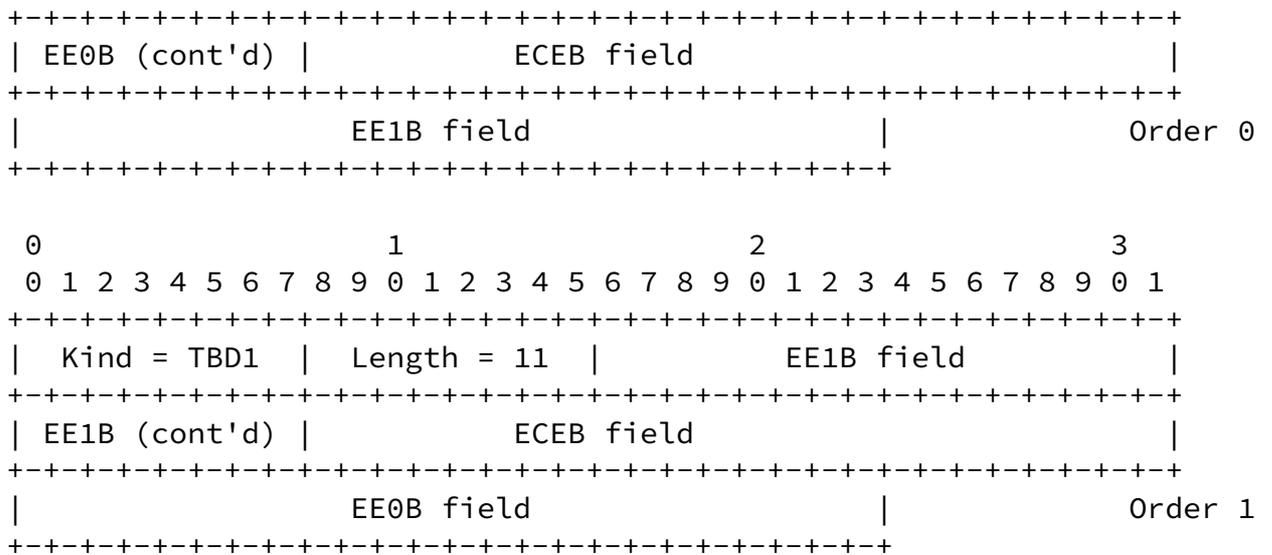retransmission(s) with a resulting unnecessary reduction in

congestion window; but only in certain circumstances. Specifically,
if TCP peer A has been sending data, then receiving, then within one
round trip it starts sending again, and the ECN-capable pure ACKs it
sent in the previous round encounter heavy enough congestion to
trigger peer B to invoke the above 'n'-CE-mark rule. Also note that
falsely considering these ACKs as duplicates would incorrectly imply
that data left the network.

3.2.2.5.2. Data Sender Safety Procedures

If the Data Sender has not received AccECN TCP Options to give it
more dependable information, and it detects that the ACE field could
have cycled, it SHOULD deem whether it cycled by taking the safest
likely case under the prevailing conditions. It can detect if the
counter could have cycled by using the jump in the acknowledgement
number since the last ACK to calculate or estimate how many segments
could have been acknowledged. An example algorithm to implement this
policy is given in Appendix A.2. An implementer MAY develop an
alternative algorithm as long as it satisfies these requirements.

If missing acknowledgement numbers arrive later (reordering) and
prove that the counter did not cycle, the Data Sender MAY attempt to
neutralize the effect of any action it took based on a conservative
assumption that it later found to be incorrect.

The Data Sender can estimate how many packets (of any marking) an ACK
acknowledges. If the ACE counter on an ACK seems to imply that the

minimum number of newly CE-marked packets is greater that the number
of newly acknowledged packets, the Data Sender SHOULD believe the ACE
counter, unless it can be sure that it is counting all control
packets correctly.

3.2.3. The AccECN Option

The AccECN Option is defined as shown in Figure 4. The initial 'E'
of each field name stands for 'Echo'.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Kind = TBD0  |  Length = 11  |          EE0B field           |
```

```
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     | EE0B (cont'd) |              ECEB field                       |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                     EE1B field                 |       Order 0
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  Kind = TBD1  |  Length = 11  |          EE1B field           |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     | EE1B (cont'd) |              ECEB field                       |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                     EE0B field                 |       Order 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                   Figure 4: The AccECN TCP Option

   Figure 4 shows two option field orders; order 0 and order 1. They
   both consists of three 24-bit fields. Order 0 provides the 24 least
   significant bits of the r.e0b, r.ceb and r.e1b counters,
   respectively. Order 1 provides the same fields, but in the opposite
   order. On each packet, the Data Receiver can use whichever order is
   more efficient.

   When a Data Receiver sends an AccECN Option, it MUST set the Kind
   field to TBD0 if using Order 0, or to TBD1 if using Order 1. These
   two new TCP Option Kinds are registered in Section 7 and called
   respectively AccECN0 and AccECN1.

   Note that there is no field to feed back Not-ECT bytes. Nonetheless
   an algorithm for the Data Sender to calculate the number of payload
   bytes received as Not-ECT is given in Appendix A.4.

   Whenever a Data Receiver sends an AccECN Option, the rules in
   Section 3.2.3.3 allow it to omit unchanged fields from the tail of
   the option, to help cope with option space limitations, as long as it
   preserves the order of the remaining fields and includes any field
   that has changed. The length field MUST indicate which fields are
   present as follows:

```
+--------+------------------+-------------------+
| Length | Type 0           | Type 1            |
+--------+------------------+-------------------+
| 11     | EE0B, ECEB, EE1B | EE1B, ECEB, EE0B  |
| 8      | EE0B, ECEB       | EE1B, ECEB        |
| 5      | EE0B             | EE1B              |
| 2      | (empty)          | (empty)           |
+--------+------------------+-------------------+
```

        Fields included in AccECN TCP Options of each length and type

   The empty option of Length=2 is provided to allow for a case where an
   AccECN Option has to be sent (e.g. on the SYN/ACK to test the path),
   but there is very limited space for the option.

   All implementations of a Data Sender that read any AccECN Option MUST
   be able to read in AccECN Options of any of the above lengths. For
   forward compatibility, if the AccECN Option is of any other length,
   implementations MUST use those whole 3-octet fields that fit within
   the length and ignore the remainder of the option, treating it as
   padding.

   The AccECN Option has to be optional to implement, because both
   sender and receiver have to be able to cope without the option anyway
   - in cases where it does not traverse a network path. It is
   RECOMMENDED to implement both sending and receiving of the AccECN
   Option. If sending of the AccECN Option is implemented, the fall-
   backs described in this document will need to be implemented as well
   (unless solely for a controlled environment where path traversal is
   not considered a problem). Even if a developer does not implement
   sending of the AccECN Option, it is RECOMMENDED that they still
   implement logic to receive and understand any AccECN Options sent by
   remote peers.

   If a Data Receiver intends to send the AccECN Option at any time
   during the rest of the connection it is strongly recommended to also
   test path traversal of the AccECN Option as specified in
   Section 3.2.3.2.

. Encoding and Decoding Feedback in the AccECN Option Fields

   Whenever the Data Receiver includes any of the counter fields (ECEB,
   EE0B, EE1B) in an AccECN Option, it MUST encode the 24 least
   significant bits of the current value of the associated counter into
   the field (respectively r.ceb, r.e0b, r.e1b).

   Whenever the Data Sender receives ACK carrying an AccECN Option, it
   first checks whether the ACK has already been superseded by another
   ACK in which case it ignores the ECN feedback. If the ACK has not
   been superseded, the Data Sender normally decodes the fields in the
   AccECN Option as follows. For each field, it takes the least
   significant 24 bits of its associated local counter (s.ceb, s.e0b or
   s.e1b) and subtracts them from the counter in the associated field of
   the incoming AccECN Option (respectively ECEB, EE0B, EE1B), to work
   out the minimum positive increment it could apply to s.ceb, s.e0b or
   s.e1b (assuming the field in the option only wrapped at most once).

   Appendix A.1 gives an example algorithm for the Data Receiver to
   encode its byte counters into the AccECN Option, and for the Data
   Sender to decode the AccECN Option fields into its byte counters.

   Note that, as specified in Section 3.2, any data on the SYN (SYN=1,
   ACK=0) is not included in any of the byte counters held locally for
   each ECN marking nor in the AccECN Option on the wire.

3.2.3.2. Path Traversal of the AccECN Option

3.2.3.2.1. Testing the AccECN Option during the Handshake

   The TCP client MUST NOT include the AccECN TCP Option on the SYN. If
   there is somehow an AccECN Option on a SYN, it MUST be ignored when
   forwarded or received. (A fall-back strategy for the loss of the
   SYN, possibly due to middlebox interference, is specified in
   Section 3.1.4.)

   A TCP server that confirms its support for AccECN (in response to an
   AccECN SYN from the client as described in Section 3.1) SHOULD
   include an AccECN TCP Option on the SYN/ACK.

   A TCP client that has successfully negotiated AccECN SHOULD include
   an AccECN Option in the first ACK at the end of the 3WHS. However,
   this first ACK is not delivered reliably, so the TCP client SHOULD
   also include an AccECN Option on the first data segment it sends (if
   it ever sends one).

   A host MAY omit the AccECN Option in any of the above three cases due
   to insufficient option space or if it has cached knowledge that the

packet would be likely to be blocked on the path to the other host if
it included an AccECN Option.

3.2.3.2.2. Testing for Loss of Packets Carrying the AccECN Option

If after the normal TCP timeout the TCP server has not received an
ACK to acknowledge its SYN/ACK, the SYN/ACK might just have been
lost, e.g. due to congestion, or a middlebox might be blocking the
AccECN Option. To expedite connection setup, the TCP server SHOULD
retransmit the SYN/ACK repeating the same AE, CWR and ECE TCP flags
as on the original SYN/ACK but with no AccECN Option. If this
retransmission times out, to expedite connection setup, the TCP
server SHOULD disable AccECN and ECN for this connection by
retransmitting the SYN/ACK with AE=CWR=ECE=0 and no AccECN Option.

Implementers MAY use other fall-back strategies if they are found to
be more effective (e.g. retrying the AccECN Option for a second time
before fall-back - most appropriate during high levels of
congestion). However, other fall-back strategies will need to follow
all the rules in Section 3.1.5, which concern behaviour when SYNs or
SYN/ACKs negotiating different types of feedback have been sent
within the same connection.

If the TCP client detects that the first data segment it sent with
the AccECN Option was lost, it SHOULD fall back to no AccECN Option
on the retransmission. Again, implementers MAY use other fall-back
strategies such as attempting to retransmit a second segment with the
AccECN Option before fall-back, and/or caching whether the AccECN
Option is blocked for subsequent connections. [RFC9040] further
discusses caching of TCP parameters and status information.

If a host falls back to not sending the AccECN Option, it will
continue to process any incoming AccECN Options as normal.

Either host MAY include the AccECN Option in a subsequent segment to
retest whether the AccECN Option can traverse the path.

If the TCP server receives a second SYN with a request for AccECN
support, it should resend the SYN/ACK, again confirming its support
for AccECN, but this time without the AccECN Option. This approach
rules out any interference by middleboxes that may drop packets with
unknown options, even though it is more likely that the SYN/ACK would
have been lost due to congestion. The TCP server MAY try to send

another packet with the AccECN Option at a later point during the
connection but should monitor if that packet got lost as well, in
which case it SHOULD disable the sending of the AccECN Option for
this half-connection.

Similarly, an AccECN end-point MAY separately memorize which data
packets carried an AccECN Option and disable the sending of AccECN
Options if the loss probability of those packets is significantly
higher than that of all other data packets in the same connection.

3.2.3.2.3. Testing for Absence of the AccECN Option

If the TCP client has successfully negotiated AccECN but does not
receive an AccECN Option on the SYN/ACK (e.g. because is has been
stripped by a middlebox or not sent by the server), the client
switches into a mode that assumes that the AccECN Option is not
available for this half connection.

Similarly, if the TCP server has successfully negotiated AccECN but
does not receive an AccECN Option on the first segment that
acknowledges sequence space at least covering the ISN, it switches
into a mode that assumes that the AccECN Option is not available for
this half connection.

While a host is in this mode that assumes incoming AccECN Options are
not available, it MUST adopt the conservative interpretation of the
ACE field discussed in Section 3.2.2.5. However, it cannot make any
assumption about support of outgoing AccECN Options on the other half
connection, so it SHOULD continue to send the AccECN Option itself
(unless it has established that sending the AccECN Option is causing
packets to be blocked as in Section 3.2.3.2.2).

If a host is in the mode that assumes incoming AccECN Options are not
available, but it receives an AccECN Option at any later point during
the connection, this clearly indicates that the AccECN Option is not
blocked on the respective path, and the AccECN endpoint MAY switch
out of the mode that assumes the AccECN Option is not available for
this half connection.

3.2.3.2.4. Test for Zeroing of the AccECN Option

For a related test for invalid initialization of the ACE field, see [Section 3.2.2.4](#)

[Section 3.2.1](#) required the Data Receiver to initialize the r.e0b and r.e1b counters to a non-zero value. Therefore, in either direction the initial value of the EE0B field or EE1B field in the AccECN Option (if one exists) ought to be non-zero. If AccECN has been negotiated:

o   the TCP server MAY check that the initial value of the EE0B field or the EE1B field is non-zero in the first segment that acknowledges sequence space that at least covers the ISN plus 1.

If it runs a test and either initial value is zero, the server will switch into a mode that ignores the AccECN Option for this half connection.

o   the TCP client MAY check the initial value of the EE0B field or the EE1B field is non-zero on the SYN/ACK. If it runs a test and either initial value is zero, the client will switch into a mode that ignores the AccECN Option for this half connection.

While a host is in the mode that ignores the AccECN Option it MUST adopt the conservative interpretation of the ACE field discussed in [Section 3.2.2.5](#).

Note that the Data Sender MUST NOT test whether the arriving byte counters in the initial AccECN Option have been initialized to specific valid values - the above checks solely test whether these fields have been incorrectly zeroed. This allows hosts to use different initial values as an additional signalling channel in future. Also note that the initial value of either field might be greater than its expected initial value, because the counters might already have been incremented. Nonetheless, the initial values of the counters have been chosen so that they cannot wrap to zero on these initial segments.

[3.2.3.2.5](#). Consistency between AccECN Feedback Fields

When the AccECN Option is available it ought to provide more unambiguous feedback. However, it supplements but does not replace the ACE field. An endpoint using AccECN feedback MUST always

reconcile the information provided in the ACE field with that in any
AccECN Option, so that the state of the ACE-related packet counter
can be relied on if future feedback does not carry the AccECN Option.

If the AccECN option is present, the s.cep counter might increase
more than expected from the increase of the s.ceb counter (e.g. due
to a CE-marked control packet). The sender's response to such a
situation is out of scope, and needs to be dealt with in a
specification that uses ECN-capable control packets. Theoretically,
this situation could also occur if a middlebox mangled the AccECN
Option but not the ACE field. However, the Data Sender has to assume
that the integrity of the AccECN Option is sound, based on the above
test of the well-known initial values and optionally other integrity
tests ([Section 5.3](#)).

If either end-point detects that the s.ceb counter has increased but
the s.cep has not (and by testing ACK coverage it is certain how much
the ACE field has wrapped), and if there is no explanation other than
an invalid protocol transition due to some form of feedback mangling,

the Data Sender MUST disable sending ECN-capable packets for the
remainder of the half-connection by setting the IP/ECN field in all
subsequent packets to Not-ECT.

[3.2.3.3](#). Usage of the AccECN TCP Option

If a Data Receiver in AccECN mode intends to use the AccECN TCP
Option to provide feedback, the rules below determine when it
includes an AccECN TCP Option, and which fields to include, given
other options might be competing for limited option space:

Importance of Congestion Control: AccECN is for congestion control,
   which SHOULD generally be considered important relative to other
   TCP options.

   If SACK has been negotiated, and the smallest recommended AccECN
   Option would leave insufficient space for two SACK blocks on a
   particular ACK, the Data Receiver MUST give precedence to the SACK
   option (total 18 octets), because loss feedback is more critical.

Recommended Simple Scheme: The Data Receiver SHOULD include an
   AccECN TCP Option on every scheduled ACK if any byte counter has

incremented since the last ACK. Whenever possible, it SHOULD
include a field for every byte counter that has changed at some
time during the connection (see examples later).

A scheduled ACK means an ACK that the Data Receiver would send by
its regular delayed ACK rules. Recall that [Section 1.3](#) defines an
'ACK' as either with data payload or without. But the above rule
is worded so that, in the common case when most of the data is
from a server to a client, the server only includes an AccECN TCP
Option while it is acknowledging data from the client.

When available TCP option space is limited on particular packets, the
recommended scheme will need to include compromises. To guide the
implementer the rules below are ranked in order of importance, but
the final decision has to be implementation-dependent, because
tradeoffs will alter as new TCP options are defined and new use-cases
arise.

Necessary Option Length: The Data Receiver MUST only include an
    AccECN TCP Option on a packet if it includes all the counter(s)
    that have incremented since the previous AccECN Option. It MUST
    only truncate unchanged fields from the right-hand tail of the
    option to preserve the order of the remaining fields (see
    [Section 3.2.3](#));

Change-Triggered AccECN TCP Options: If an arriving packet
    increments a different byte counter to that incremented by the
    previous packet, the Data Receiver SHOULD feed it back in an
    AccECN Option on the next scheduled ACK.

    For the avoidance of doubt, this rule does not concern the arrival
    of control packets with no payload, because they cannot alter any
    byte counters.

Continual Repetition: Otherwise, if arriving packets continue to
    increment the same byte counter:

    *   the Data Receiver SHOULD include a counter that has continued
        to increment on the next scheduled ACK following a change-
        triggered AccECN TCP Option;

* while the same counter continues to increment, it SHOULD
          include the counter every n ACKs as consistently as possible,
          where n can be chosen by the implementer;

        * It SHOULD always include an AccECN Option if the r.ceb counter
          is incrementing and it MAY include an AccECN Option if r.ec0b
          or r.ec1b is incrementing

        * It SHOULD, include each counter at least once for every 2^22
          bytes incremented to prevent overflow during continual
          repetition.

   The above rules complement those in [Section 3.2.2.5](#), which determine
   when to generate an ACK irrespective of whether an AccECN TCP Option
   is to be included.

   The recommended scheme is intended as a simple way to ensure that all
   the relevant byte counters will be carried on any ACK that reaches
   the Data Sender, no matter how many pure ACKs are filtered or
   coalesced along the network path, and without consuming the space
   available for payload data with counter field(s) that have never
   changed.

   As an example of the recommended scheme, if ECT(0) is the only
   codepoint that has ever arrived in the IP-ECN field, the Data
   Receiver will feed back an AccECN0 TCP Option with only the EE0B
   field on every packet. However, as soon as even one CE-marked packet
   arrives, on every packet that acknowledges new data it will start to
   include an option with two fields, EE0B and ECEB. As a second
   example, if the first packet to arrive happens to be CE-marked, the
   Data Receiver will have to arbitrarily choose whether to precede the
   ECEB field with an EE0B field or an EE1B field. If it chooses, say,

   EEB0 but it turns out never to receive ECT(0), it can start sending
   EE1B and ECEB instead - it does not have to include the EE0B field if
   the r.e0b counter has never changed during the connection.

   With the recommended scheme, if the data sending direction switches
   during a connection, there can be cases where the AccECN TCP Option
   that is meant to feed back the counter values at the end of a volley
   in one direction never reaches the other peer, due to packet loss.

ACE feedback ought to be sufficient to fill this gap, given accurate
feedback becomes moot after data transmission has paused.

[Appendix A.3](#) gives an example algorithm to estimate the number of
marked bytes from the ACE field alone, if the AccECN Option is not
available.

If a host has determined that segments with the AccECN Option always
seem to be discarded somewhere along the path, it is no longer
obliged to follow any of the rules in this section.

3.3. AccECN Compliance Requirements for TCP Proxies, Offload Engines and other Middleboxes

3.3.1. Requirements for TCP Proxies

A large class of middleboxes split TCP connections. Such a middlebox
would be compliant with the AccECN protocol if the TCP implementation
on each side complied with the present AccECN specification and each
side negotiated AccECN independently of the other side.

3.3.2. Requirements for Transparent Middleboxes and TCP Normalizers

Another large class of middleboxes intervenes to some degree at the
transport layer, but attempts to be transparent (invisible) to the
end-to-end connection. A subset of this class of middleboxes
attempts to `normalize' the TCP wire protocol by checking that all
values in header fields comply with a rather narrow interpretation of
the TCP specifications that is also not always up to date.

A middlebox that is not normalizing the TCP protocol and does not
itself act as a back-to-back pair of TCP endpoints (i.e. a middlebox
that intends to be transparent or invisible at the transport layer)
ought to forward the AccECN TCP Option unaltered, whether or not the
length value matches one of those specified in [Section 3.2.3](#), and
whether or not the initial values of the byte-counter fields match
those in [Section 3.2.1](#). This is because blocking apparently invalid
values prevents the standardized set of values being extended in
future (given outdated normalizers would block updated hosts from
using the extended AccECN standard).

A TCP normalizer is likely to block or alter an AccECN TCP Option if

the length value or the initial values of its byte-counter fields do not match one of those specified in Section 3.2.3 or Section 3.2.1. However, to comply with the present AccECN specification, a middlebox MUST NOT change the ACE field; or those fields of the AccECN Option that are currently specified in Section 3.2.3; or any AccECN field covered by integrity protection (e.g. [RFC5925]).

3.3.3. Requirements for TCP ACK Filtering

   A node that implements ACK filtering (aka. thinning or coalescing)
   SHOULD determine if an ACK is part of a connection using AccECN and
   SHOULD then preserve the correct operation of AccECN feedback. The
   following notes might help with each part of this requirement:

   o To determine whether a pure TCP ACK is part of an AccECN
      connection without resorting to connection tracking and per-flow
      state, a useful heuristic would be to check for a non-zero ECN
      field at the IP layer (because the ECN++ experiment only allows
      TCP pure ACKs to be ECN-capable if AccECN has been negotiated
      [I-D.ietf-tcpm-generalized-ecn]). This heuristic is simple and
      stateless. However, it might omit some AccECN ACKs, because it is
      only recommended but not obligatory to use ECN++ with AccECN -
      only deployment experience will tell. Also, TCP ACKs might be
      ECN-capable owing to some scheme other than AccECN, e.g. [RFC5690]
      or some future standards action. Again, only deployment
      experience will tell.

   o The main concern with preserving correct AccECN operation involves
      leaving enough ACKs for the data sender to work out whether the
      3-bit ACE field has wrapped. ACE field wrap is of less concern if
      packets also carry the AccECN TCP Option.

   Note that the present specification of AccECN in TCP does not presume
   to rely on any of the above ACK filtering behaviour in the network
   (hence the use of 'SHOULD' rather than 'MUST' above), because it has
   to be robust against pre-existing network nodes that do not
   distinguish AccECN ACKs, and robust against ACK loss during overload
   more generally.

   Section 5.2.1 of BCP 69 [RFC3449] gives best current practice on pure
   TCP ACK filtering. It gives no advice on ACKs carrying ECN feedback,
   other than that filtering ought to preserve the correct operation of
   ECN feedback, because at the time it said that "SACK and ECN remain
   areas of ongoing research". This section updates that best current
   practice for a TCP connection that supports AccECN feedback.

3.3.4. Requirements for TCP Segmentation Offload

   Hardware to offload certain TCP processing represents another large
   class of middleboxes (even though it is often a function of a host's
   network interface and rarely in its own 'box').

   The ACE field changes with every received CE marking, so today's
   receive offloading could lead to many interrupts in high congestion
   situations. Although that would be useful (because congestion
   information is received sooner), it could also significantly increase
   processor load, particularly in scenarios such as DCTCP or L4S where
   the marking rate is generally higher.

   Current offload hardware ejects a segment from the coalescing process
   whenever the TCP ECN flags change. Thus Classic ECN causes offload
   to be inefficient. In data centres it has been fortunate for this
   offload hardware that DCTCP-style feedback changes less often when
   there are long sequences of CE marks, which is more common with a
   step marking threshold (but less likely the more short flows are in
   the mix). The ACE counter approach has been designed so that
   coalescing can continue over arbitrary patterns of marking and only
   needs to stop when the counter wraps. Nonetheless, until the
   particular offload hardware in use implements this more efficient
   approach, it is likely to be more efficient for AccECN connections to
   implement this counter-style logic using software segmentation
   offload.

   ECN encodes a varying signal in the ACK stream, so it is inevitable
   that offload hardware will ultimately need to handle any form of ECN
   feedback exceptionally. The ACE field has been designed as a counter
   so that it is straightforward for offload hardware to pass on the
   highest counter, and to push a segment from its cache before the
   counter wraps. The purpose of working towards standardized TCP ECN
   feedback is to reduce the risk for hardware developers, who would
   otherwise have to guess which scheme is likely to become dominant.

   The above process has been designed to enable a continuing
   incremental deployment path - to more highly dynamic congestion
   control. Once offload hardware supports AccECN, it will be able to
   coalesce efficiently for any sequence of marks, instead of relying
   for efficiency on the long marking sequences from step marking. In
   the next stage, marking can evolve from a step to a ramp function.
   That in turn will allow host congestion control algorithms to respond
   faster to dynamics, while being backwards compatible with existing
   host algorithms.

4. Updates to RFC 3168

   Normative statements in the following sections of RFC3168 are updated
   by the present AccECN specification:

   o The whole of "6.1.1 TCP Initialization" of [RFC3168] is updated by
      Section 3.1 of the present specification.

   o In "6.1.2. The TCP Sender" of [RFC3168], all mentions of a
      congestion response to an ECN-Echo (ECE) ACK packet are updated by
      Section 3.2 of the present specification to mean an increment to
      the sender's count of CE-marked packets, s.cep. And the
      requirements to set the CWR flag no longer apply, as specified in
      Section 3.1.5 of the present specification. Otherwise, the
      remaining requirements in "6.1.2. The TCP Sender" still stand.

      It will be noted that RFC 8311 already updates, or potentially
      updates, a number of the requirements in "6.1.2. The TCP Sender".
      Section 6.1.2 of RFC 3168 extended standard TCP congestion control
      [RFC5681] to cover ECN marking as well as packet drop. Whereas,
      RFC 8311 enables experimentation with alternative responses to ECN
      marking, if specified for instance by an experimental RFC on the
      IETF document stream. RFC 8311 also strengthened the statement
      that "ECT(0) SHOULD be used" to a "MUST" (see [RFC8311] for the
      details).

   o The whole of "6.1.3. The TCP Receiver" of [RFC3168] is updated by
      Section 3.2 of the present specification, with the exception of
      the last paragraph (about congestion response to drop and ECN in
      the same round trip), which still stands. Incidentally, this last
      paragraph is in the wrong section, because it relates to TCP
      sender behaviour.

   o The following text within "6.1.5. Retransmitted TCP packets":

         "the TCP data receiver SHOULD ignore the ECN field on arriving
         data packets that are outside of the receiver's current
         window."

is updated by more stringent acceptability tests for any packet
(not just data packets) in the present specification.
Specifically, in the normative specification of AccECN (Section 3)
only 'Acceptable' packets contribute to the ECN counters at the
AccECN receiver and Section 1.3 defines an Acceptable packet as
one that passes the acceptability tests in both [RFC0793] and
[RFC5961].

   o Sections 5.2, 6.1.1, 6.1.4, 6.1.5 and 6.1.6 of [RFC3168] prohibit
      use of ECN on TCP control packets and retransmissions. The
      present specification does not update that aspect of RFC 3168, but
      it does say what feedback an AccECN Data Receiver should provide
      if it receives an ECN-capable control packet or retransmission.
      This ensures AccECN is forward compatible with any future scheme
      that allows ECN on these packets, as provided for in section 4.3
      of [RFC8311] and as proposed in [I-D.ietf-tcpm-generalized-ecn].

5. Interaction with TCP Variants

   This section is informative, not normative.

5.1. Compatibility with SYN Cookies

   A TCP server can use SYN Cookies (see Appendix A of [RFC4987]) to
   protect itself from SYN flooding attacks. It places minimal commonly
   used connection state in the SYN/ACK, and deliberately does not hold
   any state while waiting for the subsequent ACK (e.g. it closes the
   thread). Therefore it cannot record the fact that it entered AccECN
   mode for both half-connections. Indeed, it cannot even remember
   whether it negotiated the use of classic ECN [RFC3168].

   Nonetheless, such a server can determine that it negotiated AccECN as
   follows. If a TCP server using SYN Cookies supports AccECN and if it
   receives a pure ACK that acknowledges an ISN that is a valid SYN
   cookie, and if the ACK contains an ACE field with the value 0b010 to
   0b111 (decimal 2 to 7), it can assume that:

   o the TCP client must have requested AccECN support on the SYN

   o it (the server) must have confirmed that it supported AccECN

Therefore the server can switch itself into AccECN mode, and continue as if it had never forgotten that it switched itself into AccECN mode earlier.

If the pure ACK that acknowledges a SYN cookie contains an ACE field with the value 0b000 or 0b001, these values indicate that the client did not request support for AccECN and therefore the server does not enter AccECN mode for this connection. Further, 0b001 on the ACK implies that the server sent an ECN-capable SYN/ACK, which was marked CE in the network, and the non-AccECN client fed this back by setting ECE on the ACK of the SYN/ACK.

5.2. Compatibility with TCP Experiments and Common TCP Options

   AccECN is compatible (at least on paper) with the most commonly used TCP options: MSS, time-stamp, window scaling, SACK and TCP-AO. It is also compatible with the recent promising experimental TCP options TCP Fast Open (TFO [RFC7413]) and Multipath TCP (MPTCP [RFC6824]). AccECN is friendly to all these protocols, because space for TCP options is particularly scarce on the SYN, where AccECN consumes zero additional header space.

   When option space is under pressure from other options, Section 3.2.3.3 provides guidance on how important it is to send an AccECN Option relative to other options, and which fields are more important to include.

   Implementers of TFO need to take careful note of the recommendation in Section 3.2.2.1. That section recommends that, if the client has successfully negotiated AccECN, when acknowledging the SYN/ACK, even if it has data to send, it sends a pure ACK immediately before the data. Then it can reflect the IP-ECN field of the SYN/ACK on this pure ACK, which allows the server to detect ECN mangling. Note that, as specified in Section 3.2, any data on the SYN (SYN=1, ACK=0) is not included in any of the byte counters held locally for each ECN marking, nor in the AccECN Option on the wire.

. Compatibility with Feedback Integrity Mechanisms

   Three alternative mechanisms are available to assure the integrity of
   ECN and/or loss signals. AccECN is compatible with any of these
   approaches:

   o The Data Sender can test the integrity of the receiver's ECN (or
      loss) feedback by occasionally setting the IP-ECN field to a value
      normally only set by the network (and/or deliberately leaving a
      sequence number gap). Then it can test whether the Data
      Receiver's feedback faithfully reports what it expects (similar to
      para 2 of Section 20.2 of [RFC3168]). Unlike the ECN Nonce
      [RFC3540], this approach does not waste the ECT(1) codepoint in
      the IP header, it does not require standardization and it does not
      rely on misbehaving receivers volunteering to reveal feedback
      information that allows them to be detected. However, setting the
      CE mark by the sender might conceal actual congestion feedback
      from the network and should therefore only be done sparingly.

   o Networks generate congestion signals when they are becoming
      congested, so networks are more likely than Data Senders to be
      concerned about the integrity of the receiver's feedback of these
      signals. A network can enforce a congestion response to its ECN

      markings (or packet losses) using congestion exposure (ConEx)
      audit [RFC7713]. Whether the receiver or a downstream network is
      suppressing congestion feedback or the sender is unresponsive to
      the feedback, or both, ConEx audit can neutralize any advantage
      that any of these three parties would otherwise gain.

      ConEx is an experimental change to the Data Sender that would be
      most useful when combined with AccECN. Without AccECN, the ConEx
      behaviour of a Data Sender would have to be more conservative than
      would be necessary if it had the accurate feedback of AccECN.

   o The standards track TCP authentication option (TCP-AO [RFC5925])
      can be used to detect any tampering with AccECN feedback between
      the Data Receiver and the Data Sender (whether malicious or
      accidental). The AccECN fields are immutable end-to-end, so they
      are amenable to TCP-AO protection, which covers TCP options by
      default. However, TCP-AO is often too brittle to use on many end-
      to-end paths, where middleboxes can make verification fail in

their attempts to improve performance or security, e.g. by
resegmentation or shifting the sequence space.

Originally the ECN Nonce [RFC3540] was proposed to ensure integrity
of congestion feedback. With minor changes AccECN could be optimized
for the possibility that the ECT(1) codepoint might be used as an ECN
Nonce. However, given RFC 3540 has been reclassified as historic,
the AccECN design has been generalized so that it ought to be able to
support other possible uses of the ECT(1) codepoint, such as a lower
severity or a more instant congestion signal than CE.

6. Protocol Properties

This section is informative not normative. It describes how well the
protocol satisfies the agreed requirements for a more accurate ECN
feedback protocol [RFC7560].

Accuracy: From each ACK, the Data Sender can infer the number of new
   CE marked segments since the previous ACK. This provides better
   accuracy on CE feedback than classic ECN. In addition if the
   AccECN Option is present (not blocked by the network path) the
   number of bytes marked with CE, ECT(1) and ECT(0) are provided.

Overhead: The AccECN scheme is divided into two parts. The
   essential part reuses the 3 flags already assigned to ECN in the
   IP header. The supplementary part adds an additional TCP option
   consuming up to 11 bytes. However, no TCP option is consumed in
   the SYN.

Ordering: The order in which marks arrive at the Data Receiver is
   preserved in AccECN feedback, because the Data Receiver is
   expected to send an ACK immediately whenever a different mark
   arrives.

Timeliness: While the same ECN markings are arriving continually at
   the Data Receiver, it can defer ACKs as TCP does normally, but it
   will immediately send an ACK as soon as a different ECN marking
   arrives.

Timeliness vs Overhead: Change-Triggered ACKs are intended to enable

latency-sensitive uses of ECN feedback by capturing the timing of
transitions but not wasting resources while the state of the
signalling system is stable. Within the constraints of the
change-triggered ACK rules, the receiver can control how
frequently it sends the AccECN TCP Option and therefore to some
extent it can control the overhead induced by AccECN.

Resilience: All information is provided based on counters.
   Therefore if ACKs are lost, the counters on the first ACK
   following the losses allows the Data Sender to immediately recover
   the number of the ECN markings that it missed. And if data or
   ACKs are reordered, stale congestion information can be identified
   and ignored.

Resilience against Bias: Because feedback is based on repetition of
   counters, random losses do not remove any information, they only
   delay it. Therefore, even though some ACKs are change-triggered,
   random losses will not alter the proportions of the different ECN
   markings in the feedback.

Resilience vs Overhead: If space is limited in some segments
   (e.g. because more options are needed on some segments, such as
   the SACK option after loss), the Data Receiver can send AccECN
   Options less frequently or truncate fields that have not changed,
   usually down to as little as 5 bytes. However, it has to send a
   full-sized AccECN Option at least three times per RTT, which the
   Data Sender can rely on as a regular beacon or checkpoint.

Resilience vs Timeliness and Ordering: Ordering information and the
   timing of transitions cannot be communicated in three cases: i)
   during ACK loss; ii) if something on the path strips the AccECN
   Option; or iii) if the Data Receiver is unable to support Change-
   Triggered ACKs. Following ACK reordering, the Data Sender can
   reconstruct the order in which feedback was sent, but not until
   all the missing feedback has arrived.

Complexity: An AccECN implementation solely involves simple counter
   increments, some modulo arithmetic to communicate the least
   significant bits and allow for wrap, and some heuristics for
   safety against fields cycling due to prolonged periods of ACK

loss. Each host needs to maintain eight additional counters. The
hosts have to apply some additional tests to detect tampering by
middleboxes, but in general the protocol is simple to understand,
simple to implement and requires few cycles per packet to execute.

Integrity: AccECN is compatible with at least three approaches that
can assure the integrity of ECN feedback. If the AccECN Option is
stripped the resolution of the feedback is degraded, but the
integrity of this degraded feedback can still be assured.

Backward Compatibility: If only one endpoint supports the AccECN
scheme, it will fall-back to the most advanced ECN feedback scheme
supported by the other end.

Backward Compatibility: If the AccECN Option is stripped by a
middlebox, AccECN still provides basic congestion feedback in the
ACE field. Further, AccECN can be used to detect mangling of the
IP ECN field; mangling of the TCP ECN flags; blocking of ECT-
marked segments; and blocking of segments carrying the AccECN
Option. It can detect these conditions during TCP's 3WHS so that
it can fall back to operation without ECN and/or operation without
the AccECN Option.

Forward Compatibility: The behaviour of endpoints and middleboxes is
carefully defined for all reserved or currently unused codepoints
in the scheme. Then, the designers of security devices can
understand which currently unused values might appear in future.
So, even if they choose to treat such values as anomalous while
they are not widely used, any blocking will at least be under
policy control not hard-coded. Then, if previously unused values
start to appear on the Internet (or in standards), such policies
could be quickly reversed.

7. IANA Considerations

   This document reassigns bit 7 of the TCP header flags to the AccECN
   protocol. This bit was previously called the Nonce Sum (NS) flag
   [RFC3540], but RFC 3540 has been reclassified as historic [RFC8311].
   The flag will now be defined as:

```
            +-----+------------------+-----------+
            | Bit | Name             | Reference |
            +-----+------------------+-----------+
            | 7   | AE (Accurate ECN) | RFC XXXX |
            +-----+------------------+-----------+
```

                     TCP header flag reassignment

   [TO BE REMOVED: IANA is requested to update the existing entry in the
   Transmission Control Protocol (TCP) Header Flags registration
   (https://www.iana.org/assignments/tcp-header-flags/tcp-header-
   flags.xhtml#tcp-header-flags-1) for Bit 7 to "AE (Accurate ECN),
   previously used as NS (Nonce Sum) by [RFC3540], which is now Historic
   [RFC8311]" and change the reference to this RFC-to-be instead of
   RFC8311.]

   This document also defines two new TCP options for AccECN, assigned
   values of TBD0 and TBD1 (decimal) from the TCP option space. These
   values are defined as:

```
    +------+--------+------------------------------+-----------+
    | Kind | Length | Meaning                      | Reference |
    +------+--------+------------------------------+-----------+
    | TBD0 | N      | Accurate ECN Order 0 (AccECN0) | RFC XXXX |
    | TBD1 | N      | Accurate ECN Order 1 (AccECN1) | RFC XXXX |
    +------+--------+------------------------------+-----------+
```

                     New TCP Option assignments

   [TO BE REMOVED: This registration should take place at the following
   location: http://www.iana.org/assignments/tcp-parameters/tcp-
   parameters.xhtml#tcp-parameters-1]

   Early implementations using experimental option 254 per [RFC6994]
   with the single magic number 0xACCE (16 bits), as allocated in the
   IANA "TCP Experimental Option Experiment Identifiers (TCP ExIDs)"
   registry, SHOULD migrate to use these new option kinds (TBD0 & TBD1).

   [TO BE REMOVED: The description of the 0xACCE value in the TCP ExIDs
   registry should be changed to "AccECN (current and new
   implementations SHOULD use option kinds TBD0 and TBD1)" at the
   following location: https://www.iana.org/assignments/tcp-parameters/
   tcp-parameters.xhtml#tcp-exids]

8. Security Considerations

   If ever the supplementary part of AccECN based on the new AccECN TCP
   Option is unusable (due for example to middlebox interference) the
   essential part of AccECN's congestion feedback offers only limited
   resilience to long runs of ACK loss (see Section 3.2.2.5). These
   problems are unlikely to be due to malicious intervention (because if
   an attacker could strip a TCP option or discard a long run of ACKs it
   could wreak other arbitrary havoc). However, it would be of concern
   if AccECN's resilience could be indirectly compromised during a
   flooding attack. AccECN is still considered safe though, because if
   the option is not present, the AccECN Data Sender is then required to
   switch to more conservative assumptions about wrap of congestion
   indication counters (see Section 3.2.2.5 and Appendix A.2).

   Section 5.1 describes how a TCP server can negotiate AccECN and use
   the SYN cookie method for mitigating SYN flooding attacks.

   There is concern that ECN feedback could be altered or suppressed,
   particularly because a misbehaving Data Receiver could increase its
   own throughput at the expense of others. AccECN is compatible with
   the three schemes known to assure the integrity of ECN feedback (see
   Section 5.3 for details). If the AccECN Option is stripped by an
   incorrectly implemented middlebox, the resolution of the feedback
   will be degraded, but the integrity of this degraded information can
   still be assured. Assuring that Data Senders respond appropriately
   to ECN feedback is possible, but the scope of the present document is
   confined to the feedback protocol, and excludes the response to this
   feedback.

   In Section 3.2.3 a Data Sender is allowed to ignore an unrecognized
   TCP AccECN Option length and read as many whole 3-octet fields from
   it as possible up to a maximum of 3, treating the remainder as
   padding. This opens up a potential covert channel of up to 29B (40 -
   (2+3*3))B. However, it is really an overt channel (not hidden) and
   it is no different to the use of unknown TCP options with unknown
   option lengths in general. Therefore, where this is of concern, it
   can already be adequately mitigated by regular TCP normalizer
   technology (see Section 3.3.2).

The AccECN protocol is not believed to introduce any new privacy
concerns, because it merely counts and feeds back signals at the
transport layer that had already been visible at the IP layer. A
covert channel can be used to compromise privacy. However, as
explained above, undefined TCP options in general open up such
channels and common techniques are available to close them off.

There is a potential concern that a Data Receiver could deliberately
omit the AccECN Option pretending that it had been stripped by a
middlebox. No known way can yet be contrived for a receiver to take
advantage of this behaviour, which seems to always degrade its own
performance. However, the concern is mentioned here for
completeness.

9. Acknowledgements

We want to thank Koen De Schepper, Praveen Balasubramanian, Michael
Welzl, Gorry Fairhurst, David Black, Spencer Dawkins, Michael Scharf,
Michael Tuexen, Yuchung Cheng, Kenjiro Cho, Olivier Tilmans, Ilpo
Jaervinen, Neal Cardwell, Yoshifumi Nishida, Martin Duke and Jonathan
Morton for their input and discussion. The idea of using the three
ECN-related TCP flags as one field for more accurate TCP-ECN feedback
was first introduced in the re-ECN protocol that was the ancestor of
ConEx.

10. Comments Solicited

Comments and questions are encouraged and very welcome. They can be
addressed to the IETF TCP maintenance and minor modifications working
group mailing list <tcpm@ietf.org>, and/or to the authors.

11. References

11.1. Normative References

   [RFC0793] Postel, J., "Transmission Control Protocol", STD 7,
              RFC 793, DOI 10.17487/RFC0793, September 1981,
              <https://www.rfc-editor.org/info/rfc793>.

   [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>.

   [RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion
              Control", RFC 5681, DOI 10.17487/RFC5681, September 2009,
              <https://www.rfc-editor.org/info/rfc5681>.

   [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>.

11.2. Informative References

   [I-D.ietf-tcpm-generalized-ecn]
              Bagnulo, M. and B. Briscoe, "ECN++: Adding Explicit
              Congestion Notification (ECN) to TCP Control Packets",
              draft-ietf-tcpm-generalized-ecn-09 (work in progress),
              January 2022.

[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", draft-ietf-tsvwg-l4s-arch-17 (work
          in progress), March 2022.

[Mandalari18]
          Mandalari, A., Lutu, A., Briscoe, B., Bagnulo, M., and Oe.
          Alay, "Measuring ECN++: Good News for ++, Bad News for ECN
          over Mobile", IEEE Communications Magazine , March 2018.

[RFC2018] Mathis, M., Mahdavi, J., Floyd, S., and A. Romanow, "TCP
          Selective Acknowledgment Options", RFC 2018,
          DOI 10.17487/RFC2018, October 1996,
          <https://www.rfc-editor.org/info/rfc2018>.

[RFC3449] Balakrishnan, H., Padmanabhan, V., Fairhurst, G., and M.
          Sooriyabandara, "TCP Performance Implications of Network
          Path Asymmetry", BCP 69, RFC 3449, DOI 10.17487/RFC3449,
          December 2002, <https://www.rfc-editor.org/info/rfc3449>.

[RFC3540] Spring, N., Wetherall, D., and D. Ely, "Robust Explicit
          Congestion Notification (ECN) Signaling with Nonces",
          RFC 3540, DOI 10.17487/RFC3540, June 2003,
          <https://www.rfc-editor.org/info/rfc3540>.

[RFC4987] Eddy, W., "TCP SYN Flooding Attacks and Common
          Mitigations", RFC 4987, DOI 10.17487/RFC4987, August 2007,
          <https://www.rfc-editor.org/info/rfc4987>.

[RFC5562] Kuzmanovic, A., Mondal, A., Floyd, S., and K.
          Ramakrishnan, "Adding Explicit Congestion Notification
          (ECN) Capability to TCP's SYN/ACK Packets", RFC 5562,
          DOI 10.17487/RFC5562, June 2009,
          <https://www.rfc-editor.org/info/rfc5562>.

[RFC5690] Floyd, S., Arcia, A., Ros, D., and J. Iyengar, "Adding
          Acknowledgement Congestion Control to TCP", RFC 5690,
          DOI 10.17487/RFC5690, February 2010,
          <https://www.rfc-editor.org/info/rfc5690>.

   [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP
              Authentication Option", RFC 5925, DOI 10.17487/RFC5925,
              June 2010, <https://www.rfc-editor.org/info/rfc5925>.

   [RFC5961] Ramaiah, A., Stewart, R., and M. Dalal, "Improving TCP's
              Robustness to Blind In-Window Attacks", RFC 5961,
              DOI 10.17487/RFC5961, August 2010,
              <https://www.rfc-editor.org/info/rfc5961>.

   [RFC6824] Ford, A., Raiciu, C., Handley, M., and O. Bonaventure,
              "TCP Extensions for Multipath Operation with Multiple
              Addresses", RFC 6824, DOI 10.17487/RFC6824, January 2013,
              <https://www.rfc-editor.org/info/rfc6824>.

   [RFC6994] Touch, J., "Shared Use of Experimental TCP Options",
              RFC 6994, DOI 10.17487/RFC6994, August 2013,
              <https://www.rfc-editor.org/info/rfc6994>.

   [RFC7413] Cheng, Y., Chu, J., Radhakrishnan, S., and A. Jain, "TCP
              Fast Open", RFC 7413, DOI 10.17487/RFC7413, December 2014,
              <https://www.rfc-editor.org/info/rfc7413>.

   [RFC7560] Kuehlewind, M., Ed., Scheffenegger, R., and B. Briscoe,
              "Problem Statement and Requirements for Increased Accuracy
              in Explicit Congestion Notification (ECN) Feedback",
              RFC 7560, DOI 10.17487/RFC7560, August 2015,
              <https://www.rfc-editor.org/info/rfc7560>.

   [RFC7713] Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx)
              Concepts, Abstract Mechanism, and Requirements", RFC 7713,
              DOI 10.17487/RFC7713, December 2015,
              <https://www.rfc-editor.org/info/rfc7713>.

   [RFC8257] Bensley, S., Thaler, D., Balasubramanian, P., Eggert, L.,
              and G. Judd, "Data Center TCP (DCTCP): TCP Congestion
              Control for Data Centers", RFC 8257, DOI 10.17487/RFC8257,
              October 2017, <https://www.rfc-editor.org/info/rfc8257>.

   [RFC8311] Black, D., "Relaxing Restrictions on Explicit Congestion
              Notification (ECN) Experimentation", RFC 8311,

                 DOI 10.17487/RFC8311, January 2018,
                 <https://www.rfc-editor.org/info/rfc8311>.

   [RFC8511]   Khademi, N., Welzl, M., Armitage, G., and G. Fairhurst,
               "TCP Alternative Backoff with ECN (ABE)", RFC 8511,
               DOI 10.17487/RFC8511, December 2018,
               <https://www.rfc-editor.org/info/rfc8511>.

   [RFC9040]   Touch, J., Welzl, M., and S. Islam, "TCP Control Block
               Interdependence", RFC 9040, DOI 10.17487/RFC9040, July
               2021, <https://www.rfc-editor.org/info/rfc9040>.

Appendix A. Example Algorithms

   This appendix is informative, not normative. It gives example
   algorithms that would satisfy the normative requirements of the
   AccECN protocol. However, implementers are free to choose other ways

to implement the requirements.

A.1. Example Algorithm to Encode/Decode the AccECN Option

   The example algorithms below show how a Data Receiver in AccECN mode
   could encode its CE byte counter r.ceb into the ECEB field within the
   AccECN TCP Option, and how a Data Sender in AccECN mode could decode
   the ECEB field into its byte counter s.ceb. The other counters for
   bytes marked ECT(0) and ECT(1) in the AccECN Option would be
   similarly encoded and decoded.

   It is assumed that each local byte counter is an unsigned integer
   greater than 24b (probably 32b), and that the following constant has
   been assigned:

      DIVOPT = 2^24

   Every time a CE marked data segment arrives, the Data Receiver
   increments its local value of r.ceb by the size of the TCP Data.
   Whenever it sends an ACK with the AccECN Option, the value it writes
   into the ECEB field is

      ECEB = r.ceb % DIVOPT

   where '%' is the remainder operator.

   On the arrival of an AccECN Option, the Data Sender first makes sure
   the ACK has not been superseded in order to avoid winding the s.ceb
   counter backwards. It uses the TCP acknowledgement number and any
   SACK options to calculate newlyAckedB, the amount of new data that
   the ACK acknowledges in bytes (newlyAckedB can be zero but not
   negative). If newlyAckedB is zero, either the ACK has been
   superseded or CE-marked packet(s) without data could have arrived.
   To break the tie for the latter case, the Data Sender could use
   timestamps (if present) to work out newlyAckedT, the amount of new
   time that the ACK acknowledges. If the Data Sender determines that
   the ACK has been superseded it ignores the AccECN Option. Otherwise,
   the Data Sender calculates the minimum non-negative difference d.ceb
   between the ECEB field and its local s.ceb counter, using modulo
   arithmetic as follows:

```
    if ((newlyAckedB > 0) || (newlyAckedT > 0)) {
        d.ceb = (ECEB + DIVOPT - (s.ceb % DIVOPT)) % DIVOPT
        s.ceb += d.ceb
    }
```

For example, if s.ceb is 33,554,433 and ECEB is 1461 (both decimal),
then

```
    s.ceb % DIVOPT = 1
        d.ceb = (1461 + 2^24 - 1) % 2^24
              = 1460
        s.ceb = 33,554,433 + 1460
              = 33,555,893
```

In practice an implementation might use heuristics to guess the
feedback in missing ACKs, then when it subsequently receives feedback
it might find that it needs to correct its earlier heuristics as part
of the decoding process. The above decoding process does not include
any such heuristics.

A.2. Example Algorithm for Safety Against Long Sequences of ACK Loss

The example algorithms below show how a Data Receiver in AccECN mode
could encode its CE packet counter r.cep into the ACE field, and how
the Data Sender in AccECN mode could decode the ACE field into its
s.cep counter. The Data Sender's algorithm includes code to
heuristically detect a long enough unbroken string of ACK losses that
could have concealed a cycle of the congestion counter in the ACE
field of the next ACK to arrive.

Two variants of the algorithm are given: i) a more conservative
variant for a Data Sender to use if it detects that the AccECN Option
is not available (see Section 3.2.2.5 and Section 3.2.3.2); and ii) a
less conservative variant that is feasible when complementary
information is available from the AccECN Option.

A.2.1. Safety Algorithm without the AccECN Option

It is assumed that each local packet counter is a sufficiently sized
unsigned integer (probably 32b) and that the following constant has
been assigned:

```
    DIVACE = 2^3
```

Every time an Acceptable CE marked packet arrives (Section 3.2.2.2),
the Data Receiver increments its local value of r.cep by 1. It
repeats the same value of ACE in every subsequent ACK until the next
CE marking arrives, where
```

```
    ACE = r.cep % DIVACE.
```

If the Data Sender received an earlier value of the counter that had
been delayed due to ACK reordering, it might incorrectly calculate
that the ACE field had wrapped. Therefore, on the arrival of every
ACK, the Data Sender ensures the ACK has not been superseded using
the TCP acknowledgement number, any SACK options and timestamps (if
available) to calculate newlyAckedB, as in Appendix A.1. If the ACK
has not been superseded, the Data Sender calculates the minimum
difference d.cep between the ACE field and its local s.cep counter,
using modulo arithmetic as follows:

```
    if ((newlyAckedB > 0) || (newlyAckedT > 0))
        d.cep = (ACE + DIVACE - (s.cep % DIVACE)) % DIVACE
```

Section 3.2.2.5 expects the Data Sender to assume that the ACE field
cycled if it is the safest likely case under prevailing conditions.
The 3-bit ACE field in an arriving ACK could have cycled and become
ambiguous to the Data Sender if a sequence of ACKs goes missing that
covers a stream of data long enough to contain 8 or more CE marks.
We use the word `missing' rather than `lost', because some or all the
missing ACKs might arrive eventually, but out of order. Even if some
of the missing ACKs were piggy-backed on data (i.e. not pure ACKs)
retransmissions will not repair the lost AccECN information, because
AccECN requires retransmissions to carry the latest AccECN counters,
not the original ones.

The phrase `under prevailing conditions' allows for implementation-
dependent interpretation. A Data Sender might take account of the
prevailing size of data segments and the prevailing CE marking rate
just before the sequence of missing ACKs. However, we shall start
with the simplest algorithm, which assumes segments are all full-
sized and ultra-conservatively it assumes that ECN marking was 100%
on the forward path when ACKs on the reverse path started to all be
dropped. Specifically, if newlyAckedB is the amount of data that an
ACK acknowledges since the previous ACK, then the Data Sender could
assume that this acknowledges newlyAckedPkt full-sized segments,
where newlyAckedPkt = newlyAckedB/MSS. Then it could assume that the
ACE field incremented by

```
    dSafer.cep = newlyAckedPkt - ((newlyAckedPkt - d.cep) % DIVACE),
```

For example, imagine an ACK acknowledges newlyAckedPkt=9 more full-
size segments than any previous ACK, and that ACE increments by a
minimum of 2 CE marks (d.cep=2). The above formula works out that it
would still be safe to assume 2 CE marks (because 9 - ((9-2) % 8) =
2). However, if ACE increases by a minimum of 2 but acknowledges 10

full-sized segments, then it would be necessary to assume that there
could have been 10 CE marks (because 10 - ((10-2) % 8) = 10).

Note that checks would need to be added to the above pseudocode for
(d.cep > newlyAckedPkt), which could occur if newlyAckedPkt had been
wrongly estimated using an inappropriate packet size.

ACKs that acknowledge a large stretch of packets might be common in
data centres to achieve a high packet rate or might be due to ACK
thinning by a middlebox. In these cases, cycling of the ACE field
would often appear to have been possible, so the above algorithm
would be over-conservative, leading to a false high marking rate and
poor performance. Therefore it would be reasonable to only use
dSafer.cep rather than d.cep if the moving average of newlyAckedPkt
was well below 8.

Implementers could build in more heuristics to estimate prevailing
average segment size and prevailing ECN marking. For instance,
newlyAckedPkt in the above formula could be replaced with
newlyAckedPktHeur = newlyAckedPkt*p*MSS/s, where s is the prevailing
segment size and p is the prevailing ECN marking probability.
However, ultimately, if TCP's ECN feedback becomes inaccurate it
still has loss detection to fall back on. Therefore, it would seem
safe to implement a simple algorithm, rather than a perfect one.

The simple algorithm for dSafer.cep above requires no monitoring of
prevailing conditions and it would still be safe if, for example,
segments were on average at least 5% of full-sized as long as ECN
marking was 5% or less. Assuming it was used, the Data Sender would
increment its packet counter as follows:

    s.cep += dSafer.cep

If missing acknowledgement numbers arrive later (due to reordering),
Section 3.2.2.5 says "the Data Sender MAY attempt to neutralize the

effect of any action it took based on a conservative assumption that
it later found to be incorrect". To do this, the Data Sender would
have to store the values of all the relevant variables whenever it
made assumptions, so that it could re-evaluate them later. Given
this could become complex and it is not required, we do not attempt
to provide an example of how to do this.

A.2.2. Safety Algorithm with the AccECN Option

When the AccECN Option is available on the ACKs before and after the
possible sequence of ACK losses, if the Data Sender only needs CE-
marked bytes, it will have sufficient information in the AccECN
Option without needing to process the ACE field. If for some reason

it needs CE-marked packets, if dSafer.cep is different from d.cep, it
can determine whether d.cep is likely to be a safe enough estimate by
checking whether the average marked segment size (s = d.ceb/d.cep) is
less than the MSS (where d.ceb is the amount of newly CE-marked bytes
- see Appendix A.1). Specifically, it could use the following
algorithm:

```
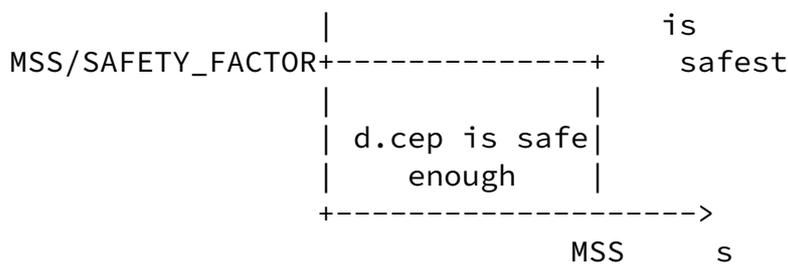    SAFETY_FACTOR = 2
    if (dSafer.cep > d.cep) {
        if (d.ceb <= MSS * d.cep) {  % Same as (s <= MSS), but no DBZ
            sSafer = d.ceb/dSafer.cep
            if (sSafer < MSS/SAFETY_FACTOR)
                dSafer.cep = d.cep    % d.cep is a safe enough estimate
        } % else
            % No need for else; dSafer.cep is already correct,
            % because d.cep must have been too small
    }
```

The chart below shows when the above algorithm will consider d.cep
can replace dSafer.cep as a safe enough estimate of the number of CE-
marked packets:

```
                  ^
           sSafer|
                 |
             MSS+
                 |
                 |           dSafer.cep
```

```
                         |                    is
     MSS/SAFETY_FACTOR+-------------+   safest
                         |             |
                         | d.cep is safe|
                         |    enough    |
                         +-------------------->
                                      MSS      s
```

   The following examples give the reasoning behind the algorithm,
   assuming MSS=1460 [B]:

   o if d.cep=0, dSafer.cep=8 and d.ceb=1460, then s=infinity and
      sSafer=182.5.
      Therefore even though the average size of 8 data segments is
      unlikely to have been as small as MSS/8, d.cep cannot have been
      correct, because it would imply an average segment size greater
      than the MSS.

   o if d.cep=2, dSafer.cep=10 and d.ceb=1460, then s=730 and
      sSafer=146.
      Therefore d.cep is safe enough, because the average size of 10
      data segments is unlikely to have been as small as MSS/10.

   o if d.cep=7, dSafer.cep=15 and d.ceb=10200, then s=1457 and
      sSafer=680.
      Therefore d.cep is safe enough, because the average data segment
      size is more likely to have been just less than one MSS, rather
      than below MSS/2.

   If pure ACKs were allowed to be ECN-capable, missing ACKs would be
   far less likely. However, because [RFC3168] currently precludes
   this, the above algorithm assumes that pure ACKs are not ECN-capable.

A.3. Example Algorithm to Estimate Marked Bytes from Marked Packets

   If the AccECN Option is not available, the Data Sender can only
   decode CE-marking from the ACE field in packets. Every time an ACK
   arrives, to convert this into an estimate of CE-marked bytes, it
   needs an average of the segment size, s_ave. Then it can add or

subtract s_ave from the value of d.ceb as the value of d.cep
increments or decrements. Some possible ways to calculate s_ave are
outlined below. The precise details will depend on why an estimate
of marked bytes is needed.

The implementation could keep a record of the byte numbers of all the
boundaries between packets in flight (including control packets), and
recalculate s_ave on every ACK. However it would be simpler to
merely maintain a counter packets_in_flight for the number of packets
in flight (including control packets), which is reset once per RTT.
Either way, it would estimate s_ave as:

    s_ave ~= flightsize / packets_in_flight,

where flightsize is the variable that TCP already maintains for the
number of bytes in flight. To avoid floating point arithmetic, it
could right-bit-shift by lg(packets_in_flight), where lg() means log
base 2.

An alternative would be to maintain an exponentially weighted moving
average (EWMA) of the segment size:

    s_ave = a * s + (1-a) * s_ave,

where a is the decay constant for the EWMA. However, then it is
necessary to choose a good value for this constant, which ought to

depend on the number of packets in flight. Also the decay constant
needs to be power of two to avoid floating point arithmetic.

A.4. Example Algorithm to Count Not-ECT Bytes

A Data Sender in AccECN mode can infer the amount of TCP payload data
arriving at the receiver marked Not-ECT from the difference between
the amount of newly ACKed data and the sum of the bytes with the
other three markings, d.ceb, d.e0b and d.e1b.

For this approach to be precise, it has to be assumed that spurious
(unnecessary) retransmissions do not lead to double counting. This
assumption is currently correct, given that RFC 3168 requires that
the Data Sender marks retransmitted segments as Not-ECT. However,

the converse is not true; necessary retransmissions will result in
under-counting.

However, such precision is unlikely to be necessary. The only known
use of a count of Not-ECT marked bytes is to test whether equipment
on the path is clearing the ECN field (perhaps due to an out-dated
attempt to clear, or bleach, what used to be the ToS field). To
detect bleaching it will be sufficient to detect whether nearly all
bytes arrive marked as Not-ECT. Therefore there should be no need to
keep track of the details of retransmissions.

Appendix B. Rationale for Usage of TCP Header Flags

B.1. Three TCP Header Flags in the SYN-SYN/ACK Handshake

AccECN uses a rather unorthodox approach to negotiate the highest
version TCP ECN feedback scheme that both ends support, as justified
below. It follows from the original TCP ECN capability negotiation
[RFC3168], in which the client set the 2 least significant of the
original reserved flags in the TCP header, and fell back to no ECN
support if the server responded with the 2 flags cleared, which had
previously been the default.

ECN originally used header flags rather than a TCP option because it
was considered more efficient to use a header flag for 1 bit of
feedback per ACK, and this bit could be overloaded to indicate
support for ECN during the handshake. During the development of ECN,
1 bit crept up to 2, in order to deliver the feedback reliably and to
work round some broken hosts that reflected the reserved flags during
the handshake.

In order to be backward compatible with RFC 3168, AccECN continues
this approach, using the 3rd least significant TCP header flag that
had previously been allocated for the ECN nonce (now historic).

Then, whatever form of server an AccECN client encounters, the
connection can fall back to the highest version of feedback protocol
that both ends support, as explained in Section 3.1.

If AccECN had used the more orthodox approach of a TCP option, it
would still have had to set the two ECN flags in the main TCP header,
in order to be able to fall back to Classic RFC 3168 ECN, or to

disable ECN support, without another round of negotiation. Then
AccECN would also have had to handle all the different ways that
servers currently respond to settings of the ECN flags in the main
TCP header, including all the conflicting cases where a server might
have said it supported one approach in the flags and another approach
in the new TCP option. And AccECN would have had to deal with all
the additional possibilities where a middlebox might have mangled the
ECN flags, or removed the TCP option. Thus, usage of the 3rd
reserved TCP header flag simplified the protocol.

The third flag was used in a way that could be distinguished from the
ECN nonce, in case any nonce deployment was encountered. Previous
usage of this flag for the ECN nonce was integrated into the original
ECN negotiation. This further justified the 3rd flag's use for
AccECN, because a non-ECN usage of this flag would have had to use it
as a separate single bit, rather than in combination with the other 2
ECN flags.

Indeed, having overloaded the original uses of these three flags for
its handshake, AccECN overloads all three bits again as a 3-bit
counter.

B.2. Four Codepoints in the SYN/ACK

Of the 8 possible codepoints that the 3 TCP header flags can indicate
on the SYN/ACK, 4 already indicated earlier (or broken) versions of
ECN support. In the early design of AccECN, an AccECN server could
use only 2 of the 4 remaining codepoints. They both indicated AccECN
support, but one fed back that the SYN had arrived marked as CE.
Even though ECN support on a SYN is not yet on the standards track,
the idea is for either end to act as a dumb reflector, so that future
capabilities can be unilaterally deployed without requiring 2-ended
deployment (justified in Section 2.5).

During traversal testing it was discovered that the ECN field in the
SYN was mangled on a non-negligible proportion of paths. Therefore
it was necessary to allow the SYN/ACK to feed all four IP/ECN
codepoints that the SYN could arrive with back to the client.
Without this, the client could not know whether to disable ECN for
the connection due to mangling of the IP/ECN field (also explained in
Section 2.5). This development consumed the remaining 2 codepoints

on the SYN/ACK that had been reserved for future use by AccECN in
earlier versions.

B.3. Space for Future Evolution

   Despite availability of usable TCP header space being extremely
   scarce, the AccECN protocol has taken all possible steps to ensure
   that there is space to negotiate possible future variants of the
   protocol, either if a variant of AccECN is required, or if a
   completely different ECN feedback approach is needed:

   Future AccECN variants: When the AccECN capability is negotiated
      during TCP's 3WHS, the rows in Table 2 tagged as 'Nonce' and
      'Broken' in the column for the capability of node B are unused by
      any current protocol in the RFC series. These could be used by
      TCP servers in future to indicate a variant of the AccECN
      protocol. In recent measurement studies in which the response of
      large numbers of servers to an AccECN SYN has been tested,
      e.g. [Mandalari18], a very small number of SYN/ACKs arrive with
      the pattern tagged as 'Nonce', and a small but more significant
      number arrive with the pattern tagged as 'Broken'. The 'Nonce'
      pattern could be a sign that a few servers have implemented the
      ECN Nonce [RFC3540], which has now been reclassified as historic
      [RFC8311], or it could be the random result of some unknown
      middlebox behaviour. The greater prevalence of the 'Broken'
      pattern suggests that some instances still exist of the broken
      code that reflects the reserved flags on the SYN.

      The requirement not to reject unexpected initial values of the ACE
      counter (in the main TCP header) in the last para of
      Section 3.2.2.4 ensures that 3 unused codepoints on the ACK of the
      SYN/ACK, 6 unused values on the first SYN=0 data packet from the
      client and 7 unused values on the first SYN=0 data packet from the
      server could be used to declare future variants of the AccECN
      protocol. The word 'declare' is used rather than 'negotiate'
      because, at this late stage in the 3WHS, it would be too late for
      a negotiation between the endpoints to be completed. A similar
      requirement not to reject unexpected initial values in the TCP
      option (Section 3.2.3.2.4) is for the same purpose. If traversal
      of the TCP option were reliable, this would have enabled a far
      wider range of future variation of the whole AccECN protocol.
      Nonetheless, it could be used to reliably negotiate a wide range
      of variation in the semantics of the AccECN Option.

   Future non-AccECN variants: Five codepoints out of the 8 possible in
      the 3 TCP header flags used by AccECN are unused on the initial
      SYN (in the order AE,CWR,ECE): 001, 010, 100, 101, 110.
      Section 3.1.3 ensures that the installed base of AccECN servers

will all assume these are equivalent to AccECN negotiation with
111 on the SYN. These codepoints would not allow fall-back to
Classic ECN support for a server that did not understand them, but
this approach ensures they are available in future, perhaps for
uses other than ECN alongside the AccECN scheme. All possible
combinations of SYN/ACK could be used in response except either
000 or reflection of the same values sent on the SYN.

Of course, other ways could be resorted to in order to extend
AccECN or ECN in future, although their traversal properties are
likely to be inferior. They include a new TCP option; using the
remaining reserved flags in the main TCP header (preferably
extending the 3-bit combinations used by AccECN to 4-bit
combinations, rather than burning one bit for just one state); a
non-zero urgent pointer in combination with the URG flag cleared;
or some other unexpected combination of fields yet to be invented.

Authors' Addresses

Bob Briscoe
Independent
UK

EMail: ietf@bobbriscoe.net
URI: http://bobbriscoe.net/

Mirja Kuehlewind
Ericsson
Germany

EMail: ietf@kuehlewind.net

Richard Scheffenegger
NetApp
Vienna
Austria

EMail: Richard.Scheffenegger@netapp.com