Internet Engineering Task Force A. Kuzmanovic
INTERNET DRAFT Northwestern University
draft-ietf-tsvwg-ecnsyn-00.txt S. Floyd
ICIR
K.K. Ramakrishnan
AT&T
October, 2005

Adding Explicit Congestion Notification (ECN) Capability to TCP's SYN/ACK Packets

Status of this Memo

Copyright Notice

Abstract

This draft specifies a modification to RFC 3168 to allow TCP SYN/ACK packets to be ECN-Capable. For TCP, RFC 3168 only specified setting

an ECN-Capable codepoint on data packets, and not on SYN and SYN/ACK packets. However, because of the high cost to the TCP transfer of having a SYN/ACK packet dropped, with the resulting retransmit timeout, this document is specifying the use of ECN for the SYN/ACK packet itself, when sent in response to a SYN packet with the two ECN flags set in the TCP header, indicating a willingness to use ECN. Setting TCP SYN/ACK packets as ECN-Capable can be of great benefit to the TCP connection, avoiding the severe penalty of a retransmit timeout for a connection that has not yet started placing a load on the network. The sender of the SYN/ACK packet must respond to an ECN mark by reducing its initial congestion window from two, three, or four segments to one segment, reducing the subsequent load from that connection on the network.

NOTE TO RFC EDITOR: PLEASE DELETE THIS NOTE UPON PUBLICATION.

Changes from draft-kuzmanovic-ecn-syn-00.txt:

* Changed name of draft to draft-ietf-twvsg-ecnsyn.

END OF NOTE TO RFC EDITOR.

1. Conventions

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 [RFC 2119].

1. Introduction

TCP's congestion control mechanism has primarily used packet loss as the congestion indication, with packets dropped when buffers overflow. With such tail-drop mechanisms, the packet delay can be high, as the queue at bottleneck routers can be fairly large. Dropping packets only when the queue overflows, and having TCP react only to such losses, results in:
1) significantly higher packet delay;
2) unnecessarily many packet losses; and
3) unfairness due to synchronization effects.

The adoption of Active Queue Management (AQM) mechanisms allows better control of bottleneck queues. This use of AQM has the following potential benefits:
1) better control of the queue, with reduced queueing delay;
2) fewer packet drops; and
3) better fairness because of fewer synchronization effects.

With the adoption of ECN, performance may be further improved. When the router detects congestion before buffer overflow, the router can provide a congestion indication either by dropping a packet, or by setting the Congestion Experienced (CE) codepoint in the Explicit Congestion Notification (ECN) field in the IP header [RFC3168]. The IETF has standardized the use of the Congestion Experienced (CE) codepoint in the IP header for routers to indicate congestion. For incremental deployment and backwards compatibility, the RFC on ECN [RFC 3168] specifies that routers may mark ECN-capable packets that would otherwise have been dropped, using the Congestion Experienced codepoint in the ECN field. The use of ECN allows TCP to react to congestion while avoiding unnecessary retransmissions and, in some cases, unnecessary retransmit timeouts. Thus, using ECN has several benefits:

1) For short transfers, a TCP connection's congestion window may be small. For example, if the current window contains only one packet, and that packet is dropped, TCP will have to wait for a retransmit timeout to recover, reducing its overall throughput. Similarly, if the current window contains only a few packets and one of those packets is dropped, there might not be enough duplicate acknowledgements for a fast retransmission, and the sender might have to wait for a delay of several round-trip times using Limited Transmit [RFC3042]. With the use of ECN, short flows are less likely to have packets dropped, sometimes avoiding unnecessary delays or costly retransmit timeouts.

2) While longer flows may not see substantially improved throughput with the use of ECN, they experience lower loss. This may benefit TCP applications that are latency- and loss-sensitive, because of the avoidance of retransmissions.

RFC 3168 only specified marking the Congestion Experienced codepoint on TCP's data packets, and not on SYN and SYN/ACK packets. RFC 3168 specified the negotiation of the use of ECN between the two TCP end-points in the TCP SYN and SYN-ACK exchange, using flags in the TCP header. Erring on the side of being conservative, RFC 3168 did not specify the use of ECN for the SYN/ACK packet itself. However, because of the high cost to the TCP transfer of having a SYN/ACK packet dropped, with the resulting retransmit timeout, this document is specifying the use of ECN for the SYN/ACK packet itself. This can be of great benefit to the TCP connection, avoiding the severe penalty of a retransmit timeout for a connection that has not yet started placing a load on the network. The sender of the SYN/ACK packet must respond to an ECN mark by reducing its initial congestion window from two, three, or four segments to one segment, reducing the subsequent load from that connection on the network.

The use of ECN for SYN/ACK packets has the following potential

benefits:
1) Avoidance of a retransmit timeout;
2) Improvement in the throughput of short connections.

This draft specifies a modification to RFC 3168 to allow TCP SYN/ACK packets to be ECN-Capable. Section 2 contains the specification of the change, while Section 3 discusses some of the issues, and Section 4 discusses related work. Section 5 contains an evaluation of the proposed change.

2. Proposal

This section specifies the modification to RFC 3168 to allow TCP SYN/ACK packets to be ECN-Capable. We use the following terminology from RFC 3168:

The ECN field in the IP header:
- o CE: the Congestion Experienced codepoint; and
- o ECT: either one of the two ECN-Capable Transport codepoints.

The ECN flags in the TCP header:
- o CWR: the Congestion Window Reduced flag; and
- o ECE: the ECN-Echo flag.

ECN-setup packets:
- o ECN-setup SYN packet: a SYN packet with the ECE and CWR flags;
- o ECN-setup SYN-ACK packet: a SYN-ACK packet with ECE but not CWR.

RFC 3168 in Section 6.1.1. states that "A host MUST NOT set ECT on SYN or SYN-ACK packets." In this section, we specify that a TCP node MAY respond to an ECN-setup SYN packet by setting ECT in the responding ECN-setup SYN/ACK packet, indicating to routers that the SYN/ACK packet is ECN-Capable. This allows a congested router along the path to mark the packet instead of dropping the packet as an indication of congestion.

Assume that TCP node A transmits to TCP node B an ECN-setup SYN packet, indicating willingness to use ECN for this connection. As specified by RFC 3168, if TCP node B is willing to use ECN, node B responds with an ECN-setup SYN-ACK packet.

Table 1 shows an interchange with the SYN/ACK packet dropped by a congested router. Node B waits for a retransmit timeout, and then retransmits the SYN/ACK packet.

```
     ---------------------------------------------------------------
        TCP Node A            Router                 TCP Node B
        ----------            ------                 ----------

        ECN-setup SYN packet --->
                                         ECN-setup SYN packet --->

                                  <--- ECN-setup SYN/ACK, possibly ECT
                                                3-second timer set
                                SYN/ACK dropped           .
                                                          .
                                                          .
                                            3-second timer expires
                                       <--- ECN-setup SYN/ACK, not ECT
        <--- ECN-setup SYN/ACK
        Data/ACK --->
                                                     Data/ACK --->
                                      <--- Data (one to four segments)
     ---------------------------------------------------------------
```

Table 1: SYN exchange with the SYN/ACK packet dropped.

If the SYN/ACK packet is dropped in the network, the TCP host (node B) responds by waiting three seconds for the retransmit timer to expire [RFC2988]. If a SYN/ACK packet with the ECT codepoint is dropped, the TCP node SHOULD resend the SYN/ACK packet without the ECN-Capable codepoint. (Although we are not aware of any middleboxes that drop SYN/ACK packets that contain an ECN-Capable codepoint in the IP header, we have learned to design our protocols defensively in this regard [RFC3360].)

Table 2 shows an interchange with the SYN/ACK packet sent as ECN-Capable, and ECN-marked instead of dropped at the congested router.

```
 ---------------------------------------------------------------
    TCP Node A            Router                TCP Node B
    ----------            ------                ----------

    ECN-setup SYN packet --->
                                ECN-setup SYN packet --->

                              <--- ECN-setup SYN/ACK, ECT
                   <--- Sets CE on SYN/ACK
    <--- ECN-setup SYN/ACK, CE

    Data/ACK, ECN-Echo --->
                                  Data/ACK, ECN-Echo --->
                             Window reduced to one segment.
                        <--- Data, CWR (one segment only)
 ---------------------------------------------------------------
```

Table 2: SYN exchange with the SYN/ACK packet marked.

If the receiving node (node A) receives a SYN/ACK packet that has been marked by the congested router, with the CE codepoint set, the receiving node MUST respond by setting the ECN-Echo flag in the TCP header of the responding ACK packet. As specified in RFC 3168, the receiving node continues to set the ECN-Echo flag in packets until it receives a packet with the CWR flag set.

When the sending node (node B) receives the ECN-Echo packet reporting the Congestion Experienced indication in the SYN/ACK packet, the node MUST set the initial congestion window to one segment, instead of two segments as allowed by [RFC2414], or three or four segments allowed by [RFC3390]. If the sending node (node B) was going to use an initial window of one segment, and receives an ECN-Echo packet informing it of a Congestion Experienced indication on its SYN/ACK packet, the sending node MAY continue to send with an initial window of one segment, without waiting for a retransmit timeout. We note that this updates RFC 3168, which specifies that "the sending TCP MUST reset the retransmit timer on receiving the ECN-Echo packet when the congestion window is one." As specified by RFC 3168, the sending node (node B) also sets the CWR flag in the TCP header of the next packet sent, to acknowledge its receipt of and reaction to the ECN-Echo flag.

3. Discussion

Motivation:
The rationale for the proposed change is the following. When node B receives a TCP SYN packet with ECN-Echo bit set in the TCP header,

this indicates that node A is ECN-capable. If node B is also ECN-capable, there are no obstacles to immediately setting one of the ECN-Capable codepoints in the IP header in the responding TCP SYN/ACK packet.

There can be a great benefit in setting an ECN-capable codepoint in SYN/ACK packets, as is discussed further in Section 4. Congestion is most likely to occur in the server-to-client direction. As a result, setting an ECN-capable codepoint in SYN/ACK packets can reduce the occurence of three-second retransmit timeouts resulting from the drop of SYN/ACK packets.

Flooding attacks:
Setting an ECN-Capable codepoint in the responding TCP SYN/ACK packets does not raise any novel security vulnerabilities. For example, provoking servers or hosts to send SYN/ACK packets to a third party in order to perform a "SYN/ACK flood" attack would be greatly inefficient. Third parties would immediately drop such packets, since they would know that they didn't generate the TCP SYN packets in the first place. Moreover, such SYN/ACK attacks would have the same signatures as the existing TCP SYN attacks. Provoking servers or hosts to reply with SYN/ACK packets in order to congest a certain link would also be highly inefficient because SYN ACK packets are small in size.

The TCP SYN packet:
There are several reasons why an ECN-Capable codepoint MUST NOT be set in the IP header of the initiating TCP SYN packet. First, when the TCP SYN packet is sent, there are no guarantees that the other TCP endpoint (node B in Table 2) is ECN-capable, or that it would be able to understand and react if the ECN CE codepoint was set by a congested router.

Second, the ECN-Capable codepoint in TCP SYN packets could be misused by malicious clients to 'improve' the well-known TCP SYN attack. By setting an ECN-Capable codepoint in TCP SYN packets, a malicious host might be able to inject a large number of TCP SYN packets through a potentially congested ECN-enabled router, congesting it even further.

For both these reasons, we continue the restriction that the TCP SYN packet MUST NOT have the ECN-Capable codepoint in the IP header set.

Backwards compatibility:
If there are some older TCP implementations that don't respond to the Congestion Experienced codepoint in a SYN/ACK packet, that would not be an insurmountable problem. It would mean that the sender of the SYN/ACK packet would not reduce the initial congestion window from two, three, or four segments down to one segment, as it should.

However, the TCP sender would still respond correctly to any subsequent CE indications on data packets later on in the connection.

SYN/ACK packets and packet size:
There are a number of router buffer architectures that have smaller dropping rates for small (SYN) packets than for large (data) packets. For example, for a Drop Tail queue in units of packets, where each packet takes a single slot in the buffer regardless of packet size, small and large packets are equally likely to be dropped. However, for a Drop Tail queue in units of bytes, small packets are less likely to be dropped than are large ones. Similarly, for RED in packet mode, small and large packets are equally likely to be dropped or marked, while for RED in byte mode, a packet's chance of being dropped or marked is proportional to the packet size in bytes.

For a congested router with an AQM mechanism in byte mode, where a packet's chance of being dropped or marked is proportional to the packet size in bytes, the drop or marking rate for TCP SYN/ACK packets should generally be low. In this case, the benefit of making SYN/ACK packets ECN-Capable should be similarly moderate. However, for a congested router with a Drop Tail queue in units of packets or with an AQM mechanism in packet mode, and with no priority queueing for smaller packets, small and large packets should have the same probability of being dropped or marked. In such a case, making SYN/ACK packets ECN-Capable should be of significant benefit.

We believe that there are a wide range of behaviors in the real world in terms of the drop or mark behavior at routers as a function of packet size [Tools, Section 10]. We note that all of these alternatives listed above are available in the NS simulator (Drop Tail queues are by default in units of packets, while the default for RED queue management has been changed from packet mode to byte mode).

4. Related Work

The addition of ECN-capability to TCP's SYN/ACK packets was proposed in [ECN+]. The paper includes an extensive set of simulation and testbed experiments to evaluate the effects of the proposal, using several Active Queue Management (AQM) mechanisms, including Random Early Detection (RED) [RED], Random Exponential Marking (REM) [REM], and Proportional Integrator (PI) [PI]. The performance measures were the end-to-end response times for each request/response pair, and the aggregate throughput on the bottleneck link. The end-to-end response time was computed as the time from the moment when the request for the file is sent to the server, until that file is successfully downloaded by the client.

The measurements from [ECN+] showed that setting an ECN-Capable codepoint in the IP packet header in TCP SYN/ACK packets systematically improves performance with all evaluated AQM schemes. When SYN/ACK packets at a congested router are ECN-marked instead of dropped, this can avoid a long initial retransmit timeout, improving the response time for the affected flow dramatically.

[ECN+] showed that the impact on aggregate throughput can also be quite significant, because marking SYN ACK packets can prevent larger flows from suffering long timeouts before being "admitted" into the network. In addition, the testbed measurements from [ECN+] showed that Web servers setting the ECN-Capable codepoint in TCP SYN/ACK packets could serve more requests.

As a final step, [ECN+] explored the co-existence of flows that do and don't set the ECN-capable codepoint in TCP SYN/ACK packets. The results in [ECN+] confirmed that both types of flows can coexist; flows that apply the change improve their end-to-end performance, while the performance degradation for flows that don't apply the change, as a result of the flows that do apply the change, is marginal.

5. Evaluation

The addition of ECN-capability to SYN/ACK packets could be of significant benefit for those ECN connections that would have had the SYN/ACK packet dropped in the network, and for which the ECN-Capability would allow the SYN/ACK to be marked rather than dropped.

The percent of SYN/ACK packets on a link can be quite high. In particular, measurements on links dominated by Web traffic indicate that 15-20% of the packets can be SYN/ACK packets [SCJO01].

The benefit of adding ECN-capability to SYN/ACK packets depends in part on the size of the data transfer. The drop of a SYN/ACK packet can increase the download time of a short file by an order of magnitude, by requiring a three-second retransmit timeout. For longer-lived flows, the effect of a dropped SYN/ACK packet on file download time is less dramatic. However, even for longer-lived flows, the addition of ECN-capability to SYN/ACK packets can improve the fairness among long-lived flows, as newly-arriving flows would be less likely to have to wait for retransmit timeouts.

The question that arises of course is what fraction of connections would see the benefit from making SYN/ACK packets ECN-capable, in a particular scenario? Specifically:

(1) What fraction of arriving SYN/ACK packets are dropped at the congested router when the SYN/ACK packets are not ECN-capable?

(2) Of those SYN/ACK packets that are dropped, what fraction of those drops would have been ECN-marks instead of drops if the SYN/ACK packets had been ECN-capable?

To answer (1), it is necessary to consider not only the level of congestion but also the queue architecture at the congested link. As described in Section 3 above, for some queue architectures small packets are less likely to be dropped than large ones. In such an environment, SYN/ACK packets would have lower packet drop rates; question (1) could not necessarily be inferred from the overall packet drop rate, but could be answered by measuring the drop rate for SYN/ACK packets directly. In such an environment, adding ECN-capability to SYN/ACK packets would be of less dramatic benefit than in environments where all packets are equally likely to be dropped regardless of packet size.

As question (2) implies, even if all of the SYN/ACK packets were ECN-capable, there could still be some SYN/ACK packets dropped instead of marked at the congested link; the full answer to question (2) depends on the details of the queue management mechanism at the router. If congestion is sufficiently bad, and the queue management mechanism cannot prevent the buffer from overflowing, then SYN/ACK packets will be dropped rather than marked upon buffer overflow whether or not they are ECN-capable.

For some AQM mechanisms, ECN-capable packets are marked instead of dropped any time this is possible, that is, any time the buffer is not yet full. For other AQM mechanisms however, such as the RED mechanism as recommended in [RED], packets are dropped rather than marked when the packet drop/mark rate exceeds a certain threshold, e.g., 10%, even if the packets are ECN-capable. For a router with such an AQM mechanism, when congestion is sufficiently severe to cause a high drop/mark rate, some SYN/ACK packets would be dropped instead of marked whether or not they were ECN-capable.

Thus, the degree of benefit of adding ECN-Capability to SYN/ACK packets depends not only on the overall packet drop rate in the network, but also on the queue management architecture at the congested link.

6. Security Considerations

TCP packets carrying the ECT codepoint in IP headers can be marked rather than dropped by ECN-capable routers. This raises several security concerns that we discuss below.

TCP SYN flooding attacks:
By setting an ECN-Capable codepoint in TCP SYN packets, a malicious

host might be able to inject a large number of TCP SYN packets through a potentially congested ECN-enabled router, congesting it even further. This is one of the reasons why an ECN-Capable codepoint MUST NOT be set in the IP header of the initiating TCP SYN packet. On the other hand, as discussed in Section 3 above, setting an ECN-Capable codepoint in the responding TCP SYN/ACK packet does not raise any novel security vulnerabilities.

"Bad" middleboxes:
While there is no evidence that any middleboxes drop SYN/ACK packets that contain an ECN-Capable codepoint in the IP header, such behavior cannot be excluded [RFC3360]. Thus, if a SYN/ACK packet with the ECT codepoint is dropped, the TCP node SHOULD resend the SYN/ACK packet without the ECN-Capable codepoint.

Congestion collapse:
Because TCP SYN/ACK packets carrying an ECT codepoint could be ECN-marked instead of dropped at an ECN-capable router, the concern is whether this can either invoke congestion, or worsen performance in highly congested scenarios. This is not a problem because after learning that the SYN/ACK packet was ECN-marked, the sender of that packet will only send one data packet; in the case that this data packet is ECN-marked, the sender will wait for a retransmission timeout. In addition, routers are free to drop rather than mark arriving packets in times of high congestion, regardless of whether the packets are ECN-capable.

7. Conclusions

This draft specifies a modification to RFC 3168 to allow TCP nodes to send SYN/ACK packets as being ECN-Capable. Making the SYN/ACK packet ECN-Capable avoids the high cost to a TCP transfer when a SYN/ACK packet is dropped by a congested router, by avoiding the resulting retransmit timeout. This improves the throughput of short connections. The sender of the SYN/ACK packet responds to an ECN mark by reducing its initial congestion window from two, three, or four segments to one segment, reducing the subsequent load from that connection on the network. The addition of ECN-capability to SYN/ACK packets is particularly beneficial in the server-to-client direction, where congestion is more likely to occur. In this case, the initial information provided by the ECN marking in the SYN/ACK packet enables the server to more appropriately adjust the initial load it places on the network.

8. Acknowledgements

9. Normative References

[RFC2414] M. Allman, S. Floyd, and C. Partridge, Increasing TCP's Initial Window, RFC 2414, September 1998.

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

[RFC3390] M. Allman, S. Floyd, and C. Partridge, Increasing TCP's Initial Window, RFC 3390, October 2002.

10. Informative References

[ECN+] A. Kuzmanovic, The Power of Explicit Congestion Notification, SIGCOMM 2005.

[PI] C. Hollot, V. Misra, W. Gong, and D. Towsley, On Designing Improved Controllers for AQM Routers Supporting TCP Flows, INFOCOM, June 2001.

[RED] S. Floyd and V. Jacobson, Random Early Detection Gateways for Congestion Avoidance, IEEE/ACM Transactions on Networking, V.1, N.4, 1993.

[REM] S. Athuraliya, V. Li, S. Low, and Q Yin, REM: Active Queue Management, IEEE Network, V.15, N. 3, May 2001.

[RFC2988] V. Paxson and M. Allman, Computing TCP's Retransmission Timer, RFC 2988, November 2000.

[RFC3042] M. Allman, H. Balakrishnan, and S. Floyd, Enhancing TCP's Loss Recovery Using Limited Transmit, RFC 3042, Proposed Standard, January 2001.

[RFC3360] S. Floyd, Inappropriate TCP Resets Considered Harmful, RFC 3360, August 2002.

[SCJO01] F. Smith, F. Campos, K. Jeffay, D. Ott, What {TCP/IP} Protocol Headers Can Tell us about the Web, SIGMETRICS, June 2001.

[Tools] S. Floyd and E. Kohler, Tools for the Evaluation of Simulation and Testbed Scenarios, Internet-draft draft-irtf-tmrg-tools-00, work in progress, September 2005.

11. IANA Considerations

There are no IANA considerations regarding this document.

AUTHORS' ADDRESSES

Aleksandar Kuzmanovic
Phone: +1 (847) 467-5519
Northwestern University
Email: akuzma@northwestern.edu
URL: http://cs.northwestern.edu/~a

Sally Floyd
Phone: +1 (510) 666-2989
ICIR (ICSI Center for Internet Research)
Email: floyd@icir.org
URL: http://www.icir.org/floyd/

K. K. Ramakrishnan
Phone: +1 (973) 360-8764
AT&T Labs Research
Email: kkrama@research.att.com
URL: http://www.research.att.com/info/kkrama

Full Copyright Statement

Intellectual Property