The Some Congestion Experienced ECN Codepoint
draft-morton-taht-tsvwg-sce-00

Abstract

   This memo reclassifies ECT(1) to be an early notification of
   congestion on ECT(0) marked packets, which can be used by AQM
   algorithms and transports as an earlier signal of congestion than CE.
   It is a simple, transparent, and backward compatible upgrade to
   existing IETF-approved AQMs, RFC3168, and nearly all congestion
   control algorithms.

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 11, 2019.

Copyright Notice

Table of Contents

1. Terminology

The keywords MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD,
SHOULD NOT, RECOMMENDED, MAY, and OPTIONAL, when they appear in this
document, are to be interpreted as described in [RFC2119].

2. Introduction

This memo reclassifies ECT(1) to be an early notification of
congestion on ECT(0) marked packets, which can be used by AQM
algorithms and transports as an earlier signal of congestion than CE
("Congestion Experienced").

This memo limits its scope to the redefinition of the ECT(1)
codepoint as SCE, "Some Congestion Experienced", with a few brief
illustrations of how it may be used.

3. Background

[RFC3168] defines the lower two bits of the (former) TOS byte in the
IPv4/6 header as the ECN field. This may take four values: Not-ECT,
ECT(0), ECT(1) or CE.

Binary Keyword References

```
   00     Not-ECT (Not ECN-Capable Transport)     [RFC 3168]
   01     ECT(1) (ECN-Capable Transport(1))       [RFC 3168]
   10     ECT(0) (ECN-Capable Transport(0))       [RFC 3168]
   11     CE (Congestion Experienced)             [RFC 3168]
```

Research has shown that the ECT(1) codepoint goes essentially unused, with the "Nonce Sum" extension to ECN having not been implemented in practice and thus subsequently obsoleted by [RFC8311] (section 3). Additionally, known [RFC3168] compliant senders do not emit ECT(1), and compliant middleboxes do not alter the field to ECT(1), while compliant receivers all interpret ECT(1) identically to ECT(0). These are useful properties which represent an opportunity for improvement.

Experience gained with 7 years of [RFC8290] deployment in the field suggests that it remains difficult to maintain the desired 100% link utilisation, whilst simultaneously strictly minimising induced delay due to excess queue depth - irrespective of whether ECN is in use. This leads to a reluctance amongst hardware vendors to implement the most effective AQM schemes because their headline benchmarks are throughput-based.

The underlying cause is the very sharp "multiplicative decrease" reaction required of transport protocols to congestion signalling (whether that be packet loss or CE marks), which tends to leave the congestion window significantly smaller than the ideal BDP when triggered at only slightly above the ideal value. The availability of this sharp response is required to assure network stability (AIMD principle), but there is presently no standardised and backwards-compatible means of providing a less drastic signal.

4. Some Congestion Experienced

As consensus has arisen that some form of ECN signaling should be an earlier signal than drop, this Internet Draft changes the meaning of ECT(1) to be SCE, meaning "Some Congestion Experienced". The above ECN-field codepoint table then becomes:

Binary Keyword References

```
   ------------------------------------------------------------

   00     Not-ECT (Not ECN-Capable Transport)     [@RFC3168]
   01     SCE (Some Congestion Experienced)       [This Internet-draft]
   10     ECT (ECN-Capable Transport)             [@RFC3168]
   11     CE (Congestion Experienced)             [@RFC3168]
```

This permits middleboxes implementing AQM to signal incipient
congestion, below the threshold required to justify setting CE, by
converting some proportion of ECT codepoints to SCE ("SCE marking").
Existing [RFC3168] compliant receivers MUST transparently ignore this
new signal, and both existing and SCE-aware middleboxes MAY convert
SCE to CE in the same circumstances as for ECT, thus ensuring
backwards compatibility with [RFC3168] ECN endpoints.

Permitted ECN codepoint packet transitions by middleboxes are:

         Not-ECT ->   Not-ECT or DROP
         ECT   ->   ECT or SCE or CE or DROP
         SCE   ->   SCE or CE or DROP
         CE   ->   CE or DROP

In other words, for ECN-aware flows, the ECN marking of an individual
packet MAY be increased by a middlebox to signal congestion, but MUST
NOT be decreased, and packets SHALL NOT be altered to appear to be
ECN-aware if they were not originally, nor vice versa. Note however
that SCE is numerically less than ECT, but semantically greater, and
the latter definition applies for this rule.

New SCE-aware receivers and transport protocols SHALL continue to
apply the [RFC3168] interpretation of the CE codepoint, that is, to
signal the sender to back off send rate to the same extent as if a
packet loss were detected. This maintains compatibility with
existing middleboxes, senders and receivers.

New SCE-aware receivers and transport protocols SHOULD interpret the
SCE codepoint as an indication of mild congestion, and respond
accordingly by applying send rates intermediate between those
resulting from a continuous sequence of ECT codepoints, and those
resulting from a CE codepoint. The ratio of ECT and SCE codepoints
received indicates the relative severity of such congestion, such
that 100% SCE is very close to the threshold of CE marking, 100% ECT
indicates that the bottleneck link may not be fully utilised, and a
1:1 balance of ECT and SCE codepoints indicates that the present send
rate is a good match to the bottleneck link.

Details of how to implement SCE awareness at the transport layer will
be left to additional Internet Drafts yet to be submitted.

To maximise the benefit of SCE, middleboxes SHOULD produce SCE
markings sooner than they produce CE markings, when the level of
congestion increases.

5. Examples of use

5.1. Cubic

   Consider a TCP transport implementing CUBIC congestion control. This
   presently exhibits exponential cwnd growth during slow-start,
   polynomial cwnd growth in steady-state, and multiplicative decrease
   upon detecting a single CE marking or packet loss in one RTT cycle.

   With SCE awareness, it might exit slow-start upon detecting a single
   SCE marking, switch from polynomial to Reno-linear cwnd growth when
   the SCE:ECT ratio exceeds 1:2, halt cwnd growth entirely when it
   exceeds 1:1, and implement a Reno-linear decline when it exceeds 2:1,
   in addition to retaining the sharp 40% decrease on detecting CE.

   In ideal circumstances, the above behaviour would result in the send
   rate stabilising at a level which produces between 50% and 66% SCE
   marking at some bottleneck on the path. The middlebox performing
   this marking can thus control the send rate smoothly to an ideal
   value, maximising throughput with minimum average queue length.

5.2. TCP receiver side handling

   SCE can potentially be handled entirely by the receiver and be
   entirely independent of any of the dozens of [RFC3168] compliant
   congestion control algorithms, for example by manipulating the TCP
   receive window in a similar manner to the sender's congestion window.

   Alternatively, some mechanism may be defined to feed back SCE signals
   to the sender explicitly. Details of this are left to future I-Ds.

5.3. Other

   New transports under development such as QUIC SHOULD implement a
   multi-bit, sub-RTT, and finer grained signal back to the sender based
   on SCE.

6. Related Work

   [RFC8087] [RFC7567] [RFC7928] [RFC8290] [RFC8289] [RFC8033] [RFC8034]

7. IANA Considerations

   There are no IANA considerations.

8. Security Considerations

There are no security considerations.

9. Acknowledgements

Many thanks to John Gilmore, the members of the ecn-sane project and the cake@lists.bufferbloat.net mailing list, and the former IETF AQM working group.

10. References

10.1. Normative References

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

[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>.

10.2. Informative References

[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>.

[RFC7567]   Baker, F., Ed. and G. Fairhurst, Ed., "IETF
            Recommendations Regarding Active Queue Management",
            BCP 197, RFC 7567, DOI 10.17487/RFC7567, July 2015,
            <https://www.rfc-editor.org/info/rfc7567>.

[RFC7928]   Kuhn, N., Ed., Natarajan, P., Ed., Khademi, N., Ed., and
            D. Ros, "Characterization Guidelines for Active Queue
            Management (AQM)", RFC 7928, DOI 10.17487/RFC7928, July
            2016, <https://www.rfc-editor.org/info/rfc7928>.

[RFC8033]   Pan, R., Natarajan, P., Baker, F., and G. White,
            "Proportional Integral Controller Enhanced (PIE): A
            Lightweight Control Scheme to Address the Bufferbloat
            Problem", RFC 8033, DOI 10.17487/RFC8033, February 2017,
            <https://www.rfc-editor.org/info/rfc8033>.

   [RFC8034] White, G. and R. Pan, "Active Queue Management (AQM) Based
              on Proportional Integral Controller Enhanced PIE) for
              Data-Over-Cable Service Interface Specifications (DOCSIS)
              Cable Modems", RFC 8034, DOI 10.17487/RFC8034, February
              2017, <https://www.rfc-editor.org/info/rfc8034>.

   [RFC8087] Fairhurst, G. and M. Welzl, "The Benefits of Using
              Explicit Congestion Notification (ECN)", RFC 8087,
              DOI 10.17487/RFC8087, March 2017,
              <https://www.rfc-editor.org/info/rfc8087>.

   [RFC8289] Nichols, K., Jacobson, V., McGregor, A., Ed., and J.
              Iyengar, Ed., "Controlled Delay Active Queue Management",
              RFC 8289, DOI 10.17487/RFC8289, January 2018,
              <https://www.rfc-editor.org/info/rfc8289>.

   [RFC8290] Hoeiland-Joergensen, T., McKenney, P., Taht, D., Gettys,
              J., and E. Dumazet, "The Flow Queue CoDel Packet Scheduler
              and Active Queue Management Algorithm", RFC 8290,
              DOI 10.17487/RFC8290, January 2018,
              <https://www.rfc-editor.org/info/rfc8290>.

Authors' Addresses

   Jonathan Morton
   Bufferbloat.net
   Koekkoenranta 21
   PITKAeJAeRVI 31520
   FINLAND

   Phone: +358 44 927 2377
   Email: chromatix99@gmail.com

   David M. Taeht
   TekLibre
   20600 Aldercroft Heights Rd
   Los Gatos, Ca 95033
   USA

   Phone: +18312059740
   Email: dave@taht.net