TCPM Working Group
Internet-Draft
Intended status: Experimental
Expires: March 14, 2014

M. Kuehlewind
University of Stuttgart
B. Trammell
ETH Zurich

September 10, 2013

A Mechanism for ECN Path Probing and Fallback
draft-kuehlewind-tcpm-ecn-fallback-01.txt

Abstract

Explicit Congestion Notification (ECN) is a TCP/IP extension that is widely implemented but hardly used due to the perceived unusablilty of ECN on many paths through the Internet caused by ECN-ignorant routers and middleboxes. This document specifies an ECN probing and fall-back mechanism in case ECN has be successfully negotiated between two connection endpoints, but might not be usable on the path.

Status of This Memo

Copyright Notice

1. Introduction

The deployment of Explicit Congestion Notification (ECN) [RFC3168] and AQM would arguably improve end-to-end performance in the Internet, by providing a congestion signal to the transport layer without relying on queue tail drop and packet loss. However, though ECN has been standardized since 2000, implementation and deployment have lagged significantly, in part due to the perceived unusablilty of ECN on many paths through the Internet caused by ECN-ignorant routers and middleboxes.

Recent research by the authors [KuNeTr13] has shown accelerating deployment of ECN-capable servers in the Internet, due to the deployment of TCP stacks for which ECN is enabled by default. In addition, ECN is usable end-to-end on the vast majority of paths measured in this study: that is, a Congestion Experienced mark will cause a ECN Echo on the associated ACK. However, there still exist a non-negligible number of paths on which a successfully negotiated usage of ECN will not result in a connection on which congestion will be correctly echoed, or worse, leads to the loss of packets with CE or ECE set.

This document presents an experimental, in-band, runtime method for determining the usability of ECN by a given traffic flow, based on the active measurement method described in [KuNeTr13]. If ECN is successfully negotiated but found by this method to be unusable, it can be disabled on subsequent packets in the flow in order to avoid connectivity problems caused by ECN-unusability on the path.

2. ECN Path Capability Probing

A TCP sender can determine whether or not its path to the receiver is usable for ECN using the procedure detailed below.

1. The sender attempts to negotiate ECN usage as per Section 6.1.1 of [RFC3168]. If ECN is not successfully negotiated, the procedure ends, and ECN is not used for the duration of the connection.

2. The sender disables the normal usage of ECN for the duration of the procedure, as the ECN codepoints are used for path probing. This means all segments are sent with the Non-ECN-Capable codepoint during this procedure unless otherwise stated. Moreover, the sender will only take loss as a congestion signal

and will not react with window reductions to the ECN-Echo (ECE) feedback signal from the receiver during this procedure.

3. The sender sets the Non-ECN-Capable codepoint in the IP header until it has completed sending the first N data segments, where N is the size of the initial congestion window. Loss is used to discover congestion for these segments.

4. The next three data segments sent consist of the "CE probe": these three segments are sent with the Congestion Experienced codepoint set.

5. If all three of the CE probe segments are lost and must be retransmitted, the path is deemed not ECN-usable and the sender falls back as in Section 3.

6. If the ECE flag is not set on the ACK segment(s) sent by the receiver acknowledging the CE probe segments, the path may or may not be usable, as that there might be middleboxes/gateways that clear CE on segments from end hosts, because they assume that congestion can not have occurred up to this point on the path. In this case, the sender may continue using ECN, because while it may not work for detecting congestion, the use of ECN does not negatively affect connectivity.

7. While the sender does not reduce the congestion window for the ECE ACKS acknowledging the the CE probe segments, it does set CWR on the subsequent segment sent.

8. If no fallback has occurred by the time the ACK of the final CE probe segment is received, the path is deemed ECN usable, and the sender ends the probing procedure and proceeds to use ECN normally as in [RFC3168].

The operation of this procedure on an ECN-usable path, with an initial window size of 3, bulk data transfer initiated by the sender, where the receiver does not perform probing, is shown in Figure 1.

```
   Sender                    Receiver
     |                            |
     |----SYN ECE CWR----------->|  connection
     |<-----------SYN ACK ECE----|  establishment
     |----ACK------------------->|
     |                            |
     |----data 0---------------->|  initial window
     |<------------ECT0 ACK 1----|  transmission
     |----data 1---------------->|
     |----data 2---------------->|
     |<------------ECT0 ACK 3----|
     |                            |
     |----data 3 CE------------->|  CE probe
     |----data 4 CE------------->|
     |----data 5 CE------------->|
     |----data 6 --------------->|
     |<--------ECT0 ECE ACK 7----|  CE probe ACK: ECN OK
     |                            |
     |----data 7  ECT0 CWR------>|  CE probe CWR
     |----data 8  ECT0---------->|  and transition to
     |----data 9  ECT0---------->|  normal ECN usage
```

Figure 1: CE probing on ECN-usable path

Conversely, the operation of the probe and fallback procedure on an ECN-unusable path with an initial window size of 3, bulk data transfer initiated by the sender, where the receiver does not perform probing, is shown in Figure 2.

```
   Sender                     Receiver
     |                            |
     |----SYN ECE CWR----------->|  connection
     |<-----------SYN ACK ECE----|  establishment
     |----ACK------------------->|
     |                            |
     |----data 0---------------->|  initial window
     |<-----------------ACK 1----|  transmission
     |----data 1---------------->|
     |----data 2---------------->|
     |<-----------------ACK 3----|
     |                            |
     |----data 3 CE----X          |  CE probe
     |----data 4 CE----X          |  (blocked)
     |----data 5 CE----X          |
     |----data 6 --------------->|
     |<-----------------ACK 3----|  dupack
     |----data 7 --------------->|
     |----data 8---------------->|
     |<-----------------ACK 3----|  dupack
     |                            |
     |----data 3---------------->|  retransmit without CE
     |<-----------------ACK 4----|
     |----data 4---------------->|
     |<-----------------ACK 5----|
     |----data 5---------------->|
     |<-----------------ACK 9----|  and fallback (no ECN)
     |----data 9---------------->|
```

Figure 2: Fallback on ECN-unusable path

As the probing begins after all the segments in the initial congestion window have been sent, it requires more than an initial congestion window plus 6 segments (3 CE probe + 3 duplicated ACKs) of available data to send. TCP implementations can use socket send buffer occupancy as a signal as to whether sufficient data is available to use the probing procedure. If enough data is already in the buffer by the time the initial congestion window is sent, then the probe segments SHOULD be sent. Otherwise, the implementation can use additional heuristics, outside the scope of this document to define, to determine if significant data is likely to be available.

3. ECN Fallback

If ECN is found to be unusable on a given flow by path capability probing as in Section 2 above, the sender simply stops setting any ECN-Capable-Transport codepoint on subsequent packets in the flow. The receiver MUST, however, still set ECE on any ACK for a packet

with CE set. Note that this behavior is consistent with section 6.1.1 of [RFC3168].

A sender may keep a cache of paths found to be unusable for ECN and disable ECN for subsequent connections on a per-destination basis. In this case, the reciever should periodically (i.e., on the order of hours or days) expire these cache entries to cause re-probing to occur in order to account for routing changes in the network.

[EDITOR'S NOTE: what to do on RTO?]

4. Discussion

[EDITOR'S NOTE: the general case of this algorithm is J ECT segments followed by K CE segments; we chose (J,K) = (0,3). We should justify this choice.]

[EDITOR'S NOTE: need to think about how this would interact with conex; an analysis comparing the delay caused by path probing as opposed to the delay caused by ECN failure would be interesting.]

[EDITOR'S NOTE: initial implementation results go here?.]

5. Security Considerations

[FIXME: we'll have to explore attacks against this mechanism which could affect network or connection stability, so the following is wrong...]

This document has no security considerations.

6. IANA Considerations

This document has no IANA considerations.

7. Acknowledgements

Thanks to Michael Welzl and Richard Scheffenegger for the comments and discussions. This work is partially materially supported by the European Commission under grant agreement FP7-ICT-318627 mPlane.

8. References

8.1. Normative References

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

8.2. Informative References

[KuNeTr13]
Kuehlewind, M., Neuner, S., and B. Trammell, "On the state of ECN and TCP Options on the Internet", Mar 2013.

(In LNCS 7799, Proceedings of PAM 2013, Hong Kong)

Authors' Addresses

Mirja Kuehlewind
University of Stuttgart
Pfaffenwaldring 47
70569 Stuttgart
Germany

Email: mirja.kuehlewind@ikr.uni-stuttgart.de

Brian Trammell
Swiss Federal Institute of Technology Zurich
Gloriastrasse 35
8092 Zurich
Switzerland

Phone: +41 44 632 70 13
Email: trammell@tik.ee.ethz.ch