RACK: a time-based fast loss detection algorithm for TCP
draft-cheng-tcpm-rack-00

Abstract

This document presents a new TCP loss detection algorithm called RACK ("Recent ACKnowledgment"). RACK uses the notion of time, instead of packet or sequence counts, to detect losses, for modern TCP implementations that can support per-packet timestamps and the selective acknowledgment (SACK) option. It is intended to replace the conventional DUPACK threshold approach and its variants, as well as other nonstandard approaches.

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 http://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 April 21, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of
the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

1. Introduction

This document presents a new loss detection algorithm called RACK ("Recent ACKnowledgment"). RACK uses the notion of time instead of the conventional packet or sequence counting approaches for detecting losses. RACK deems a packet lost if some packet sent sufficiently later has been cumulatively or selectively acknowledged. It does this by recording packet transmission times and inferring losses using cumulative acknowledgments or selective acknowledgment (SACK) TCP options.

The main motivation for RACK is to replace both the standard and nonstandard loss detection algorithms [RFC5681][RFC6675][RFC5827][RFC4653][FACK][THIN-STREAM] to simplify TCP development.

Another motivation is to improve loss detection for modern traffic patterns and underlying network changes. First, the prevalence of interactive request-response traffic means TCP is often application-limited. Second, wide deployment of traffic policers results in frequent lost retransmissions and losses at the tail of transactions. Third, mobile wireless and router load-balancing cause frequent occurrences of small degrees of reordering.

These three factors together make existing packet or sequence counting approaches inefficient. This is because mechanisms based purely on counting packets in sequence order can either detect loss quickly or accurately, but it is hard to achieve both, especially when the sender is application-limited and reordering is unpredictable. And under these conditions none of them can detect lost retransmission well.

2. Requirements

The reader is expected to be familiar with the definitions given in the TCP congestion control [RFC5681] and selective acknowledgment [RFC2018] RFCs. Familiarity with the conservative SACK-based recovery for TCP [RFC6675] is not expected but helps.

RACK has three requirements:

1. The connection MUST use selective acknowledgment (SACK) options [RFC2018].
2. For each packet sent, the sender MUST store its most recent transmission time with (at least) millisecond granularity. For round-trip times lower than a millisecond (e.g., intra-datacenter communications) microsecond granularity would significantly help the detection latency but is not required.

3. For each packet sent, the sender MUST store whether the packet has been retransmitted or not.

We assume that requirement 1 implies the sender keeps a SACK scoreboard, which is a data structure to store selective acknowledgment information on a per-connection basis. For the ease of explaining the algorithm, we use a pseudo-scoreboard that manages the data in sequence number ranges. But the specifics of the data structure are left to the implementor.

RACK does not need any change on the receiver.

3. Definitions of variables

A sender needs to store these new RACK variables:

"Packet.xmit_time" is the time of the last transmission of a data packet, including any retransmissions, if any. The sender needs to record the transmission time for each packet sent and not yet acknowledged. The time MUST be stored at millisecond granularity or finer.

"RACK.xmit_time" is the most recent Packet.xmit_time among all the packets that were delivered (either cumulatively acknowledged or selectively acknowledged) on the connection.

"RACK.RTT" is the associated RTT measured when RACK.xmit_time, above, was changed. It is the RTT of the most recently transmitted packet that has been delivered (either cumulatively acknowledged or selectively acknowledged) on the connection.

"RACK.reoWnd" is a reordering window for the connection, computed in the unit of time used for recording packet transmission times. It is used to defer the moment at which RACK marks a packet lost.

"RACK.min_RTT" is the estimated minimum round-trip time (RTT) of the connection.

Note that the Packet.xmit_time variable is per packet in flight. The RACK.xmit_time, RACK.RTT, RACK.reoWnd, and RACK.min_RTT variables are per connection.
4. Algorithm Details

4.1. Transmitting a data packet

Upon transmitting or retransmitting a packet, record the time in Packet.xmit_time.

4.2. Upon receiving an ACK

Step 1: Update RACK.min_RTT.

Use the RTT measurements obtained in [RFC6298] or [RFC7323] to update the estimated minimum RTT in RACK.min_RTT. The sender can track a simple global minimum of all RTT measurements from the connection, or a windowed min-filtered value of recent RTT measurements. This document does not specify an exact approach.

Step 2: Update RACK.reo_wnd.

To handle the prevalent small degree of reordering, RACK.reo_wnd serves as an allowance for settling time before marking a packet lost. By default it is 1 millisecond. We RECOMMEND implementing the reordering detection in [REORDER-DETECT][RFC4737] to dynamically adjust the reordering window. When the sender detects packet reordering RACK.reo_wnd MAY be changed to RACK.min_RTT/4. We discuss more about the reordering window in the next section.

Step 3: Advance RACK.xmit_time and update RACK.RTT.

Given the information provided in an ACK, each packet cumulatively ACKed or SACKed is marked as delivered in the scoreboard. Among all the packets ACKed or SACKed so far in the connection, record the most recent Packet.xmit_time in RACK.xmit_time if it is ahead of RACK.xmit_time, unless the retransmission is considered as likely spurious by the following check. Ignore the packet if it has been retransmitted and either of two condition is true:

1. The Timestamp Echo Reply field (TSecr) of the ACK’s timestamp option [RFC7323], if available, indicates the ACK was not acknowledging the last retransmission of the packet.

2. The packet was last retransmitted less than RACK.min_rtt ago. While it is still possible the packet is spuriously retransmitted because of a recent RTT decrease, we believe that our experience suggests this is a reasonable heuristic.

If this ACK causes a change to RACK.xmit_time then record the RTT implied by this ACK: set RACK.RTT = now - RACK.xmit_time.
Exit here and omit step 3 if RACK.xmit_time has not changed.

Step 3: Detect losses.

For each packet that has not been fully SACKed, if RACK.xmit_time is after Packet.xmit_time + RACK.reo_wnd, then mark the packet (or its corresponding sequence range) lost in the scoreboard. The rationale is that if another packet that was sent later has been delivered, and the reordering window or "reordering settling time" has already passed, the packet was likely lost.

If a packet that was sent later has been delivered, but the reordering window has not passed, then it is not yet safe to deem the given packet lost. Using the basic algorithm above, the sender would wait for the next ACK to further advance RACK.xmit_time; but this risks a timeout (RTO) if no more ACKs come back (e.g., due to losses or application limit). For timely loss detection, the sender MAY install a "reordering settling" timer set to fire at the earliest moment at which it is safe to conclude that some packet is lost. The earliest moment is the time it takes to expire the reordering window of the earliest unacked packet in flight, which is the minimum value of (Packet.xmit_time + RACK.RTT + RACK.reo_wnd + 1ms) across all unacknowledged packets.

This timer expiration value can be derived as follows. As a starting point, we consider that the reordering window has passed if the RACK packet was sent sufficiently after the packet in question, or a sufficient time has elapsed since the RACK packet was S/ACKed, or some combination of the two. More precisely, RACK marks a packet as lost if the reordering window for a packet has elapsed through the sum of:

1. delta in transmit time between a packet and the RACK packet
2. delta in time between the S/ACK of the RACK packet (RACK.sacked_time) and now

So we mark a packet as lost if:

RACK.xmit_time > Packet.xmit_time
AND
(RACK.xmit_time - Packet.xmit_time) + (now - RACK.sacked_time) > RACK.reo_wnd

If we solve this second condition for "now", the moment at which we can declare a packet lost, then we get:

now > Packet.xmit_time + RACK.reo_wnd + (RACK.sacked_time - RACK.xmit_time)
Then \((RACK.\text{sacked\_time} - RACK.\text{xmit\_time})\) is just the RTT of the packet we used to set \(RACK.\text{xmit\_time}\), so this reduces to:

\[
\text{now} > \text{Packet.\text{xmit\_time}} + RACK.\text{RTT} + RACK.\text{reo\_wnd}
\]

The following pseudocode implements the algorithm above. When an ACK is received or the RACK timer expires, call \(RACK\_\text{detect\_loss()}\):

\[
\text{RACK\_detect\_loss():}
\begin{align*}
\text{min\_timeout} &= 0 \\
\text{For each packet, Packet, in the scoreboard:} \\
& \quad \text{If Packet is already SACKed, ACKed,} \\
& \quad \quad \text{or marked lost and not yet retransmitted:} \\
& \quad \quad \quad \text{Skip to the next packet} \\
& \quad \text{If Packet.\text{xmit\_time} > RACK.\text{xmit\_time}:} \\
& \quad \quad \text{Skip to the next packet} \\
& \quad \text{timeout = Packet.\text{xmit\_time} + RACK.\text{RTT} + RACK.\text{reo\_wnd} + 1} \\
& \quad \text{If now >= timeout} \\
& \quad \quad \text{Mark Packet lost} \\
& \quad \text{Else If (min\_timeout == 0) or (timeout is before min\_timeout):} \\
& \quad \quad \text{min\_timeout = timeout} \\
& \text{If min\_timeout != 0} \\
& \quad \text{Arm the RACK timer to call \(RACK\_\text{detect\_loss()}\) at the time min\_timeout}
\end{align*}
\]

5. Algorithm Analysis

5.1. Advantages

The biggest advantage of RACK is that every data packet, whether it is an original data transmission or a retransmission, can be used to detect losses of the packets sent prior to it.

Example: tail drop. Consider a sender that transmits a window of three data packets (P1, P2, P3), and P1 and P3 are lost. Suppose the transmission of each packet is at least RACK.\text{reo\_wnd} (1 millisecond by default) after the transmission of the previous packet. RACK will mark P1 as lost when the SACK of P2 is received, and this will trigger the retransmission of P1 as R1. When R1 is cumulatively acknowledged, RACK will mark P3 as lost and the sender will retransmit P3 as R3. This example illustrates how RACK is able to repair certain drops at the tail of a transaction without any timer. Notice that neither the conventional duplicate ACK threshold [RFC5681], nor [RFC6675], nor the Forward Acknowledgment [FACK]
algorithm can detect such losses, because of the required packet or sequence count.

Example: lost retransmit. Consider a window of three data packets (P1, P2, P3) that are sent; P1 and P2 are dropped. Suppose the transmission of each packet is at least RACK.reo_wnd (1 millisecond by default) after the transmission of the previous packet. When P3 is SACKed, RACK will mark P1 and P2 lost and they will be retransmitted as R1 and R2. Suppose R1 is lost again (as a tail drop) but R2 is SACKed; RACK will mark R1 lost for retransmission again. Again, neither the conventional three duplicate ACK threshold approach, nor [RFC6675], nor the Forward Acknowledgment [FACK] algorithm can detect such losses. And such a lost retransmission is very common when TCP is being rate-limited, particularly by token bucket policers with large bucket depth and low rate limit. Retransmissions are often lost repeatedly because standard congestion control requires multiple round trips to reduce the rate below the policed rate.

Example: (small) degree of reordering. Consider a common reordering event: a window of packets are sent as (P1, P2, P3). P1 and P2 carry a full payload of MSS octets, but P3 has only a 1-octet payload due to application-limited behavior. Suppose the sender has detected reordering previously (e.g., by implementing the algorithm in [REORDER-DETECT]) and thus RACK.reo_wnd is min_RTT/4. Now P3 is reordered and delivered first, before P1 and P2. As long as P1 and P2 are delivered within min_RTT/4, RACK will not consider P1 and P2 lost. But if P1 and P2 are delivered outside the reordering window, then RACK will still falsely mark P1 and P2 lost. We discuss how to reduce the false positives in the end of this section.

The examples above show that RACK is particularly useful when the sender is limited by the application, which is common for interactive, request/response traffic. Similarly, RACK still works when the sender is limited by the receive window, which is common for applications that use the receive window to throttle the sender.

5.2. Disadvantages

RACK requires the sender to record the transmission time of each packet sent at a clock granularity of one millisecond or finer. TCP implementations that record this already for RTT estimation do not require any new per-packet state. But implementations that are not yet recording packet transmission times will need to add per-packet internal state (commonly either 4 or 8 octets per packet) to track transmission times. In contrast, the conventional approach requires one variable to track number of duplicate ACK threshold.
5.3. Adjusting the reordering window

RACK uses a reordering window of \( \text{min}_\text{rtt} / 4 \). It uses the minimum RTT to accommodate reordering introduced by packets traversing slightly different paths (e.g., router-based parallelism schemes) or out-of-order deliveries in the lower link layer (e.g., wireless links using link-layer retransmission). Alternatively, RACK can use the smoothed RTT used in RTT estimation [RFC6298]. However, smoothed RTT can be significantly inflated by orders of magnitude due to congestion and buffer-bloat, which would result in an overly conservative reordering window and slow loss detection. Furthermore, RACK uses a quarter of minimum RTT because Linux TCP uses the same factor in its implementation to delay Early Retransmit [RFC5827] to reduce spurious loss detections in the presence of reordering, and experience shows that this seems to work reasonably well.

One potential improvement is to further adapt the reordering window by measuring the degree of reordering in time, instead of packet distances. But that requires storing the delivery timestamp of each packet. Some scoreboard implementations currently merge SACKed packets together to support TSO (TCP Segmentation Offload) for faster scoreboard indexing. Supporting per-packet delivery timestamps is difficult in such implementations. However, we acknowledge that the current metric can be improved by further research.

5.4. Relationships with other loss recovery algorithms

The primary motivation of RACK is to ultimately provide a simple and general replacement for some of the standard loss recovery algorithms [RFC5681][RFC6675][RFC5827][RFC4653] and nonstandard ones [FACK][THIN-STREAM]. While RACK can be a supplemental loss detection on top of these algorithms, this is not necessary, because the RACK implicitly subsumes most of them.

~~[RFC5827][RFC4653][THIN-STREAM] dynamically adjusts the duplicate ACK threshold based on the current or previous flight sizes. RACK takes a different approach, by using only one ACK event and a reordering window. RACK can be seen as an extended Early Retransmit [RFC5827] without a FlightSize limit but with an additional reordering window. [FACK] considers an original packet to be lost when its sequence range is sufficiently far below the highest SACKed sequence. In some sense RACK can be seen as a generalized form of FACK that operates in time space instead of sequence space, enabling it to better handle reordering, application-limited traffic, and lost retransmissions.~~

Neverthelss RACK is still an experimental algorithm. Since the oldest loss detection algorithm, the 3 duplicate ACK threshold [RFC5681], has been standardized and widely deployed, we RECOMMEND
TCP implementations use both RACK and the algorithm specified in Section 3.2 in [RFC5681] for compatibility.

RACK is compatible with and does not interfere with the the standard RTO [RFC6298], RTO-restart [RTO-RESTART], F-RTO [RFC5682] and Eifel algorithms [RFC3522]. This is because RACK only detects loss by using ACK events. It neither changes the timer calculation nor detects spurious timeouts.

Furthermore, RACK naturally works well with Tail Loss Probe [TLP] because a tail loss probe solicits either an ACK or SACK, which can be used by RACK to detect more losses. RACK can be used to relax TLP’s requirement for using FACK and retransmitting the highest-sequenced packet, because RACK is agnostic to packet sequence numbers, and uses transmission time instead. Thus TLP can be modified to retransmit the first unacknowledged packet, which can improve application latency.

5.5. Interaction with congestion control

RACK intentionally decouples loss detection from congestion control. RACK only detects losses; it does not modify the congestion control algorithm [RFC5681][RFC6937]. However, RACK may detect losses earlier or later than the conventional duplicate ACK threshold approach does. A packet marked lost by RACK SHOULD NOT be retransmitted until congestion control deems this appropriate (e.g. using [RFC6937]).

RACK is applicable for both fast recovery and recovery after a retransmission timeout (RTO) in [RFC5681]. The distinction between fast recovery or RTO recovery is not necessary because RACK is purely based on the transmission time order of packets. When a packet retransmitted by RTO is acknowledged, RACK will mark any unacked packet sent sufficiently prior to the RTO as lost, because at least one RTT has elapsed since these packets were sent.

6. Security Considerations

RACK does not change the risk profile for TCP.

An interesting scenario is ACK-splitting attacks [SCWA99]: for an MSS-size packet sent, the receiver or the attacker might send MSS ACKs that SACK or acknowledge one additional byte per ACK. This would not fool RACK. RACK.xmit_time would not advance because all the sequences of the packet are transmitted at the same time (carry the same transmission timestamp). In other words, SACKing only one byte of a packet or SACKing the packet in entirety have the same effect on RACK.
7. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

8. Acknowledgments

The authors wish to thank Matt Mathis for his insights in FACK that inspired this work. Nandita Dukkipati, Eric Dumazet, Van Jacobson, Ian Swett, and Jana Iyengar contributed to the algorithm and the implementations in TCP and QUIC. We thank the authors of RFC3517 for a great document on an excellent loss recovery algorithm that this draft is trying to improve upon.

9. References

9.1. Normative References


9.2. Informative References


Authors' Addresses

Yuchung Cheng
Google, Inc
1600 Amphitheater Parkway
Mountain View, California  93117
USA

Email: ycheng@google.com

Neal Cardwell
Google, Inc
76 Ninth Avenue
New York, NY  10011
USA

Email: ncardwell@google.com
Abstract

This informational memo describes Datacenter TCP (DCTCP), an improvement to TCP congestion control for datacenter traffic. DCTCP uses improved Explicit Congestion Notification (ECN) processing to estimate the fraction of bytes that encounter congestion, rather than simply detecting that some congestion has occurred. DCTCP then scales the TCP congestion window based on this estimate. This method achieves high burst tolerance, low latency, and high throughput with shallow-buffered switches. This memo also discusses deployment issues related to the coexistence of DCTCP and conventional TCP, the lack of a negotiating mechanism between sender and receiver, and presents some possible mitigations.

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 http://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 May 5, 2016.
1. Introduction

Large datacenters necessarily need many network switches to interconnect its many servers. Therefore, a datacenter can greatly reduce its capital expenditure by leveraging low-cost switches. However, such low-cost switches tend to have limited queue capacities and are thus more susceptible to packet loss due to congestion.

Network traffic in a datacenter is often a mix of short and long flows, where the short flows require low latencies and the long flows require high throughputs. Datacenters also experience incast bursts, where many servers send traffic to a single server at the same time.
For example, this traffic pattern is a natural consequence of MapReduce workload: The worker nodes complete at approximately the same time, and all reply to the master node concurrently.

These factors place some conflicting demands on the queue occupancy of a switch:

- The queue must be short enough that it does not impose excessive latency on short flows.
- The queue must be long enough to buffer sufficient data for the long flows to saturate the path capacity.
- The queue must be short enough to absorb incast bursts without excessive packet loss.

Standard TCP congestion control [RFC5681] relies on packet loss to detect congestion. This does not meet the demands described above. First, short flows will start to experience unacceptable latencies before packet loss occurs. Second, by the time TCP congestion control kicks in on the senders, most of the incast burst has already been dropped.

[RFC3168] describes a mechanism for using Explicit Congestion Notification (ECN) from the switches for early detection of congestion, rather than waiting for packet loss to occur. However, this method only detects the presence of congestion, not its extent. In the presence of mild congestion, the TCP congestion window is reduced too aggressively and this unnecessarily reduces the throughput of long flows.

Datacenter TCP (DCTCP) improves traditional ECN processing by estimating the fraction of bytes that encounter congestion, rather than simply detecting that some congestion has occurred. DCTCP then scales the TCP congestion window based on this estimate. This method achieves high burst tolerance, low latency, and high throughput with shallow-buffered switches.

It is recommended that DCTCP be deployed in a datacenter environment where the endpoints and the switching fabric are under a single administrative domain. This protocol is not meant for uncontrolled deployment in the global Internet.

2. Terminology

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 [RFC2119].
3. DCTCP Algorithm

There are three components involved in the DCTCP algorithm:

- The switches (or other intermediate devices in the network) detect congestion and set the Congestion Encountered (CE) codepoint in the IP header.
- The receiver echoes the congestion information back to the sender, using the ECN-Echo (ECE) flag in the TCP header.
- The sender computes a congestion estimate and reacts, by reducing the TCP congestion window accordingly (cwnd).

3.1. Marking Congestion on the Switches

The switches in a datacenter fabric indicate congestion to the end nodes by setting the CE codepoint in the IP header as specified in Section 5 of [RFC3168]. For example, the switches may be configured with a congestion threshold. When a packet arrives at a switch and its queue length is greater than the congestion threshold, the switch sets the CE codepoint in the packet. For example, Section 3.4 of [DCTCP10] suggests threshold marking with a threshold \( K > (\text{RTT} \times C)/7 \), where \( C \) is the sending rate in packets per second. However, the actual algorithm for marking congestion is an implementation detail of the switch and will generally not be known to the sender and receiver. Therefore, sender and receiver MUST NOT assume that a particular marking algorithm is implemented by the switching fabric.

3.2. Echoing Congestion Information on the Receiver

According to Section 6.1.3 of [RFC3168], the receiver sets the ECE flag if any of the packets being acknowledged had the CE code point set. The receiver then continues to set the ECE flag until it receives a packet with the Congestion Window Reduced (CWR) flag set. However, the DCTCP algorithm requires more detailed congestion information. In particular, the sender must be able to determine the number of bytes sent that encountered congestion. Thus, the scheme described in [RFC3168] does not suffice.

One possible solution is to ACK every packet and set the ECE flag in the ACK if and only if the CE code point was set in the packet being acknowledged. However, this prevents the use of delayed ACKs, which are an important performance optimization in datacenters.

Instead, DCTCP introduces a new Boolean TCP state variable, "DCTCP Congestion Encountered" (DCTCP.CE), which is initialized to false and stored in the Transmission Control Block (TCB). When sending an ACK,
the ECE flag MUST be set if and only if DCTCP.CE is true. When receiving packets, the CE codepoint MUST be processed as follows:

1. If the CE codepoint is set and DCTCP.CE is false, send an ACK for any previously unacknowledged packets and set DCTCP.CE to true.

2. If the CE codepoint is not set and DCTCP.CE is true, send an ACK for any previously unacknowledged packets and set DCTCP.CE to false.

3. Otherwise, ignore the CE codepoint.

The handling of the "Congestion Window Reduced" (CWR) bit is also exactly as per [RFC3168] including [RFC3168-ERRATA3639]. That is, on receipt of a segment with both the CE and CWR bits set, CWR is processed first and then ECE is processed.

```
<table>
<thead>
<tr>
<th>Send immediate ACK with ECE=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Send 1 ACK</td>
</tr>
<tr>
<td>for every m packets with ECE=0</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
chosen to be approximately the Round Trip Time (RTT). In particular, an observation window ends when all bytes in flight at the beginning of the window have been acknowledged.

In order to update DCTCP.Alpha, the TCP state variables defined in [RFC0793] are used, and three additional TCP state variables are introduced:

- DCTCP.WindowEnd: The TCP sequence number threshold for beginning a new observation window; initialized to SND.UNA.
- DCTCP.BytesSent: The number of bytes sent during the current observation window; initialized to zero.
- DCTCP.BytesMarked: The number of bytes sent during the current observation window that encountered congestion; initialized to zero.

The congestion estimator on the sender MUST process acceptable ACKs as follows:

1. Compute the bytes acknowledged (TCP SACK options [RFC2018] are ignored):
   \[\text{Bytes Acked} = \text{SEG.ACK} - \text{SND.UNA}\]

2. Update the bytes sent:
   \[\text{DCTCP.BytesSent} += \text{Bytes Acked}\]

3. If the ECE flag is set, update the bytes marked:
   \[\text{DCTCP.BytesMarked} += \text{Bytes Acked}\]

4. If the acknowledgment number is less than or equal to DCTCP.WindowEnd, stop processing. Otherwise, the end of the observation window has been reached, so proceed to update the congestion estimate as follows:

5. Compute the congestion level for the current observation window:
   \[M = \frac{\text{DCTCP.BytesMarked}}{\text{DCTCP.BytesSent}}\]

6. Update the congestion estimate:
   \[\text{DCTCP.Alpha} = \text{DCTCP.Alpha} \times (1 - g) + g \times M\]

7. Determine the end of the next observation window:
DCTCP.WindowEnd = SND.NXT

8. Reset the byte counters:

DCTCP.BytesSent = DCTCP.BytesMarked = 0

Rather than always halving the congestion window as described in [RFC3168], when the sender receives an indication of congestion (ECE), the sender MUST update cwnd as follows:

\[
cwnd = cwnd \times (1 - \text{DCTCP.Alpha} / 2)
\]

Thus, when no bytes sent experienced congestion, DCTCP.Alpha equals zero, and cwnd is left unchanged. When all sent bytes experienced congestion, DCTCP.Alpha equals one, and cwnd is reduced by half. Lower levels of congestion will result in correspondingly smaller reductions to cwnd.

Just as specified in [RFC3168], TCP should not react to congestion indications more than once for every window of data. The setting of the "Congestion Window Reduced" (CWR) bit is also exactly as per [RFC3168].

3.4. Handling of SYN, SYN-ACK, RST Packets

[RFC3168] requires that a compliant TCP MUST NOT set ECT on SYN or SYN-ACK packets. [RFC5562] proposes setting ECT on SYN-ACK packets, but maintains the restriction of no ECT on SYN packets. Both these RFCs prohibit ECT in SYN packets due to security concerns regarding malicious SYN packets with ECT set. These RFCs, however, are intended for general Internet use, and do not directly apply to a controlled datacenter environment. The switching fabric can drop TCP packets that do not have the ECT set in the IP header. If SYN and SYN-ACK packets for DCTCP connections do not have ECT set, they will be dropped with high probability. For DCTCP connections, the sender SHOULD set ECT for SYN, SYN-ACK and RST packets.

4. Implementation Issues

As noted in Section 3.3, the implementation MUST choose a suitable estimation gain. [DCTCP10] provides a theoretical basis for selecting the gain. However, it may be more practical to use experimentation to select a suitable gain for a particular network and workload. The Microsoft implementation of DCTCP in Windows Server 2012 uses a fixed estimation gain of 1/16.

The implementation must also decide when to use DCTCP. Datacenter servers may need to communicate with endpoints outside the
datacenter, where DCTCP is unsuitable or unsupported. Thus, a global configuration setting to enable DCTCP will generally not suffice. DCTCP provides no mechanism for negotiating its use. Thus, there is additional management and configuration overhead required to ensure that DCTCP is not used with non-DCTCP endpoints.

Potential solutions rely on either configuration or heuristics. Heuristics need to allow endpoints to individually enable DCTCP, to ensure a DCTCP sender is always paired with a DCTCP receiver. One approach is to enable DCTCP based on the IP address of the remote endpoint. Another approach is to detect connections that transmit within the bounds a datacenter. For example, Microsoft Windows Server 2012 (and later versions) supports automatic selection of DCTCP if the estimated RTT is less than 10 msec and ECN is successfully negotiated, under the assumption that if the RTT is low, then the two endpoints are likely in the same datacenter network.

It is RECOMMENDED that an implementation deal with loss episodes in the same way as conventional TCP. In case of a timeout or fast retransmit or any change in delay (for delay based congestion control), the cwnd and other state variables like ssthresh must be changed in the same way that a conventional TCP would have changed them. It would be useful to implement DCTCP as additional actions on top of an existing congestion control algorithm like NewReno. The DCTCP implementation MAY also allow configuration of resetting the value of DCTCP.Alpha as part of processing any loss episodes.

To prevent incast throughput collapse, the minimum RTO (MinRTO) used by TCP should be lowered significantly. The default value of MinRTO in Windows is 300 msec, which is much greater than the maximum latencies inside a datacenter. In Microsoft Windows Server 2012 (and later), the MinRTO value is configurable, allowing values as low as 10 msec on a per-subnet or per-port basis (or even globally.) A lower MinRTO value requires a correspondingly lower delayed ACK timeout on the receiver. It is RECOMMENDED that an implementation allow configuration of lower timeouts for DCTCP connections.

In the same vein, it is also RECOMMENDED that an implementation allow configuration of restarting the congestion window (cwnd) of idle DCTCP connections as described in [RFC5681], since network conditions can change rapidly in datacenters.

[RFC3168] forbids the ECN-marking of pure ACK packets, because of the inability of TCP to mitigate ACK-path congestion and protocol-wise preferential treatment by routers. However, dropping pure ACKs – rather than ECN marking them – has disadvantages for typical datacenter traffic patterns. Because of the prevalence of bursty traffic patterns that feature transient congestion, dropping of ACKs...
causes subsequent retransmissions. It is RECOMMENDED that an implementation provide a configuration knob that forces ECT to be set on pure ACKs.

The DCTCP.Alpha calculation as per the formula in Section 3.3 involves fractions. An efficient kernel implementation MAY scale the DCTCP.Alpha value for efficient computation using shift operations. For example, if the implementation chooses \( g = 1/16 \), multiplications of DCTCP.Alpha by \( g \) become right-shifts by 4. A scaling implementation SHOULD ensure that DCTCP.Alpha is able to reach zero once it falls below the smallest shifted value (16 in the above example). At the other extreme, a scaled update MUST also ensure DCTCP.Alpha does not exceed the scaling factor, which would be equivalent to greater than 100% congestion. So, DCTCP.Alpha MUST be clamped after an update.

This results in the following computations replacing steps 5 and 6 in Section 3.3, where SCF is the chosen scaling factor (65536 in the example) and SHF is the shift factor (4 in the example):

1. Compute the congestion level for the current observation window:
   \[
   \text{ScaledM} = \text{SCF} \times \frac{\text{DCTCP.BytesMarked}}{\text{DCTCP.BytesSent}}
   \]

2. Update the congestion estimate:
   \[
   \text{if (DCTCP.Alpha >> SHF) == 0 then DCTCP.Alpha = 0}
   \]
   \[
   \text{DCTCP.Alpha += (ScaledM >> SHF) - (DCTCP.Alpha >> SHF)}
   \]
   \[
   \text{if DCTCP.Alpha > SCF then DCTCP.Alpha = SCF}
   \]

5. Deployment Issues

DCTCP and conventional TCP congestion control do not coexist well in the same network. In DCTCP, the marking threshold is set to a very low value to reduce queueing delay, and a relatively small amount of congestion will exceed the marking threshold. During such periods of congestion, conventional TCP will suffer packet loss and quickly and drastically reduce cwnd. DCTCP, on the other hand, will use the fraction of marked packets to reduce cwnd more gradually. Thus, the rate reduction in DCTCP will be much slower than that of conventional TCP, and DCTCP traffic will gain a larger share of the capacity compared to conventional TCP traffic traversing the same path. If the traffic in the datacenter is a mix of conventional TCP and DCTCP, it is RECOMMENDED that DCTCP traffic be segregated from conventional TCP traffic. [MORGANSTANLEY] describes a deployment that uses the IP
DSCP bits to segregate the network such that AQM is applied to DCTCP traffic, whereas TCP traffic is managed via drop-tail queueing.

Today’s commodity switches allow configuration of different marking/drop profiles for non-TCP and non-IP packets. Non-TCP and non-IP packets should be able to pass through such switches, unless they really run out of buffer space. If the datacenter traffic consists of such traffic (including UDP), one possible mitigation would be to mark IP packets as ECT even when there is no transport that is reacting to the marking.

Since DCTCP relies on congestion marking by the switches, DCTCP can only be deployed in datacenters where the entire network infrastructure supports ECN. The switches may also support configuration of the congestion threshold used for marking. The proposed parameterization can be configured with switches that implement RED. [DCTCP10] provides a theoretical basis for selecting the congestion threshold, but as with the estimation gain, it may be more practical to rely on experimentation or simply to use the default configuration of the device. DCTCP will degrade to loss-based congestion control when transiting a congested drop-tail link.

DCTCP requires changes on both the sender and the receiver, so both endpoints must support DCTCP. Furthermore, DCTCP provides no mechanism for negotiating its use, so both endpoints must be configured through some out-of-band mechanism to use DCTCP. A variant of DCTCP that can be deployed unilaterally and only requires standard ECN behavior has been described in [ODCTCP][BSDCAN], but requires additional experimental evaluation.

6. Known Issues

DCTCP relies on the sender’s ability to reconstruct the stream of CE codepoints received by the remote endpoint. To accomplish this, DCTCP avoids using a single ACK packet to acknowledge segments received both with and without the CE codepoint set. However, if one or more ACK packets are dropped, it is possible that a subsequent ACK will cumulatively acknowledge a mix of CE and non-CE segments. This will, of course, result in a less accurate congestion estimate. There are some potential considerations:

- Even with an inaccurate congestion estimate, DCTCP may still perform better than [RFC3168].
- If the estimation gain is small relative to the packet loss rate, the estimate may not be too inaccurate.
If packet loss mostly occurs under heavy congestion, most drops will occur during an unbroken string of CE packets, and the estimate will be unaffected.

However, the effect of packet drops on DCTCP under real world conditions has not been analyzed.

DCTCP provides no mechanism for negotiating its use. The effect of using DCTCP with a standard ECN endpoint has been analyzed in [ODCTCP][BSDCAN]. Furthermore, it is possible that other implementations may also modify [RFC3168] behavior without negotiation, causing further interoperability issues.

Much like standard TCP, DCTCP is biased against flows with longer RTTs. A method for improving the fairness of DCTCP has been proposed in [ADCTCP], but requires additional experimental evaluation.

7. Implementation Status

This section documents the implementation status of the specification in this document, as recommended by [RFC6982].

This document describes DCTCP as implemented in Microsoft Windows Server 2012. Since publication of the first versions of this document, the Linux [LINUX] and FreeBSD [FREEBSD] operating systems have also implemented support for DCTCP in a way that is believed to follow this document.

8. Security Considerations

DCTCP enhances ECN and thus inherits the security considerations discussed in [RFC3168]. The processing changes introduced by DCTCP do not exacerbate these considerations or introduce new ones. In particular, with either algorithm, the network infrastructure or the remote endpoint can falsely report congestion and thus cause the sender to reduce cwnd. However, this is no worse than what can be achieved by simply dropping packets.

9. IANA Considerations

This document has no actions for IANA.

10. Acknowledgements

The DCTCP algorithm was originally proposed and analyzed in [DCTCP10] by Mohammad Alizadeh, Albert Greenberg, Dave Maltz, Jitu Padhye, Parveen Patel, Balaji Prabhakar, Sudipta Sengupta, and Murari Sridharan.

We would like to thank Andrew Shewmaker for identifying the problem of clamping DCTCP.Alpha and proposing a solution for it.

Lars Eggert has received funding from the European Union’s Horizon 2020 research and innovation program 2014–2018 under grant agreement No. 644866 ("SSICLOPS"). This document reflects only the authors’ views and the European Commission is not responsible for any use that may be made of the information it contains.

11. References

11.1. Normative References


11.2. Informative References


Authors’ Addresses

Stephen Bensley
Microsoft
One Microsoft Way
Redmond, WA  98052
USA

Phone: +1 425 703 5570
Email: sbens@microsoft.com

Lars Eggert
NetApp
Sonnenallee 1
Kirchheim  85551
Germany

Phone: +49 151 120 55791
Email: lars@netapp.com
URI:   http://eggert.org/

Dave Thaler
Microsoft

Phone: +1 425 703 8835
Email: dthaler@microsoft.com

Praveen Balasubramanian
Microsoft

Phone: +1 425 538 2782
Email: pravb@microsoft.com

Glenn Judd
Morgan Stanley

Phone: +1 973 979 6481
Email: glenn.judd@morganstanley.com
Transmission Control Protocol Specification

draft-ietf-tcpm-rfc793bis-01

Abstract

This document specifies the Internet’s Transmission Control Protocol
(TCP). TCP is an important transport layer protocol in the Internet
stack, and has continuously evolved over decades of use and growth of
the Internet. Over this time, a number of changes have been made to
TCP as it was specified in RFC 793, though these have only been
documented in a piecemeal fashion. This document collects and brings
those changes together with the protocol specification from RFC 793.
This document obsoletes RFC 793 and several other RFCs (TODO: list
all actual RFCs when finished).

RFC EDITOR NOTE: If approved for publication as an RFC, this should
be marked additionally as "STD: 7" and replace RFC 793 in that role.

Requirements Language

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 [3].

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 http://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 March 24, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the
document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal
Provisions Relating to IETF Documents
(http://trustee.ietf.org/license-info) in effect on the date of
publication of this document. Please review these documents
carefully, as they describe your rights and restrictions with respect
to this document. Code Components extracted from this document must
include Simplified BSD License text as described in Section 4.e of
the Trust Legal Provisions and are provided without warranty as
described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF
Contributions published or made publicly available before November
10, 2008. The person(s) controlling the copyright in some of this
material may not have granted the IETF Trust the right to allow
modifications of such material outside the IETF Standards Process.
Without obtaining an adequate license from the person(s) controlling
the copyright in such materials, this document may not be modified
outside the IETF Standards Process, and derivative works of it may
not be created outside the IETF Standards Process, except to format
it for publication as an RFC or to translate it into languages other
than English.

Table of Contents

1. Purpose and Scope .................................. 3
2. Introduction ..................................... 4
3. Functional Specification .......................... 4
  3.1. Header Format ................................ 4
  3.2. Terminology .................................. 9
  3.3. Sequence Numbers .............................. 14
  3.4. Establishing a connection .................... 20
  3.5. Closing a Connection ........................ 27
    3.5.1. Half-Closed Connections ................. 30
  3.6. Precedence and Security ..................... 30
  3.7. Segmentation ................................. 31
    3.7.1. Maximum Segment Size Option ............ 32
    3.7.2. Path MTU Discovery ...................... 33
    3.7.3. Interfaces with Variable MTU Values .... 34
    3.7.4. Nagle Algorithm ........................ 34
    3.7.5. IPv6 Jumbograms ........................ 35
  3.8. Data Communication .......................... 35
1. Purpose and Scope

In 1981, RFC 793 [6] was released, documenting the Transmission Control Protocol (TCP), and replacing earlier specifications for TCP that had been published in the past.

Since then, TCP has been implemented many times, and has been used as a transport protocol for numerous applications on the Internet.

For several decades, RFC 793 plus a number of other documents have combined to serve as the specification for TCP [16]. Over time, a number of errata have been identified on RFC 793, as well as deficiencies in security, performance, and other aspects. A number of enhancements has grown and been documented separately. These were never accumulated together into an update to the base specification.

The purpose of this document is to bring together all of the IETF Standards Track changes that have been made to the basic TCP functional specification and unify them into an update of the RFC 793 protocol specification. Some companion documents are referenced for important algorithms that TCP uses (e.g. for congestion control), but have not been attempted to include in this document. This is a conscious choice, as this base specification can be used with multiple additional algorithms that are developed and incorporated separately, but all TCP implementations need to implement this specification as a common basis in order to interoperate. As some additional TCP features have become quite complicated themselves (e.g. advanced loss recovery and congestion control), future companion documents may attempt to similarly bring these together.

In addition to the protocol specification that describes the TCP segment format, generation, and processing rules that are to be implemented in code, RFC 793 and other updates also contain
informative and descriptive text for human readers to understand aspects of the protocol design and operation. This document does not attempt to alter or update this informative text, and is focused only on updating the normative protocol specification. We preserve references to the documentation containing the important explanations and rationale, where appropriate.

This document is intended to be useful both in checking existing TCP implementations for conformance, as well as in writing new implementations.

2. Introduction

RFC 793 contains a discussion of the TCP design goals and provides examples of its operation, including examples of connection establishment, closing connections, and retransmitting packets to repair losses.

This document describes the basic functionality expected in modern implementations of TCP, and replaces the protocol specification in RFC 793. It does not replicate or attempt to update the examples and other discussion in RFC 793. Other documents are referenced to provide explanation of the theory of operation, rationale, and detailed discussion of design decisions. This document only focuses on the normative behavior of the protocol.

TEMPORARY EDITOR’S NOTE: This is an early revision in the process of updating RFC 793. Many planned changes are not yet incorporated.

***Please do not use this revision as a basis for any work or reference.***

A list of changes from RFC 793 is contained in Section 4.

TEMPORARY EDITOR’S NOTE: the current revision of this document does not yet collect all of the changes that will be in the final version. The set of content changes planned for future revisions is kept in Section 4.

3. Functional Specification

3.1. Header Format

TCP segments are sent as internet datagrams. The Internet Protocol header carries several information fields, including the source and destination host addresses [2]. A TCP header follows the internet header, supplying information specific to the TCP protocol. This
division allows for the existence of host level protocols other than TCP.

TCP Header Format

Note that one tick mark represents one bit position.

Source Port: 16 bits

The source port number.

Destination Port: 16 bits

The destination port number.

Sequence Number: 32 bits

The sequence number of the first data octet in this segment (except when SYN is present). If SYN is present the sequence number is the initial sequence number (ISN) and the first data octet is ISN+1.

Acknowledgment Number: 32 bits
If the ACK control bit is set this field contains the value of the next sequence number the sender of the segment is expecting to receive. Once a connection is established this is always sent.

Data Offset: 4 bits

The number of 32 bit words in the TCP Header. This indicates where the data begins. The TCP header (even one including options) is an integral number of 32 bits long.

Reserved: 4 bits

Reserved for future use. Must be zero.

Control Bits: 8 bits (from left to right):

- CWR: Congestion Window Reduced
- ECE: ECN-Echo
- URG: Urgent Pointer field significant
- ACK: Acknowledgment field significant
- PSH: Push Function
- RST: Reset the connection
- SYN: Synchronize sequence numbers
- FIN: No more data from sender

Window: 16 bits

The number of data octets beginning with the one indicated in the acknowledgment field which the sender of this segment is willing to accept.

The window size MUST be treated as an unsigned number, or else large window sizes will appear like negative windows and TCP will not work. It is RECOMMENDED that implementations will reserve 32-bit fields for the send and receive window sizes in the connection record and do all window computations with 32 bits.

Checksum: 16 bits

The checksum field is the 16 bit one’s complement of the one’s complement sum of all 16 bit words in the header and text. If a segment contains an odd number of header and text octets to be checksummed, the last octet is padded on the right with zeros to form a 16 bit word for checksum purposes. The pad is not transmitted as part of the segment. While computing the checksum, the checksum field itself is replaced with zeros.
The checksum also covers a 96 bit pseudo header conceptually prefixed to the TCP header. This pseudo header contains the Source Address, the Destination Address, the Protocol, and TCP length. This gives the TCP protection against misrouted segments. This information is carried in the Internet Protocol and is transferred across the TCP/Network interface in the arguments or results of calls by the TCP on the IP.

+--------+--------+--------+--------+
|           Source Address          |
+--------+--------+--------+--------+
|         Destination Address       |
+--------+--------+--------+--------+
|  zero  |  PTCL  |    TCP Length   |
+--------+--------+--------+--------+

The TCP Length is the TCP header length plus the data length in octets (this is not an explicitly transmitted quantity, but is computed), and it does not count the 12 octets of the pseudo header.

The TCP checksum is never optional. The sender MUST generate it and the receiver MUST check it.

Urgent Pointer: 16 bits

This field communicates the current value of the urgent pointer as a positive offset from the sequence number in this segment. The urgent pointer points to the sequence number of the octet following the urgent data. This field is only be interpreted in segments with the URG control bit set.

Options: variable

Options may occupy space at the end of the TCP header and are a multiple of 8 bits in length. All options are included in the checksum. An option may begin on any octet boundary. There are two cases for the format of an option:

Case 1: A single octet of option-kind.

Case 2: An octet of option-kind, an octet of option-length, and the actual option-data octets.

The option-length counts the two octets of option-kind and option-length as well as the option-data octets.
Note that the list of options may be shorter than the data offset field might imply. The content of the header beyond the End-of-Option option must be header padding (i.e., zero).

Currently defined options include (kind indicated in octal):

<table>
<thead>
<tr>
<th>Kind</th>
<th>Length</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>End of option list.</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>No-Operation.</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>Maximum Segment Size.</td>
</tr>
</tbody>
</table>

A TCP MUST be able to receive a TCP option in any segment. A TCP MUST ignore without error any TCP option it does not implement, assuming that the option has a length field (all TCP options except End of option list and No-Operation have length fields). TCP MUST be prepared to handle an illegal option length (e.g., zero) without crashing; a suggested procedure is to reset the connection and log the reason.

Specific Option Definitions

End of Option List

\[
\begin{array}{c}
+--------+ \\
|00000000| \\
+--------+ \\
\end{array}
\]

Kind=0

This option code indicates the end of the option list. This might not coincide with the end of the TCP header according to the Data Offset field. This is used at the end of all options, not the end of each option, and need only be used if the end of the options would not otherwise coincide with the end of the TCP header.

No-Operation

\[
\begin{array}{c}
+--------+ \\
|00000001| \\
+--------+ \\
\end{array}
\]

Kind=1

This option code may be used between options, for example, to align the beginning of a subsequent option on a word boundary. There is no guarantee that senders will use this option, so receivers must be prepared to process options even if they do not begin on a word boundary.
Maximum Segment Size (MSS)

+--------+--------+---------+--------+
|00000010|00000100|   max seg size   |
+--------+--------+---------+--------+
Kind=2   Length=4

Maximum Segment Size Option Data: 16 bits

If this option is present, then it communicates the maximum receive segment size at the TCP which sends this segment. This field may be sent in the initial connection request (i.e., in segments with the SYN control bit set) and must not be sent in other segments. If this option is not used, any segment size is allowed. A more complete description of this option is in Section 3.7.1.

Padding: variable

The TCP header padding is used to ensure that the TCP header ends and data begins on a 32 bit boundary. The padding is composed of zeros.

3.2. Terminology

Before we can discuss very much about the operation of the TCP we need to introduce some detailed terminology. The maintenance of a TCP connection requires the remembering of several variables. We conceive of these variables being stored in a connection record called a Transmission Control Block or TCB. Among the variables stored in the TCB are the local and remote socket numbers, the security and precedence of the connection, pointers to the user's send and receive buffers, pointers to the retransmit queue and to the current segment. In addition several variables relating to the send and receive sequence numbers are stored in the TCB.
Send Sequence Variables

SND.UNA - send unacknowledged
SND.NXT - send next
SND.WND - send window
SND.UP - send urgent pointer
SND.WL1 - segment sequence number used for last window update
SND.WL2 - segment acknowledgment number used for last window update
ISS - initial send sequence number

Receive Sequence Variables

RCV.NXT - receive next
RCV.WND - receive window
RCV.UP - receive urgent pointer
IRS - initial receive sequence number

The following diagrams may help to relate some of these variables to the sequence space.

Send Sequence Space

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>SND.UNA</td>
<td>SND.NXT</td>
<td>SND.UNA</td>
<td>+SND.WND</td>
</tr>
</tbody>
</table>

1 - old sequence numbers which have been acknowledged
2 - sequence numbers of unacknowledged data
3 - sequence numbers allowed for new data transmission
4 - future sequence numbers which are not yet allowed

Send Sequence Space

Figure 2

The send window is the portion of the sequence space labeled 3 in Figure 2.
Receive Sequence Space

\[
\begin{array}{c|c|c}
1 & 2 & 3 \\
-\hline
\text{RCV.NXT} & \text{RCV.NXT} & \text{RCV.NXT} \\
\text{+RCV.WND} & & \\
\end{array}
\]

1 - old sequence numbers which have been acknowledged
2 - sequence numbers allowed for new reception
3 - future sequence numbers which are not yet allowed

Receive Sequence Space

Figure 3

The receive window is the portion of the sequence space labeled 2 in Figure 3.

There are also some variables used frequently in the discussion that take their values from the fields of the current segment.

Current Segment Variables

- \text{SEG.SEQ} - segment sequence number
- \text{SEG.ACK} - segment acknowledgment number
- \text{SEG.LEN} - segment length
- \text{SEG.WND} - segment window
- \text{SEG.UP} - segment urgent pointer
- \text{SEG.PRC} - segment precedence value

A connection progresses through a series of states during its lifetime. The states are: \text{LISTEN}, \text{SYN-SENT}, \text{SYN-RECEIVED}, \text{ESTABLISHED}, \text{FIN-WAIT-1}, \text{FIN-WAIT-2}, \text{CLOSE-WAIT}, \text{CLOSING}, \text{LAST-ACK}, \text{TIME-WAIT}, and the fictional state \text{CLOSED}. CLOSED is fictional because it represents the state when there is no TCB, and therefore, no connection. Briefly the meanings of the states are:

\text{LISTEN} - represents waiting for a connection request from any remote TCP and port.

\text{SYN-SENT} - represents waiting for a matching connection request after having sent a connection request.

\text{SYN-RECEIVED} - represents waiting for a confirming connection request acknowledgment after having both received and sent a connection request.
ESTABLISHED - represents an open connection, data received can be delivered to the user. The normal state for the data transfer phase of the connection.

FIN-WAIT-1 - represents waiting for a connection termination request from the remote TCP, or an acknowledgment of the connection termination request previously sent.

FIN-WAIT-2 - represents waiting for a connection termination request from the remote TCP.

CLOSE-WAIT - represents waiting for a connection termination request from the local user.

CLOSING - represents waiting for a connection termination request acknowledgment from the remote TCP.

LAST-ACK - represents waiting for an acknowledgment of the connection termination request previously sent to the remote TCP (this termination request sent to the remote TCP already included an acknowledgment of the termination request sent from the remote TCP).

TIME-WAIT - represents waiting for enough time to pass to be sure the remote TCP received the acknowledgment of its connection termination request.

CLOSED - represents no connection state at all.

A TCP connection progresses from one state to another in response to events. The events are the user calls, OPEN, SEND, RECEIVE, CLOSE, ABORT, and STATUS; the incoming segments, particularly those containing the SYN, ACK, RST and FIN flags; and timeouts.

The state diagram in Figure 4 illustrates only state changes, together with the causing events and resulting actions, but addresses neither error conditions nor actions which are not connected with state changes. In a later section, more detail is offered with respect to the reaction of the TCP to events.

NOTA BENE: this diagram is only a summary and must not be taken as the total specification.
note 1: The transition from SYN-RCVD to LISTEN on receiving a RST is conditional on having reached SYN-RCVD after a passive open.

note 2: An unshown transition exists from FIN-WAIT-1 to TIME-WAIT if a FIN is received and the local FIN is also acknowledged.
3.3. Sequence Numbers

A fundamental notion in the design is that every octet of data sent over a TCP connection has a sequence number. Since every octet is sequenced, each of them can be acknowledged. The acknowledgment mechanism employed is cumulative so that an acknowledgment of sequence number X indicates that all octets up to but not including X have been received. This mechanism allows for straightforward duplicate detection in the presence of retransmission. Numbering of octets within a segment is that the first data octet immediately following the header is the lowest numbered, and the following octets are numbered consecutively.

It is essential to remember that the actual sequence number space is finite, though very large. This space ranges from 0 to $2^{32} - 1$. Since the space is finite, all arithmetic dealing with sequence numbers must be performed modulo $2^{32}$. This unsigned arithmetic preserves the relationship of sequence numbers as they cycle from $2^{32} - 1$ to 0 again. There are some subtleties to computer modulo arithmetic, so great care should be taken in programming the comparison of such values. The symbol "=\=" means "less than or equal" (modulo $2^{32}$).

The typical kinds of sequence number comparisons which the TCP must perform include:

(a) Determining that an acknowledgment refers to some sequence number sent but not yet acknowledged.

(b) Determining that all sequence numbers occupied by a segment have been acknowledged (e.g., to remove the segment from a retransmission queue).

(c) Determining that an incoming segment contains sequence numbers which are expected (i.e., that the segment "overlaps" the receive window).

In response to sending data the TCP will receive acknowledgments. The following comparisons are needed to process the acknowledgments.

\[ \text{SND.UNA} = \text{oldest unacknowledged sequence number} \]
\[ \text{SND.NXT} = \text{next sequence number to be sent} \]
\[ \text{SEG.ACK} = \text{acknowledgment from the receiving TCP (next sequence number expected by the receiving TCP)} \]
SEG.SEQ = first sequence number of a segment
SEG.LEN = the number of octets occupied by the data in the segment (counting SYN and FIN)
SEG.SEQ+SEG.LEN-1 = last sequence number of a segment

A new acknowledgment (called an "acceptable ack"), is one for which the inequality below holds:

SND.UNA < SEG.ACK =< SND.NXT

A segment on the retransmission queue is fully acknowledged if the sum of its sequence number and length is less or equal than the acknowledgment value in the incoming segment.

When data is received the following comparisons are needed:

RCV.NXT = next sequence number expected on an incoming segment, and is the left or lower edge of the receive window
RCV.NXT+RCV.WND-1 = last sequence number expected on an incoming segment, and is the right or upper edge of the receive window
SEG.SEQ = first sequence number occupied by the incoming segment
SEG.SEQ+SEG.LEN-1 = last sequence number occupied by the incoming segment

A segment is judged to occupy a portion of valid receive sequence space if

RCV.NXT =< SEG.SEQ < RCV.NXT+RCV.WND
or
RCV.NXT =< SEG.SEQ+SEG.LEN-1 < RCV.NXT+RCV.WND

The first part of this test checks to see if the beginning of the segment falls in the window, the second part of the test checks to see if the end of the segment falls in the window; if the segment passes either part of the test it contains data in the window.

Actually, it is a little more complicated than this. Due to zero windows and zero length segments, we have four cases for the acceptability of an incoming segment:
Segment Length | Receive Window | Test
---------------|---------------|-----------------
 0             | 0             | SEG.SEQ = RCV.NXT
 0             | >0            | RCV.NXT =< SEG.SEQ < RCV.NXT+RCV.WND
>0             | 0             | not acceptable
>0             | >0            | RCV.NXT =< SEG.SEQ < RCV.NXT+RCV.WND
               |               | or RCV.NXT =< SEG.SEQ+SEG.LEN-1 < RCV.NXT+RCV.WND

Note that when the receive window is zero no segments should be acceptable except ACK segments. Thus, it is be possible for a TCP to maintain a zero receive window while transmitting data and receiving ACKs. However, even when the receive window is zero, a TCP must process the RST and URG fields of all incoming segments.

We have taken advantage of the numbering scheme to protect certain control information as well. This is achieved by implicitly including some control flags in the sequence space so they can be retransmitted and acknowledged without confusion (i.e., one and only one copy of the control will be acted upon). Control information is not physically carried in the segment data space. Consequently, we must adopt rules for implicitly assigning sequence numbers to control. The SYN and FIN are the only controls requiring this protection, and these controls are used only at connection opening and closing. For sequence number purposes, the SYN is considered to occur before the first actual data octet of the segment in which it occurs, while the FIN is considered to occur after the last actual data octet in a segment in which it occurs. The segment length (SEG.LEN) includes both data and sequence space occupying controls. When a SYN is present then SEG.SEQ is the sequence number of the SYN.

Initial Sequence Number Selection

The protocol places no restriction on a particular connection being used over and over again. A connection is defined by a pair of sockets. New instances of a connection will be referred to as incarnations of the connection. The problem that arises from this is -- "how does the TCP identify duplicate segments from previous incarnations of the connection?" This problem becomes apparent if the connection is being opened and closed in quick succession, or if the connection breaks with loss of memory and is then reestablished.

To avoid confusion we must prevent segments from one incarnation of a connection from being used while the same sequence numbers may still
be present in the network from an earlier incarnation. We want to assure this, even if a TCP crashes and loses all knowledge of the sequence numbers it has been using. When new connections are created, an initial sequence number (ISN) generator is employed which selects a new 32 bit ISN. There are security issues that result if an off-path attacker is able to predict or guess ISN values.

The recommended ISN generator is based on the combination of a (possibly fictitious) 32 bit clock whose low order bit is incremented roughly every 4 microseconds, and a pseudorandom hash function (PRF). The clock component is intended to insure that with a Maximum Segment Lifetime (MSL), generated ISNs will be unique, since it cycles approximately every 4.55 hours, which is much longer than the MSL. This recommended algorithm is further described in RFC 1948 and builds on the basic clock-driven algorithm from RFC 793.

A TCP MUST use a clock-driven selection of initial sequence numbers, and SHOULD generate its Initial Sequence Numbers with the expression:

\[
\text{ISN} = M + F(\text{localip, localport, remoteip, remoteport, secretkey})
\]

where M is the 4 microsecond timer, and F() is a pseudorandom function (PRF) of the connection’s identifying parameters ("localip, localport, remoteip, remoteport") and a secret key ("secretkey"). F() MUST NOT be computable from the outside, or an attacker could still guess at sequence numbers from the ISN used for some other connection. The PRF could be implemented as a cryptographic hash of the concatenation of the TCP connection parameters and some secret data. For discussion of the selection of a specific hash algorithm and management of the secret key data, please see Section 3 of [14].

For each connection there is a send sequence number and a receive sequence number. The initial send sequence number (ISS) is chosen by the data sending TCP, and the initial receive sequence number (IRS) is learned during the connection establishing procedure.

For a connection to be established or initialized, the two TCPs must synchronize on each other's initial sequence numbers. This is done in an exchange of connection establishing segments carrying a control bit called "SYN" (for synchronize) and the initial sequence numbers. As a shorthand, segments carrying the SYN bit are also called "SYNs". Hence, the solution requires a suitable mechanism for picking an initial sequence number and a slightly involved handshake to exchange the ISN’s.

The synchronization requires each side to send its own initial sequence number and to receive a confirmation of it in acknowledgment.
from the other side. Each side must also receive the other side’s initial sequence number and send a confirming acknowledgment.

1) A --> B  SYN my sequence number is X
2) A <-- B  ACK your sequence number is X
3) A <-- B  SYN my sequence number is Y
4) A --> B  ACK your sequence number is Y

Because steps 2 and 3 can be combined in a single message this is called the three way (or three message) handshake.

A three way handshake is necessary because sequence numbers are not tied to a global clock in the network, and TCPs may have different mechanisms for picking the ISN’s. The receiver of the first SYN has no way of knowing whether the segment was an old delayed one or not, unless it remembers the last sequence number used on the connection (which is not always possible), and so it must ask the sender to verify this SYN. The three way handshake and the advantages of a clock-driven scheme are discussed in [3].

Knowing When to Keep Quiet

To be sure that a TCP does not create a segment that carries a sequence number which may be duplicated by an old segment remaining in the network, the TCP must keep quiet for an MSL before assigning any sequence numbers upon starting up or recovering from a crash in which memory of sequence numbers in use was lost. For this specification the MSL is taken to be 2 minutes. This is an engineering choice, and may be changed if experience indicates it is desirable to do so. Note that if a TCP is reinitialized in some sense, yet retains its memory of sequence numbers in use, then it need not wait at all; it must only be sure to use sequence numbers larger than those recently used.

The TCP Quiet Time Concept

This specification provides that hosts which "crash" without retaining any knowledge of the last sequence numbers transmitted on each active (i.e., not closed) connection shall delay emitting any TCP segments for at least the agreed MSL in the internet system of which the host is a part. In the paragraphs below, an explanation for this specification is given. TCP implementors may violate the "quiet time" restriction, but only at the risk of causing some old data to be accepted as new or new data rejected as old duplicated by some receivers in the internet system.

TCPs consume sequence number space each time a segment is formed and entered into the network output queue at a source host. The
duplicate detection and sequencing algorithm in the TCP protocol relies on the unique binding of segment data to sequence space to the extent that sequence numbers will not cycle through all $2^{32}$ values before the segment data bound to those sequence numbers has been delivered and acknowledged by the receiver and all duplicate copies of the segments have "drained" from the internet. Without such an assumption, two distinct TCP segments could conceivably be assigned the same or overlapping sequence numbers, causing confusion at the receiver as to which data is new and which is old. Remember that each segment is bound to as many consecutive sequence numbers as there are octets of data and SYN or FIN flags in the segment.

Under normal conditions, TCPs keep track of the next sequence number to emit and the oldest awaiting acknowledgment so as to avoid mistakenly using a sequence number over before its first use has been acknowledged. This alone does not guarantee that old duplicate data is drained from the net, so the sequence space has been made very large to reduce the probability that a wandering duplicate will cause trouble upon arrival. At 2 megabits/sec, it takes 4.5 hours to use up $2^{32}$ octets of sequence space. Since the maximum segment lifetime in the net is not likely to exceed a few tens of seconds, this is deemed ample protection for foreseeable nets, even if data rates escalate to 10’s of megabits/sec. At 100 megabits/sec, the cycle time is 5.4 minutes which may be a little short, but still within reason.

The basic duplicate detection and sequencing algorithm in TCP can be defeated, however, if a source TCP does not have any memory of the sequence numbers it last used on a given connection. For example, if the TCP were to start all connections with sequence number 0, then upon crashing and restarting, a TCP might re-form an earlier connection (possibly after half-open connection resolution) and emit packets with sequence numbers identical to or overlapping with packets still in the network which were emitted on an earlier incarnation of the same connection. In the absence of knowledge about the sequence numbers used on a particular connection, the TCP specification recommends that the source delay for MSL seconds before emitting segments on the connection, to allow time for segments from the earlier connection incarnation to drain from the system.

Even hosts which can remember the time of day and used it to select initial sequence number values are not immune from this problem (i.e., even if time of day is used to select an initial sequence number for each new connection incarnation).

Suppose, for example, that a connection is opened starting with sequence number $S$. Suppose that this connection is not used much and that eventually the initial sequence number function (ISN(t)) takes
on a value equal to the sequence number, say $S_1$, of the last segment sent by this TCP on a particular connection. Now suppose, at this instant, the host crashes, recovers, and establishes a new incarnation of the connection. The initial sequence number chosen is $S_1 = \text{ISN}(t)$ -- last used sequence number on old incarnation of connection! If the recovery occurs quickly enough, any old duplicates in the net bearing sequence numbers in the neighborhood of $S_1$ may arrive and be treated as new packets by the receiver of the new incarnation of the connection.

The problem is that the recovering host may not know for how long it crashed nor does it know whether there are still old duplicates in the system from earlier connection incarnations.

One way to deal with this problem is to deliberately delay emitting segments for one MSL after recovery from a crash-- this is the "quiet time" specification. Hosts which prefer to avoid waiting are willing to risk possible confusion of old and new packets at a given destination may choose not to wait for the "quite time". Implementors may provide TCP users with the ability to select on a connection by connection basis whether to wait after a crash, or may informally implement the "quite time" for all connections. Obviously, even where a user selects to "wait," this is not necessary after the host has been "up" for at least MSL seconds.

To summarize: every segment emitted occupies one or more sequence numbers in the sequence space, the numbers occupied by a segment are "busy" or "in use" until MSL seconds have passed, upon crashing a block of space-time is occupied by the octets and SYN or FIN flags of the last emitted segment, if a new connection is started too soon and uses any of the sequence numbers in the space-time footprint of the last segment of the previous connection incarnation, there is a potential sequence number overlap area which could cause confusion at the receiver.

3.4. Establishing a connection

The "three-way handshake" is the procedure used to establish a connection. This procedure normally is initiated by one TCP and responded to by another TCP. The procedure also works if two TCP simultaneously initiate the procedure. When simultaneous attempt occurs, each TCP receives a "SYN" segment which carries no acknowledgment after it has sent a "SYN". Of course, the arrival of an old duplicate "SYN" segment can potentially make it appear, to the recipient, that a simultaneous connection initiation is in progress. Proper use of "reset" segments can disambiguate these cases.
Several examples of connection initiation follow. Although these examples do not show connection synchronization using data-carrying segments, this is perfectly legitimate, so long as the receiving TCP doesn’t deliver the data to the user until it is clear the data is valid (i.e., the data must be buffered at the receiver until the connection reaches the ESTABLISHED state). The three-way handshake reduces the possibility of false connections. It is the implementation of a trade-off between memory and messages to provide information for this checking.

The simplest three-way handshake is shown in Figure 5 below. The figures should be interpreted in the following way. Each line is numbered for reference purposes. Right arrows (-->) indicate departure of a TCP segment from TCP A to TCP B, or arrival of a segment at B from A. Left arrows (<--), indicate the reverse. Ellipsis (...) indicates a segment which is still in the network (delayed). An "XXX" indicates a segment which is lost or rejected. Comments appear in parentheses. TCP states represent the state AFTER the departure or arrival of the segment (whose contents are shown in the center of each line). Segment contents are shown in abbreviated form, with sequence number, control flags, and ACK field. Other fields such as window, addresses, lengths, and text have been left out in the interest of clarity.

TCP A                                                TCP B
1.  CLOSED                                               LISTEN
2.  SYN-SENT    --> <SEQ=100><CTL=SYN>               --> SYN-RECEIVED
3.  ESTABLISHED <-- <SEQ=300><ACK=101><CTL=SYN,ACK>  <-- SYN-RECEIVED
4.  ESTABLISHED --> <SEQ=101><ACK=301><CTL=ACK>       --> ESTABLISHED
5.  ESTABLISHED --> <SEQ=101><ACK=301><CTL=ACK><DATA> --> ESTABLISHED

   Basic 3-Way Handshake for Connection Synchronization

   Figure 5

In line 2 of Figure 5, TCP A begins by sending a SYN segment indicating that it will use sequence numbers starting with sequence number 100. In line 3, TCP B sends a SYN and acknowledges the SYN it received from TCP A. Note that the acknowledgment field indicates TCP B is now expecting to hear sequence 101, acknowledging the SYN which occupied sequence 100.
At line 4, TCP A responds with an empty segment containing an ACK for
TCP B’s SYN; and in line 5, TCP A sends some data. Note that the
sequence number of the segment in line 5 is the same as in line 4
because the ACK does not occupy sequence number space (if it did, we
would wind up ACKing ACK’s!).

Simultaneous initiation is only slightly more complex, as is shown in
Figure 6. Each TCP cycles from CLOSED to SYN-SENT to SYN-RECEIVED to
ESTABLISHED.

TCP A                        TCP B
1. CLOSED                    CLOSED
2. SYN-SENT --> <SEQ=100><CTL=SYN> ...
3. SYN-RECEIVED <-- <SEQ=300><CTL=SYN> <-- SYN-SENT
4. ... <SEQ=100><CTL=SYN> --> SYN-RECEIVED
5. SYN-RECEIVED --> <SEQ=100><ACK=301><CTL=SYN,ACK> ...
6. ESTABLISHED <-- <SEQ=300><ACK=101><CTL=SYN,ACK> <-- SYN-RECEIVED
7. ... <SEQ=100><ACK=301><CTL=SYN,ACK> --> ESTABLISHED

Simultaneous Connection Synchronization

Figure 6

A TCP MUST support simultaneous open attempts.

Note that a TCP implementation MUST keep track of whether a
connection has reached SYN_RCVD state as the result of a passive OPEN
or an active OPEN.

The principle reason for the three-way handshake is to prevent old
duplicate connection initiations from causing confusion. To deal
with this, a special control message, reset, has been devised. If
the receiving TCP is in a non-synchronized state (i.e., SYN-SENT,
SYN-RECEIVED), it returns to LISTEN on receiving an acceptable reset.
If the TCP is in one of the synchronized states (ESTABLISHED, FIN-
WAIT-1, FIN-WAIT-2, CLOSE-WAIT, CLOSING, LAST-ACK, TIME-WAIT), it
aborts the connection and informs its user. We discuss this latter
case under "half-open" connections below.
1. CLOSED

2. SYN-SENT --> <SEQ=100><CTL=SYN>  ...  

3. (duplicate) ... <SEQ=90><CTL=SYN>  --> SYN-RECEIVED  

4. SYN-SENT <-- <SEQ=300><ACK=91><CTL=SYN,ACK>  <-- SYN-RECEIVED  

5. SYN-SENT  --> <SEQ=91><CTL=RST>  --> LISTEN  

6. ... <SEQ=100><CTL=SYN>  --> SYN-RECEIVED  

7. SYN-SENT  <-- <SEQ=400><ACK=101><CTL=SYN,ACK>  <-- SYN-RECEIVED  

8. ESTABLISHED  --> <SEQ=101><ACK=401><CTL=ACK>  --> ESTABLISHED  

Recovery from Old Duplicate SYN

As a simple example of recovery from old duplicates, consider Figure 7. At line 3, an old duplicate SYN arrives at TCP B. TCP B cannot tell that this is an old duplicate, so it responds normally (line 4). TCP A detects that the ACK field is incorrect and returns a RST (reset) with its SEQ field selected to make the segment believable. TCP B, on receiving the RST, returns to the LISTEN state. When the original SYN (pun intended) finally arrives at line 6, the synchronization proceeds normally. If the SYN at line 6 had arrived before the RST, a more complex exchange might have occurred with RST’s sent in both directions.

Half-Open Connections and Other Anomalies

An established connection is said to be "half-open" if one of the TCPs has closed or aborted the connection at its end without the knowledge of the other, or if the two ends of the connection have become desynchronized owing to a crash that resulted in loss of memory. Such connections will automatically become reset if an attempt is made to send data in either direction. However, half-open connections are expected to be unusual, and the recovery procedure is mildly involved.

If at site A the connection no longer exists, then an attempt by the user at site B to send any data on it will result in the site B TCP receiving a reset control message. Such a message indicates to the
site B TCP that something is wrong, and it is expected to abort the connection.

Assume that two user processes A and B are communicating with one another when a crash occurs causing loss of memory to A’s TCP. Depending on the operating system supporting A’s TCP, it is likely that some error recovery mechanism exists. When the TCP is up again, A is likely to start again from the beginning or from a recovery point. As a result, A will probably try to OPEN the connection again or try to SEND on the connection it believes open. In the latter case, it receives the error message "connection not open" from the local (A’s) TCP. In an attempt to establish the connection, A’s TCP will send a segment containing SYN. This scenario leads to the example shown in Figure 8. After TCP A crashes, the user attempts to re-open the connection. TCP B, in the meantime, thinks the connection is open.

TCP A                                           TCP B
1.  (CRASH)                               (send 300, receive 100)
2.  CLOSED                                           ESTABLISHED
3.  SYN-SENT --> <SEQ=400><CTL=SYN>              --> (??)
4.  (!!)     <-- <SEQ=300><ACK=100><CTL=ACK>     <-- ESTABLISHED
5.  SYN-SENT --> <SEQ=100><CTL=RST>              --> (Abort!!)
6.  SYN-SENT                                         CLOSED
7.  SYN-SENT --> <SEQ=400><CTL=SYN>              -->

Half-Open Connection Discovery

Figure 8

When the SYN arrives at line 3, TCP B, being in a synchronized state, and the incoming segment outside the window, responds with an acknowledgment indicating what sequence it next expects to hear (ACK 100). TCP A sees that this segment does not acknowledge anything it sent and, being unsynchronized, sends a reset (RST) because it has detected a half-open connection. TCP B aborts at line 5. TCP A will continue to try to establish the connection; the problem is now reduced to the basic 3-way handshake of Figure 5.

An interesting alternative case occurs when TCP A crashes and TCP B tries to send data on what it thinks is a synchronized connection.

Eddy                     Expires March 24, 2016                [Page 24]
This is illustrated in Figure 9. In this case, the data arriving at TCP A from TCP B (line 2) is unacceptable because no such connection exists, so TCP A sends a RST. The RST is acceptable so TCP B processes it and aborts the connection.

TCP A                                              TCP B
1.  (CRASH)                              (send 300, receive 100)
2.  (??)    <-- <SEQ=300><ACK=100><DATA=10><CTL=ACK> <-- ESTABLISHED
3.          --> <SEQ=100><CTL=RST>                   --> (ABORT!!)

Active Side Causes Half-Open Connection Discovery

In Figure 10, we find the two TCPs A and B with passive connections waiting for SYN. An old duplicate arriving at TCP B (line 2) stirs B into action. A SYN-ACK is returned (line 3) and causes TCP A to generate a RST (the ACK in line 3 is not acceptable). TCP B accepts the reset and returns to its passive LISTEN state.

TCP A                                         TCP B
1.  LISTEN                                        LISTEN
2.       ... <SEQ=Z><CTL=SYN>                -->  SYN-RECEIVED
3.  (??) <-- <SEQ=X><ACK=Z+1><CTL=SYN,ACK>   <--  SYN-RECEIVED
4.       --> <SEQ=Z+1><CTL=RST>              -->  (return to LISTEN!)
5.  LISTEN                                        LISTEN

Old Duplicate SYN Initiates a Reset on two Passive Sockets

A variety of other cases are possible, all of which are accounted for by the following rules for RST generation and processing.

Reset Generation
As a general rule, reset (RST) must be sent whenever a segment arrives which apparently is not intended for the current connection. A reset must not be sent if it is not clear that this is the case.

There are three groups of states:

1. If the connection does not exist (CLOSED) then a reset is sent in response to any incoming segment except another reset. In particular, SYN s addressed to a non-existent connection are rejected by this means.

   If the incoming segment has the ACK bit set, the reset takes its sequence number from the ACK field of the segment, otherwise the reset has sequence number zero and the ACK field is set to the sum of the sequence number and segment length of the incoming segment. The connection remains in the CLOSED state.

2. If the connection is in any non-synchronized state (LISTEN, SYN-SENT, SYN RECEIVED), and the incoming segment acknowledges something not yet sent (the segment carries an unacceptable ACK), or if an incoming segment has a security level or compartment which does not exactly match the level and compartment requested for the connection, a reset is sent.

   If our SYN has not been acknowledged and the precedence level of the incoming segment is higher than the precedence level requested then either raise the local precedence level (if allowed by the user and the system) or send a reset; or if the precedence level of the incoming segment is lower than the precedence level requested then continue as if the precedence matched exactly (if the remote TCP cannot raise the precedence level to match ours this will be detected in the next segment it sends, and the connection will be terminated then). If our SYN has been acknowledged (perhaps in this incoming segment) the precedence level of the incoming segment must match the local precedence level exactly, if it does not a reset must be sent.

   If the incoming segment has an ACK field, the reset takes its sequence number from the ACK field of the segment, otherwise the reset has sequence number zero and the ACK field is set to the sum of the sequence number and segment length of the incoming segment. The connection remains in the same state.

3. If the connection is in a synchronized state (ESTABLISHED, FIN WAIT 1, FIN WAIT 2, CLOSE WAIT, CLOSING, LAST ACK, TIME WAIT), any unacceptable segment (out of window sequence number or unacceptable acknowledgment number) must elicit only an empty acknowledgment segment containing the current send sequence number.
and an acknowledgment indicating the next sequence number expected to be received, and the connection remains in the same state.

If an incoming segment has a security level, or compartment, or precedence which does not exactly match the level, and compartment, and precedence requested for the connection, a reset is sent and the connection goes to the CLOSED state. The reset takes its sequence number from the ACK field of the incoming segment.

Reset Processing

In all states except SYN-SENT, all reset (RST) segments are validated by checking their SEQ-fields. A reset is valid if its sequence number is in the window. In the SYN-SENT state (a RST received in response to an initial SYN), the RST is acceptable if the ACK field acknowledges the SYN.

The receiver of a RST first validates it, then changes state. If the receiver was in the LISTEN state, it ignores it. If the receiver was in SYN-RECEIVED state and had previously been in the LISTEN state, then the receiver returns to the LISTEN state, otherwise the receiver aborts the connection and goes to the CLOSED state. If the receiver was in any other state, it aborts the connection and advises the user and goes to the CLOSED state.

TCP SHOULD allow a received RST segment to include data.

3.5. Closing a Connection

CLOSE is an operation meaning "I have no more data to send." The notion of closing a full-duplex connection is subject to ambiguous interpretation, of course, since it may not be obvious how to treat the receiving side of the connection. We have chosen to treat CLOSE in a simplex fashion. The user who CLOSES may continue to RECEIVE until he is told that the other side has CLOSED also. Thus, a program could initiate several SENDs followed by a CLOSE, and then continue to RECEIVE until signaled that a RECEIVE failed because the other side has CLOSED. We assume that the TCP will signal a user, even if no RECEIVES are outstanding, that the other side has closed, so the user can terminate his side gracefully. A TCP will reliably deliver all buffers SENT before the connection was CLOSED so a user who expects no data in return need only wait to hear the connection was CLOSED successfully to know that all his data was received at the destination TCP. Users must keep reading connections they close for sending until the TCP says no more data.

There are essentially three cases:
1) The user initiates by telling the TCP to CLOSE the connection

2) The remote TCP initiates by sending a FIN control signal

3) Both users CLOSE simultaneously

Case 1: Local user initiates the close

In this case, a FIN segment can be constructed and placed on the outgoing segment queue. No further SENDs from the user will be accepted by the TCP, and it enters the FIN-WAIT-1 state. RECEIVEs are allowed in this state. All segments preceding and including FIN will be retransmitted until acknowledged. When the other TCP has both acknowledged the FIN and sent a FIN of its own, the first TCP can ACK this FIN. Note that a TCP receiving a FIN will ACK but not send its own FIN until its user has CLOSED the connection also.

Case 2: TCP receives a FIN from the network

If an unsolicited FIN arrives from the network, the receiving TCP can ACK it and tell the user that the connection is closing. The user will respond with a CLOSE, upon which the TCP can send a FIN to the other TCP after sending any remaining data. The TCP then waits until its own FIN is acknowledged whereupon it deletes the connection. If an ACK is not forthcoming, after the user timeout the connection is aborted and the user is told.

Case 3: both users close simultaneously

A simultaneous CLOSE by users at both ends of a connection causes FIN segments to be exchanged. When all segments preceding the FINs have been processed and acknowledged, each TCP can ACK the FIN it has received. Both will, upon receiving these ACKs, delete the connection.
A TCP connection may terminate in two ways: (1) the normal TCP close sequence using a FIN handshake, and (2) an "abort" in which one or more RST segments are sent and the connection state is immediately discarded. If a TCP connection is closed by the remote site, the
local application MUST be informed whether it closed normally or was aborted.

3.5.1. Half-Closed Connections

The normal TCP close sequence delivers buffered data reliably in both directions. Since the two directions of a TCP connection are closed independently, it is possible for a connection to be "half closed," i.e., closed in only one direction, and a host is permitted to continue sending data in the open direction on a half-closed connection.

A host MAY implement a "half-duplex" TCP close sequence, so that an application that has called CLOSE cannot continue to read data from the connection. If such a host issues a CLOSE call while received data is still pending in TCP, or if new data is received after CLOSE is called, its TCP SHOULD send a RST to show that data was lost.

When a connection is closed actively, it MUST linger in TIME-WAIT state for a time 2xMSL (Maximum Segment Lifetime). However, it MAY accept a new SYN from the remote TCP to reopen the connection directly from TIME-WAIT state, if it:

(1) assigns its initial sequence number for the new connection to be larger than the largest sequence number it used on the previous connection incarnation, and

(2) returns to TIME-WAIT state if the SYN turns out to be an old duplicate.

3.6. Precedence and Security

The intent is that connection be allowed only between ports operating with exactly the same security and compartment values and at the higher of the precedence level requested by the two ports.

The precedence and security parameters used in TCP are exactly those defined in the Internet Protocol (IP) [2]. Throughout this TCP specification the term "security/compartment" is intended to indicate the security parameters used in IP including security, compartment, user group, and handling restriction.

A connection attempt with mismatched security/compartment values or a lower precedence value must be rejected by sending a reset. Rejecting a connection due to too low a precedence only occurs after an acknowledgment of the SYN has been received.
Note that TCP modules which operate only at the default value of precedence will still have to check the precedence of incoming segments and possibly raise the precedence level they use on the connection.

The security parameters may be used even in a non-secure environment (the values would indicate unclassified data), thus hosts in non-secure environments must be prepared to receive the security parameters, though they need not send them.

3.7. Segmentation

The term "segmentation" refers to the activity TCP performs when ingesting a stream of bytes from a sending application and packetizing that stream of bytes into TCP segments. Individual TCP segments often do not correspond one-for-one to individual send (or socket write) calls from the application. Applications may perform writes at the granularity of messages in the upper layer protocol, but TCP guarantees no boundary coherence between the TCP segments sent and received versus user application data read or write buffer boundaries. In some specific protocols, such as RDMA using DDP and MPA [10], there are performance optimizations possible when the relation between TCP segments and application data units can be controlled, and MPA includes a specific mechanism for detecting and verifying this relationship between TCP segments and application message data structures, but this is specific to applications like RDMA. In general, multiple goals influence the sizing of TCP segments created by a TCP implementation.

Goals driving the sending of larger segments include:

- Reducing the number of packets in flight within the network.
- Increasing processing efficiency and potential performance by enabling a smaller number of interrupts and inter-layer interactions.
- Limiting the overhead of TCP headers.

Note that the performance benefits of sending larger segments may decrease as the size increases, and there may be boundaries where advantages are reversed. For instance, on some machines 1025 bytes within a segment could lead to worse performance than 1024 bytes, due purely to data alignment on copy operations.

Goals driving the sending of smaller segments include:
o Avoiding sending segments larger than the smallest MTU within an
IP network path, because this results in either packet loss or
fragmentation. Making matters worse, some firewalls or
middleboxes may drop fragmented packets or ICMP messages related
to fragmentation.

o Preventing delays to the application data stream, especially when
TCP is waiting on the application to generate more data, or when
the application is waiting on an event or input from its peer in
order to generate more data.

o Enabling "fate sharing" between TCP segments and lower-layer data
units (e.g. below IP, for links with cell or frame sizes smaller
than the IP MTU).

Towards meeting these competing sets of goals, TCP includes several
mechanisms, including the Maximum Segment Size option, Path MTU
Discovery, the Nagle algorithm, and support for IPv6 Jumbograms, as
discussed in the following subsections.

3.7.1. Maximum Segment Size Option

TCP MUST implement both sending and receiving the MSS option.

TCP SHOULD send an MSS option in every SYN segment when its receive
MSS differs from the default 536 for IPv4 or 1220 for IPv6, and MAY
send it always.

If an MSS option is not received at connection setup, TCP MUST assume
a default send MSS of 536 (576 - 40) for IPv4 or 1220 (1280 - 60) for
IPv6.

The maximum size of a segment that TCP really sends, the "effective
send MSS," MUST be the smaller of the send MSS (which reflects the
available reassembly buffer size at the remote host) and the largest
size permitted by the IP layer:

\[ \text{Eff.snd.MSS} = \min(\text{SendMSS} + 20, \text{MMS}_S) - \text{TCPhdrsize} - \text{IPoptionsize} \]

where:

o SendMSS is the MSS value received from the remote host, or the
default 536 for IPv4 or 1220 for IPv6, if no MSS option is
received.
o MMS_S is the maximum size for a transport-layer message that TCP may send.

o TCPPhdrsize is the size of the fixed TCP header and any options. This is 20 in the (rare) case that no options are present, but may be larger if TCP options are to be sent. Note that some options may not be included on all segments, but that for each segment sent, the sender should adjust the data length accordingly, within the Eff_snd.MSS.

o IPOptionsize is the size of any IP options associated with a TCP connection. Note that some options may not be included on all packets, but that for each segment sent, the sender should adjust the data length accordingly, within the Eff_snd.MSS.

The MSS value to be sent in an MSS option should be equal to the effective MTU minus the fixed IP and TCP headers. By ignoring both IP and TCP options when calculating the value for the MSS option, if there are any IP or TCP options to be sent in a packet, then the sender must decrease the size of the TCP data accordingly. RFC 6691 [15] discusses this in greater detail.

The MSS value to be sent in an MSS option must be less than or equal to:

\[ \text{MMS}_R - 20 \]

where MMS_R is the maximum size for a transport-layer message that can be received (and reassembled). TCP obtains MMS_R and MMS_S from the IP layer; see the generic call GET_MAXSIZES in Section 3.4 of RFC 1122.

When TCP is used in a situation where either the IP or TCP headers are not fixed, the sender must reduce the amount of TCP data in any given packet by the number of octets used by the IP and TCP options. This has been a point of confusion historically, as explained in RFC 6691, Section 3.1.

3.7.2. Path MTU Discovery

A TCP implementation may be aware of the MTU on directly connected links, but will rarely have insight about MTUs across an entire network path. For IPv4, RFC 1122 provides an IP-layer recommendation on the default effective MTU for sending to be less than or equal to 576 for destinations not directly connected. For IPv6, this would be 1280. In all cases, however, implementation of Path MTU Discovery (PMTUD) and Packetization Layer Path MTU Discovery (PLPMTUD) is
strongly recommended in order for TCP to improve segmentation decisions.

PMTUD for IPv4 [1] or IPv6 [2] is implemented in conjunction between TCP, IP, and ICMP protocols. Several adjustments to a TCP implementation with PMTUD are described in RFC 2923 in order to deal with problems experienced in practice [5]. PLPMTUD [9] is a Standards Track improvement to PMTUD that relaxes the requirement for ICMP support across a path, and improves performance in cases where ICMP is not consistently conveyed. The mechanisms in all four of these RFCs are recommended to be included in TCP implementations.

The TCP MSS option specifies an upper bound for the size of packets that can be received. Hence, setting the value in the MSS option too small can impact the ability for PMTUD or PLPMTUD to find a larger path MTU. RFC 1191 discusses this implication of many older TCP implementations setting MSS to 536 for non-local destinations, rather than deriving it from the MTUs of connected interfaces as recommended.

3.7.3. Interfaces with Variable MTU Values

The effective MTU can sometimes vary, as when used with variable compression, e.g., ROBust Header Compression (ROHC) [12]. It is tempting for TCP to want to advertise the largest possible MSS, to support the most efficient use of compressed payloads. Unfortunately, some compression schemes occasionally need to transmit full headers (and thus smaller payloads) to resynchronize state at their endpoint compressors/decompressors. If the largest MTU is used to calculate the value to advertise in the MSS option, TCP retransmission may interfere with compressor resynchronization.

As a result, when the effective MTU of an interface varies, TCP SHOULD use the smallest effective MTU of the interface to calculate the value to advertise in the MSS option.

3.7.4. Nagle Algorithm

The "Nagle algorithm" was described in RFC 896 [7] and was recommended in RFC 1122 [8] for mitigation of an early problem of too many small packets being generated. It has been implemented in most current TCP code bases, sometimes with minor variations.

If there is unacknowledged data (i.e., SND.NXT > SND.UNA), then the sending TCP buffers all user data (regardless of the PSH bit), until the outstanding data has been acknowledged or until the TCP can send a full-sized segment (Eff.snd.MSS bytes).
TODO - see if SEND description later should be updated to reflect this

A TCP SHOULD implement the Nagle Algorithm to coalesce short segments. However, there MUST be a way for an application to disable the Nagle algorithm on an individual connection. In all cases, sending data is also subject to the limitation imposed by the Slow Start algorithm [11].

3.7.5. IPv6 Jumbograms

In order to support TCP over IPv6 jumbograms, implementations need to be able to send TCP segments larger than the 64KB limit that the MSS option can convey. RFC 2675 [4] defines that an MSS value of 65,535 bytes is to be treated as infinity, and Path MTU Discovery [2] is used to determine the actual MSS.

3.8. Data Communication

Once the connection is established data is communicated by the exchange of segments. Because segments may be lost due to errors (checksum test failure), or network congestion, TCP uses retransmission (after a timeout) to ensure delivery of every segment. Duplicate segments may arrive due to network or TCP retransmission. As discussed in the section on sequence numbers the TCP performs certain tests on the sequence and acknowledgment numbers in the segments to verify their acceptability.

The sender of data keeps track of the next sequence number to use in the variable SND.NXT. The receiver of data keeps track of the next sequence number to expect in the variable RCV.NXT. The sender of data keeps track of the oldest unacknowledged sequence number in the variable SND.UNA. If the data flow is momentarily idle and all data sent has been acknowledged then the three variables will be equal.

When the sender creates a segment and transmits it the sender advances SND.NXT. When the receiver accepts a segment it advances RCV.NXT and sends an acknowledgment. When the data sender receives an acknowledgment it advances SND.UNA. The extent to which the values of these variables differ is a measure of the delay in the communication. The amount by which the variables are advanced is the length of the data and SYN or FIN flags in the segment. Note that once in the ESTABLISHED state all segments must carry current acknowledgment information.

The CLOSE user call implies a push function, as does the FIN control flag in an incoming segment.
Retransmission Timeout

NOTE: TODO this needs to be updated in light of 1122 4.2.2.15 and errata 573; this will be done as part of RFC 1122 incorporation into this document.

Because of the variability of the networks that compose an internetwork system and the wide range of uses of TCP connections the retransmission timeout must be dynamically determined. One procedure for determining a retransmission timeout is given here as an illustration.

An Example Retransmission Timeout Procedure

Measure the elapsed time between sending a data octet with a particular sequence number and receiving an acknowledgment that covers that sequence number (segments sent do not have to match segments received). This measured elapsed time is the Round Trip Time (RTT). Next compute a Smoothed Round Trip Time (SRTT) as:

$$SRTT = (\text{ALPHA} \times SRTT) + ((1-\text{ALPHA}) \times RTT)$$

and based on this, compute the retransmission timeout (RTO) as:

$$RTO = \min[\text{UBOUND}, \max[\text{LBOUND}, (\text{BETA} \times SRTT)]]$$

where UBOUND is an upper bound on the timeout (e.g., 1 minute), LBOUND is a lower bound on the timeout (e.g., 1 second), ALPHA is a smoothing factor (e.g., .8 to .9), and BETA is a delay variance factor (e.g., 1.3 to 2.0).

The Communication of Urgent Information

As a result of implementation differences and middlebox interactions, new applications SHOULD NOT employ the TCP urgent mechanism. However, TCP implementations MUST still include support for the urgent mechanism. Details can be found in RFC 6093 [13].

The objective of the TCP urgent mechanism is to allow the sending user to stimulate the receiving user to accept some urgent data and to permit the receiving TCP to indicate to the receiving user when all the currently known urgent data has been received by the user.

This mechanism permits a point in the data stream to be designated as the end of urgent information. Whenever this point is in advance of the receive sequence number (RCV.NXT) at the receiving TCP, that TCP must tell the user to go into "urgent mode"; when the receive sequence number catches up to the urgent pointer, the TCP must tell user to go into "normal mode". If the urgent pointer is updated
while the user is in "urgent mode", the update will be invisible to the user.

The method employs a urgent field which is carried in all segments transmitted. The URG control flag indicates that the urgent field is meaningful and must be added to the segment sequence number to yield the urgent pointer. The absence of this flag indicates that there is no urgent data outstanding.

To send an urgent indication the user must also send at least one data octet. If the sending user also indicates a push, timely delivery of the urgent information to the destination process is enhanced.

A TCP MUST support a sequence of urgent data of any length. [8]

A TCP MUST inform the application layer asynchronously whenever it receives an Urgent pointer and there was previously no pending urgent data, or whenever the Urgent pointer advances in the data stream. There MUST be a way for the application to learn how much urgent data remains to be read from the connection, or at least to determine whether or not more urgent data remains to be read. [8]

Managing the Window

The window sent in each segment indicates the range of sequence numbers the sender of the window (the data receiver) is currently prepared to accept. There is an assumption that this is related to the currently available data buffer space available for this connection.

Indicating a large window encourages transmissions. If more data arrives than can be accepted, it will be discarded. This will result in excessive retransmissions, adding unnecessarily to the load on the network and the TCPs. Indicating a small window may restrict the transmission of data to the point of introducing a round trip delay between each new segment transmitted.

The mechanisms provided allow a TCP to advertise a large window and to subsequently advertise a much smaller window without having accepted that much data. This, so called "shrinking the window," is strongly discouraged. The robustness principle dictates that TCPs will not shrink the window themselves, but will be prepared for such behavior on the part of other TCPs.

The sending TCP must be prepared to accept from the user and send at least one octet of new data even if the send window is zero. The sending TCP must regularly retransmit to the receiving TCP even when
the window is zero. Two minutes is recommended for the retransmission interval when the window is zero. This retransmission is essential to guarantee that when either TCP has a zero window the re-opening of the window will be reliably reported to the other.

When the receiving TCP has a zero window and a segment arrives it must still send an acknowledgment showing its next expected sequence number and current window (zero).

The sending TCP packages the data to be transmitted into segments which fit the current window, and may repackage segments on the retransmission queue. Such repackaging is not required, but may be helpful.

In a connection with a one-way data flow, the window information will be carried in acknowledgment segments that all have the same sequence number so there will be no way to reorder them if they arrive out of order. This is not a serious problem, but it will allow the window information to be on occasion temporarily based on old reports from the data receiver. A refinement to avoid this problem is to act on the window information from segments that carry the highest acknowledgment number (that is segments with acknowledgment number equal or greater than the highest previously received).

The window management procedure has significant influence on the communication performance. The following comments are suggestions to implementers.

Window Management Suggestions

Allocating a very small window causes data to be transmitted in many small segments when better performance is achieved using fewer large segments.

One suggestion for avoiding small windows is for the receiver to defer updating a window until the additional allocation is at least X percent of the maximum allocation possible for the connection (where X might be 20 to 40).

Another suggestion is for the sender to avoid sending small segments by waiting until the window is large enough before sending data. If the user signals a push function then the data must be sent even if it is a small segment.

Note that the acknowledgments should not be delayed or unnecessary retransmissions will result. One strategy would be to send an acknowledgment when a small segment arrives (with out updating the
window information), and then to send another acknowledgment with new window information when the window is larger.

The segment sent to probe a zero window may also begin a break up of transmitted data into smaller and smaller segments. If a segment containing a single data octet sent to probe a zero window is accepted, it consumes one octet of the window now available. If the sending TCP simply sends as much as it can whenever the window is non zero, the transmitted data will be broken into alternating big and small segments. As time goes on, occasional pauses in the receiver making window allocation available will result in breaking the big segments into a small and not quite so big pair. And after a while the data transmission will be in mostly small segments.

The suggestion here is that the TCP implementations need to actively attempt to combine small window allocations into larger windows, since the mechanisms for managing the window tend to lead to many small windows in the simplest minded implementations.

3.9. Interfaces

There are of course two interfaces of concern: the user/TCP interface and the TCP/lower-level interface. We have a fairly elaborate model of the user/TCP interface, but the interface to the lower level protocol module is left unspecified here, since it will be specified in detail by the specification of the lower level protocol. For the case that the lower level is IP we note some of the parameter values that TCPs might use.

3.9.1. User/TCP Interface

The following functional description of user commands to the TCP is, at best, fictional, since every operating system will have different facilities. Consequently, we must warn readers that different TCP implementations may have different user interfaces. However, all TCPs must provide a certain minimum set of services to guarantee that all TCP implementations can support the same protocol hierarchy. This section specifies the functional interfaces required of all TCP implementations.

TCP User Commands

The following sections functionally characterize a USER/TCP interface. The notation used is similar to most procedure or function calls in high level languages, but this usage is not meant to rule out trap type service calls (e.g., SVCs, UUOs, EMTs).
The user commands described below specify the basic functions the TCP must perform to support interprocess communication. Individual implementations must define their own exact format, and may provide combinations or subsets of the basic functions in single calls. In particular, some implementations may wish to automatically OPEN a connection on the first SEND or RECEIVE issued by the user for a given connection.

In providing interprocess communication facilities, the TCP must not only accept commands, but must also return information to the processes it serves. The latter consists of:

(a) general information about a connection (e.g., interrupts, remote close, binding of unspecified foreign socket).

(b) replies to specific user commands indicating success or various types of failure.

Open

Format: OPEN (local port, foreign socket, active/passive [, timeout] [, precedence] [, security/compartment] [local IP address,] [, options]) -> local connection name

We assume that the local TCP is aware of the identity of the processes it serves and will check the authority of the process to use the connection specified. Depending upon the implementation of the TCP, the local network and TCP identifiers for the source address will either be supplied by the TCP or the lower level protocol (e.g., IP). These considerations are the result of concern about security, to the extent that no TCP be able to masquerade as another one, and so on. Similarly, no process can masquerade as another without the collusion of the TCP.

If the active/passive flag is set to passive, then this is a call to LISTEN for an incoming connection. A passive open may have either a fully specified foreign socket to wait for a particular connection or an unspecified foreign socket to wait for any call. A fully specified passive call can be made active by the subsequent execution of a SEND.

A transmission control block (TCB) is created and partially filled in with data from the OPEN command parameters.

On an active OPEN command, the TCP will begin the procedure to synchronize (i.e., establish) the connection at once.
The timeout, if present, permits the caller to set up a timeout for all data submitted to TCP. If data is not successfully delivered to the destination within the timeout period, the TCP will abort the connection. The present global default is five minutes.

The TCP or some component of the operating system will verify the users authority to open a connection with the specified precedence or security/compartment. The absence of precedence or security/compartment specification in the OPEN call indicates the default values must be used.

TCP will accept incoming requests as matching only if the security/compartment information is exactly the same and only if the precedence is equal to or higher than the precedence requested in the OPEN call.

The precedence for the connection is the higher of the values requested in the OPEN call and received from the incoming request, and fixed at that value for the life of the connection. Implementers may want to give the user control of this precedence negotiation. For example, the user might be allowed to specify that the precedence must be exactly matched, or that any attempt to raise the precedence be confirmed by the user.

A local connection name will be returned to the user by the TCP. The local connection name can then be used as a short hand term for the connection defined by the <local socket, foreign socket> pair.

The optional "local IP address" parameter MUST be supported to allow the specification of the local IP address. This enables applications that need to select the local IP address used when multihoming is present.

A passive OPEN call with a specified "local IP address" parameter will await an incoming connection request to that address. If the parameter is unspecified, a passive OPEN will await an incoming connection request to any local IP address, and then bind the local IP address of the connection to the particular address that is used.

For an active OPEN call, a specified "local IP address" parameter will be used for opening the connection. If the parameter is unspecified, the TCP will choose an appropriate local IP address (see RFC 1122 section 3.3.4.2).
Send

Format: SEND (local connection name, buffer address, byte count, PUSH flag, URGENT flag [,timeout])

This call causes the data contained in the indicated user buffer to be sent on the indicated connection. If the connection has not been opened, the SEND is considered an error. Some implementations may allow users to SEND first; in which case, an automatic OPEN would be done. If the calling process is not authorized to use this connection, an error is returned.

If the PUSH flag is set, the data must be transmitted promptly to the receiver, and the PUSH bit will be set in the last TCP segment created from the buffer. If the PUSH flag is not set, the data may be combined with data from subsequent SENDs for transmission efficiency.

New applications SHOULD NOT set the URGENT flag [13] due to implementation differences and middlebox issues.

If the URGENT flag is set, segments sent to the destination TCP will have the urgent pointer set. The receiving TCP will signal the urgent condition to the receiving process if the urgent pointer indicates that data preceding the urgent pointer has not been consumed by the receiving process. The purpose of urgent is to stimulate the receiver to process the urgent data and to indicate to the receiver when all the currently known urgent data has been received. The number of times the sending user’s TCP signals urgent will not necessarily be equal to the number of times the receiving user will be notified of the presence of urgent data.

If no foreign socket was specified in the OPEN, but the connection is established (e.g., because a LISTENing connection has become specific due to a foreign segment arriving for the local socket), then the designated buffer is sent to the implied foreign socket. Users who make use of OPEN with an unspecified foreign socket can make use of SEND without ever explicitly knowing the foreign socket address.

However, if a SEND is attempted before the foreign socket becomes specified, an error will be returned. Users can use the STATUS call to determine the status of the connection. In some implementations the TCP may notify the user when an unspecified socket is bound.
If a timeout is specified, the current user timeout for this connection is changed to the new one.

In the simplest implementation, SEND would not return control to the sending process until either the transmission was complete or the timeout had been exceeded. However, this simple method is both subject to deadlocks (for example, both sides of the connection might try to do SENDs before doing any RECEIVES) and offers poor performance, so it is not recommended. A more sophisticated implementation would return immediately to allow the process to run concurrently with network I/O, and, furthermore, to allow multiple SENDs to be in progress. Multiple SENDs are served in first come, first served order, so the TCP will queue those it cannot service immediately.

We have implicitly assumed an asynchronous user interface in which a SEND later elicits some kind of SIGNAL or pseudo-interrupt from the serving TCP. An alternative is to return a response immediately. For instance, SENDs might return immediate local acknowledgment, even if the segment sent had not been acknowledged by the distant TCP. We could optimistically assume eventual success. If we are wrong, the connection will close anyway due to the timeout. In implementations of this kind (synchronous), there will still be some asynchronous signals, but these will deal with the connection itself, and not with specific segments or buffers.

In order for the process to distinguish among error or success indications for different SENDs, it might be appropriate for the buffer address to be returned along with the coded response to the SEND request. TCP-to-user signals are discussed below, indicating the information which should be returned to the calling process.

Receive

Format: RECEIVE (local connection name, buffer address, byte count) → byte count, urgent flag, push flag

This command allocates a receiving buffer associated with the specified connection. If no OPEN precedes this command or the calling process is not authorized to use this connection, an error is returned.

In the simplest implementation, control would not return to the calling program until either the buffer was filled, or some error occurred, but this scheme is highly subject to deadlocks.
A more sophisticated implementation would permit several RECEIVEs to be outstanding at once. These would be filled as segments arrive. This strategy permits increased throughput at the cost of a more elaborate scheme (possibly asynchronous) to notify the calling program that a PUSH has been seen or a buffer filled.

If enough data arrive to fill the buffer before a PUSH is seen, the PUSH flag will not be set in the response to the RECEIVE. The buffer will be filled with as much data as it can hold. If a PUSH is seen before the buffer is filled the buffer will be returned partially filled and PUSH indicated.

If there is urgent data the user will have been informed as soon as it arrived via a TCP-to-user signal. The receiving user should thus be in "urgent mode". If the URGENT flag is on, additional urgent data remains. If the URGENT flag is off, this call to RECEIVE has returned all the urgent data, and the user may now leave "urgent mode". Note that data following the urgent pointer (non-urgent data) cannot be delivered to the user in the same buffer with preceding urgent data unless the boundary is clearly marked for the user.

To distinguish among several outstanding RECEIVEs and to take care of the case that a buffer is not completely filled, the return code is accompanied by both a buffer pointer and a byte count indicating the actual length of the data received.

Alternative implementations of RECEIVE might have the TCP allocate buffer storage, or the TCP might share a ring buffer with the user.

Close

Format: CLOSE (local connection name)

This command causes the connection specified to be closed. If the connection is not open or the calling process is not authorized to use this connection, an error is returned. Closing connections is intended to be a graceful operation in the sense that outstanding SENDs will be transmitted (and retransmitted), as flow control permits, until all have been serviced. Thus, it should be acceptable to make several SEND calls, followed by a CLOSE, and expect all the data to be sent to the destination. It should also be clear that users should continue to RECEIVE on CLOSING connections, since the other side may be trying to transmit the last of its data. Thus, CLOSE means "I have no more to send" but does not mean "I will
not receive any more." It may happen (if the user level protocol is not well thought out) that the closing side is unable to get rid of all its data before timing out. In this event, CLOSE turns into ABORT, and the closing TCP gives up.

The user may CLOSE the connection at any time on his own initiative, or in response to various prompts from the TCP (e.g., remote close executed, transmission timeout exceeded, destination inaccessible).

Because closing a connection requires communication with the foreign TCP, connections may remain in the closing state for a short time. Attempts to reopen the connection before the TCP replies to the CLOSE command will result in error responses.

Close also implies push function.

Status

Format: STATUS (local connection name) -> status data

This is an implementation dependent user command and could be excluded without adverse effect. Information returned would typically come from the TCB associated with the connection.

This command returns a data block containing the following information:

- local socket,
- foreign socket,
- local connection name,
- receive window,
- send window,
- connection state,
- number of buffers awaiting acknowledgment,
- number of buffers pending receipt,
- urgent state,
- precedence,
- security/compartment,
- and transmission timeout.

Depending on the state of the connection, or on the implementation itself, some of this information may not be available or meaningful. If the calling process is not authorized to use this connection, an error is returned. This prevents unauthorized processes from gaining information about a connection.
Abort

Format: ABORT (local connection name)

This command causes all pending SENDs and RECEIVES to be aborted, the TCB to be removed, and a special RESET message to be sent to the TCP on the other side of the connection. Depending on the implementation, users may receive abort indications for each outstanding SEND or RECEIVE, or may simply receive an ABORT-acknowledgment.

Flush

Some TCP implementations have included a FLUSH call, which will empty the TCP send queue of any data for which the user has issued SEND calls but which is still to the right of the current send window. That is, it flushes as much queued send data as possible without losing sequence number synchronization.

Set TOS

The application layer MUST be able to specify the Type-of-Service (TOS) for segments that are sent on a connection. It not required, but the application SHOULD be able to change the TOS during the connection lifetime. TCP SHOULD pass the current TOS value without change to the IP layer, when it sends segments on the connection.

The TOS will be specified independently in each direction on the connection, so that the receiver application will specify the TOS used for ACK segments.

TCP MAY pass the most recently received TOS up to the application.

TCP-to-User Messages

It is assumed that the operating system environment provides a means for the TCP to asynchronously signal the user program. When the TCP does signal a user program, certain information is passed to the user. Often in the specification the information will be an error message. In other cases there will be information relating to the completion of processing a SEND or RECEIVE or other user call.

The following information is provided:
3.9.2. TCP/Lower-Level Interface

The TCP calls on a lower level protocol module to actually send and receive information over a network. One case is that of the ARPA internetwork system where the lower level module is the Internet Protocol (IP) [2].

If the lower level protocol is IP it provides arguments for a type of service and for a time to live. TCP uses the following settings for these parameters:

- **Type of Service** = Precedence: given by user, Delay: normal, Throughput: normal, Reliability: normal; or binary XXX000000, where XXX are the three bits determining precedence, e.g. 000 means routine precedence.
- **Time to Live** = one minute, or 00111100.

Note that the assumed maximum segment lifetime is two minutes. Here we explicitly ask that a segment be destroyed if it cannot be delivered by the internet system within one minute.

If the lower level is IP (or other protocol that provides this feature) and source routing is used, the interface must allow the route information to be communicated. This is especially important so that the source and destination addresses used in the TCP checksum be the originating source and ultimate destination. It is also important to preserve the return route to answer connection requests.

Any lower level protocol will have to provide the source address, destination address, and protocol fields, and some way to determine the "TCP length", both to provide the functional equivalent service of IP and to be used in the TCP checksum.

3.10. Event Processing

The processing depicted in this section is an example of one possible implementation. Other implementations may have slightly different processing sequences, but they should differ from those in this section only in detail, not in substance.
The activity of the TCP can be characterized as responding to events. The events that occur can be cast into three categories: user calls, arriving segments, and timeouts. This section describes the processing the TCP does in response to each of the events. In many cases the processing required depends on the state of the connection.

Events that occur:

User Calls
- OPEN
- SEND
- RECEIVE
- CLOSE
- ABORT
- STATUS

Arriving Segments
- SEGMENT ARRIVES

Timeouts
- USER TIMEOUT
- RETRANSMISSION TIMEOUT
- TIME-WAIT TIMEOUT

The model of the TCP/user interface is that user commands receive an immediate return and possibly a delayed response via an event or pseudo interrupt. In the following descriptions, the term "signal" means cause a delayed response.

Error responses are given as character strings. For example, user commands referencing connections that do not exist receive "error: connection not open".

Please note in the following that all arithmetic on sequence numbers, acknowledgment numbers, windows, et cetera, is modulo 2**32 the size of the sequence number space. Also note that "=<" means less than or equal to (modulo 2**32).

A natural way to think about processing incoming segments is to imagine that they are first tested for proper sequence number (i.e., that their contents lie in the range of the expected "receive window" in the sequence number space) and then that they are generally queued and processed in sequence number order.
When a segment overlaps other already received segments we reconstruct the segment to contain just the new data, and adjust the header fields to be consistent.

Note that if no state change is mentioned the TCP stays in the same state.
OPEN Call

CLOSED STATE (i.e., TCB does not exist)

Create a new transmission control block (TCB) to hold connection state information. Fill in local socket identifier, foreign socket, precedence, security/compartment, and user timeout information. Note that some parts of the foreign socket may be unspecified in a passive OPEN and are to be filled in by the parameters of the incoming SYN segment. Verify the security and precedence requested are allowed for this user, if not return "error: precedence not allowed" or "error: security/compartment not allowed." If passive enter the LISTEN state and return. If active and the foreign socket is unspecified, return "error: foreign socket unspecified"; if active and the foreign socket is specified, issue a SYN segment. An initial send sequence number (ISS) is selected. A SYN segment of the form <SEQ=ISS><CTL=SYN> is sent. Set SND.UNA to ISS, SND.NXT to ISS+1, enter SYN-SENT state, and return.

If the caller does not have access to the local socket specified, return "error: connection illegal for this process". If there is no room to create a new connection, return "error: insufficient resources".

LISTEN STATE

If active and the foreign socket is specified, then change the connection from passive to active, select an ISS. Send a SYN segment, set SND.UNA to ISS, SND.NXT to ISS+1. Enter SYN-SENT state. Data associated with SEND may be sent with SYN segment or queued for transmission after entering ESTABLISHED state. The urgent bit if requested in the command must be sent with the data segments sent as a result of this command. If there is no room to queue the request, respond with "error: insufficient resources". If Foreign socket was not specified, then return "error: foreign socket unspecified".
SYN-SENT STATE
SYN-RECEIVED STATE
ESTABLISHED STATE
FIN-WAIT-1 STATE
FIN-WAIT-2 STATE
CLOSE-WAIT STATE
CLOSING STATE
LAST-ACK STATE
TIME-WAIT STATE

Return "error: connection already exists".
SEND Call

CLOSED STATE (i.e., TCB does not exist)

If the user does not have access to such a connection, then return "error: connection illegal for this process".

Otherwise, return "error: connection does not exist".

LISTEN STATE

If the foreign socket is specified, then change the connection from passive to active, select an ISS. Send a SYN segment, set SND.UNA to ISS, SND.NXT to ISS+1. Enter SYN-SENT state. Data associated with SEND may be sent with SYN segment or queued for transmission after entering ESTABLISHED state. The urgent bit if requested in the command must be sent with the data segments sent as a result of this command. If there is no room to queue the request, respond with "error: insufficient resources". If Foreign socket was not specified, then return "error: foreign socket unspecified".

SYN-SENT STATE
SYN-RECEIVED STATE

Queue the data for transmission after entering ESTABLISHED state. If no space to queue, respond with "error: insufficient resources".

ESTABLISHED STATE
CLOSE-WAIT STATE

Segmentize the buffer and send it with a piggybacked acknowledgment (acknowledgment value = RCV.NXT). If there is insufficient space to remember this buffer, simply return "error: insufficient resources".

If the urgent flag is set, then SND.UP <- SND.NXT and set the urgent pointer in the outgoing segments.

FIN-WAIT-1 STATE
FIN-WAIT-2 STATE
CLOSING STATE
LAST-ACK STATE
TIME-WAIT STATE

Return "error: connection closing" and do not service request.
RECEIVE Call

CLOSED STATE (i.e., TCB does not exist)

If the user does not have access to such a connection, return "error: connection illegal for this process".

Otherwise return "error: connection does not exist".

LISTEN STATE
SYN-SENT STATE
SYN-RECEIVED STATE

Queue for processing after entering ESTABLISHED state. If there is no room to queue this request, respond with "error: insufficient resources".

ESTABLISHED STATE
FIN-WAIT-1 STATE
FIN-WAIT-2 STATE

If insufficient incoming segments are queued to satisfy the request, queue the request. If there is no queue space to remember the RECEIVE, respond with "error: insufficient resources".

Reassemble queued incoming segments into receive buffer and return to user. Mark "push seen" (PUSH) if this is the case.

If RCV.UP is in advance of the data currently being passed to the user notify the user of the presence of urgent data.

When the TCP takes responsibility for delivering data to the user that fact must be communicated to the sender via an acknowledgment. The formation of such an acknowledgment is described below in the discussion of processing an incoming segment.

CLOSE-WAIT STATE

Since the remote side has already sent FIN, RECEIVES must be satisfied by text already on hand, but not yet delivered to the user. If no text is awaiting delivery, the RECEIVE will get a "error: connection closing" response. Otherwise, any remaining text can be used to satisfy the RECEIVE.

CLOSING STATE
LAST-ACK STATE
TIME-WAIT STATE

Return "error: connection closing".
CLOSE Call

CLOSED STATE (i.e., TCB does not exist)

If the user does not have access to such a connection, return "error: connection illegal for this process".

Otherwise, return "error: connection does not exist".

LISTEN STATE

Any outstanding RECEIVEs are returned with "error: closing" responses. Delete TCB, enter CLOSED state, and return.

SYN-SENT STATE

Delete the TCB and return "error: closing" responses to any queued SENDs, or RECEIVEs.

SYN-RECEIVED STATE

If no SENDs have been issued and there is no pending data to send, then form a FIN segment and send it, and enter FIN-WAIT-1 state; otherwise queue for processing after entering ESTABLISHED state.

ESTABLISHED STATE

Queue this until all preceding SENDs have been segmentized, then form a FIN segment and send it. In any case, enter FIN-WAIT-1 state.

FIN-WAIT-1 STATE

FIN-WAIT-2 STATE

Strictly speaking, this is an error and should receive a "error: connection closing" response. An "ok" response would be acceptable, too, as long as a second FIN is not emitted (the first FIN may be retransmitted though).

CLOSE-WAIT STATE

Queue this request until all preceding SENDs have been segmentized; then send a FIN segment, enter LAST-ACK state.

CLOSING STATE

LAST-ACK STATE

TIME-WAIT STATE
Respond with "error: connection closing".
ABORT Call

CLOSED STATE (i.e., TCB does not exist)

If the user should not have access to such a connection, return "error: connection illegal for this process".

Otherwise return "error: connection does not exist".

LISTEN STATE

Any outstanding RECEIVEs should be returned with "error: connection reset" responses. Delete TCB, enter CLOSED state, and return.

SYN-SENT STATE

All queued SENDs and RECEIVEs should be given "connection reset" notification, delete the TCB, enter CLOSED state, and return.

SYN-RECEIVED STATE

ESTABLISHED STATE

FIN-WAIT-1 STATE

FIN-WAIT-2 STATE

CLOSE-WAIT STATE

Send a reset segment:

<SEQ=SND.NXT><CTL=RST>

All queued SENDs and RECEIVEs should be given "connection reset" notification; all segments queued for transmission (except for the RST formed above) or retransmission should be flushed, delete the TCB, enter CLOSED state, and return.

CLOSING STATE LAST-ACK STATE TIME-WAIT STATE

Respond with "ok" and delete the TCB, enter CLOSED state, and return.
STATUS Call

CLOSED STATE (i.e., TCB does not exist)
If the user should not have access to such a connection, return "error: connection illegal for this process".
Otherwise return "error: connection does not exist".

LISTEN STATE
Return "state = LISTEN", and the TCB pointer.

SYN-SENT STATE
Return "state = SYN-SENT", and the TCB pointer.

SYN-RECEIVED STATE
Return "state = SYN-RECEIVED", and the TCB pointer.

ESTABLISHED STATE
Return "state = ESTABLISHED", and the TCB pointer.

FIN-WAIT-1 STATE
Return "state = FIN-WAIT-1", and the TCB pointer.

FIN-WAIT-2 STATE
Return "state = FIN-WAIT-2", and the TCB pointer.

CLOSE-WAIT STATE
Return "state = CLOSE-WAIT", and the TCB pointer.

CLOSING STATE
Return "state = CLOSING", and the TCB pointer.

LAST-ACK STATE
Return "state = LAST-ACK", and the TCB pointer.

TIME-WAIT STATE
Return "state = TIME-WAIT", and the TCB pointer.
SEGMENT ARRIVES

If the state is CLOSED (i.e., TCB does not exist) then

all data in the incoming segment is discarded. An incoming
segment containing a RST is discarded. An incoming segment not
containing a RST causes a RST to be sent in response. The
acknowledgment and sequence field values are selected to make
the reset sequence acceptable to the TCP that sent the
offending segment.

If the ACK bit is off, sequence number zero is used,

<SEQ=0><ACK=SEG.SEQ+SEG.LEN><CTL=RST,ACK>

If the ACK bit is on,

<SEQ=SEG.ACK><CTL=RST>

Return.

If the state is LISTEN then

first check for an RST

An incoming RST should be ignored. Return.

second check for an ACK

Any acknowledgment is bad if it arrives on a connection
still in the LISTEN state. An acceptable reset segment
should be formed for any arriving ACK-bearing segment. The
RST should be formatted as follows:

<SEQ=SEG.ACK><CTL=RST>

Return.

third check for a SYN

If the SYN bit is set, check the security. If the security/
compartment on the incoming segment does not exactly match
the security/compartment in the TCB then send a reset and
return.

<SEQ=0><ACK=SEG.SEQ+SEG.LEN><CTL=RST,ACK>
If the SEG.PRC is greater than the TCB.PRC then if allowed by the user and the system set TCB.PRC<SEG.PRC, if not allowed send a reset and return.

<SEQ=0><ACK=SEG.SEQ+SEG.LEN><CTL=RST,ACK>

If the SEG.PRC is less than the TCB.PRC then continue.

Set RCV.NXT to SEG.SEQ+1, IRS is set to SEG.SEQ and any other control or text should be queued for processing later. ISS should be selected and a SYN segment sent of the form:

<SEQ=ISS><ACK=RCV.NXT><CTL=SYN,ACK>

SND.NXT is set to ISS+1 and SND.UNA to ISS. The connection state should be changed to SYN-RECEIVED. Note that any other incoming control or data (combined with SYN) will be processed in the SYN-RECEIVED state, but processing of SYN and ACK should not be repeated. If the listen was not fully specified (i.e., the foreign socket was not fully specified), then the unspecified fields should be filled in now.

Any other control or text-bearing segment (not containing SYN) must have an ACK and thus would be discarded by the ACK processing. An incoming RST segment could not be valid, since it could not have been sent in response to anything sent by this incarnation of the connection. So you are unlikely to get here, but if you do, drop the segment, and return.

If the state is SYN-SENT then

first check the ACK bit

If the ACK bit is set

If SEG.ACK =< ISS, or SEG.ACK > SND.NXT, send a reset (unless the RST bit is set, if so drop the segment and return)

<SEQ=SEG.ACK><CTL=RST>

and discard the segment. Return.
If $\text{SND.UNA} < \text{SEG.ACK} =< \text{SND.NXT}$ then the ACK is acceptable. (TODO: in processing Errata ID 3300, it was noted that some stacks in the wild that do not send data on the SYN are just checking that $\text{SEG.ACK} = \text{SND.NXT}$ ... think about whether anything should be said about that here)

Second check the RST bit

If the RST bit is set

If the ACK was acceptable then signal the user "error: connection reset", drop the segment, enter CLOSED state, delete TCB, and return. Otherwise (no ACK) drop the segment and return.

Third check the security and precedence

If the security/compartment in the segment does not exactly match the security/compartment in the TCB, send a reset

If there is an ACK

$<\text{SEQ}=\text{SEG.ACK}><\text{CTL}=\text{RST}>$

Otherwise

$<\text{SEQ}=0><\text{ACK}=\text{SEG.SEQ}+\text{SEG.LEN}><\text{CTL}=\text{RST,ACK}>$

If there is an ACK

The precedence in the segment must match the precedence in the TCB, if not, send a reset

$<\text{SEQ}=\text{SEG.ACK}><\text{CTL}=\text{RST}>$

If there is no ACK

If the precedence in the segment is higher than the precedence in the TCB then if allowed by the user and the system raise the precedence in the TCB to that in the segment, if not allowed to raise the prec then send a reset.

$<\text{SEQ}=0><\text{ACK}=\text{SEG.SEQ}+\text{SEG.LEN}><\text{CTL}=\text{RST,ACK}>$

If the precedence in the segment is lower than the precedence in the TCB continue.
If a reset was sent, discard the segment and return.

fourth check the SYN bit

This step should be reached only if the ACK is ok, or there is no ACK, and it the segment did not contain a RST.

If the SYN bit is on and the security/compartment and precedence are acceptable then, RCV.NXT is set to SEG.SEQ+1, IRS is set to SEG.SEQ. SND.UNA should be advanced to equal SEG.ACK (if there is an ACK), and any segments on the retransmission queue which are thereby acknowledged should be removed.

If SND.UNA > ISS (our SYN has been ACKed), change the connection state to ESTABLISHED, form an ACK segment

<SEQ=SND.NXT><ACK=RCV.NXT><CTL=ACK>

and send it. Data or controls which were queued for transmission may be included. If there are other controls or text in the segment then continue processing at the sixth step below where the URG bit is checked, otherwise return.

Otherwise enter SYN-RECEIVED, form a SYN,ACK segment

<SEQ=ISS><ACK=RCV.NXT><CTL=SYN,ACK>

and send it. Set the variables:

\[
\begin{align*}
\text{SND.WND} & \leftarrow \text{SEG.WND} \\
\text{SND.WL1} & \leftarrow \text{SEG.SEQ} \\
\text{SND.WL2} & \leftarrow \text{SEG.ACK}
\end{align*}
\]

If there are other controls or text in the segment, queue them for processing after the ESTABLISHED state has been reached, return.

fifth, if neither of the SYN or RST bits is set then drop the segment and return.

Otherwise,

first check sequence number

SYN-RECEIVED STATE
ESTABLISHED STATE
FIN-WAIT-1 STATE
FIN-WAIT-2 STATE  
CLOSE-WAIT STATE  
CLOSING STATE  
LAST-ACK STATE  
TIME-WAIT STATE  

Segments are processed in sequence. Initial tests on arrival are used to discard old duplicates, but further processing is done in SEG.SEQ order. If a segment’s contents straddle the boundary between old and new, only the new parts should be processed.

There are four cases for the acceptability test for an incoming segment:

<table>
<thead>
<tr>
<th>Segment Length</th>
<th>Receive Window</th>
<th>Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>SEG.SEQ = RCV.NXT</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>&gt;0</td>
<td>RCV.NXT =&lt; SEG.SEQ &lt; RCV.NXT+RCV.WND</td>
<td></td>
</tr>
<tr>
<td>&gt;0</td>
<td>0</td>
<td>not acceptable</td>
<td></td>
</tr>
<tr>
<td>&gt;0</td>
<td>&gt;0</td>
<td>RCV.NXT =&lt; SEG.SEQ &lt; RCV.NXT+RCV.WND or RCV.NXT =&lt; SEG.SEQ+SEG.LEN-1 &lt; RCV.NXT+RCV.WND</td>
<td></td>
</tr>
</tbody>
</table>

If the RCV.WND is zero, no segments will be acceptable, but special allowance should be made to accept valid ACKs, URGs and RSTs.

If an incoming segment is not acceptable, an acknowledgment should be sent in reply (unless the RST bit is set, if so drop the segment and return):

```
<SEQ=SND.NXT><ACK=RCV.NXT><CTL=ACK>
```

After sending the acknowledgment, drop the unacceptable segment and return.

In the following it is assumed that the segment is the idealized segment that begins at RCV.NXT and does not exceed the window. One could tailor actual segments to fit this assumption by trimming off any portions that lie outside the window (including SYN and FIN), and only processing further
if the segment then begins at RCV.NXT. Segments with higher beginning sequence numbers should be held for later processing.

second check the RST bit,

SYN-RECEIVED STATE

If the RST bit is set

If this connection was initiated with a passive OPEN (i.e., came from the LISTEN state), then return this connection to LISTEN state and return. The user need not be informed. If this connection was initiated with an active OPEN (i.e., came from SYN-SENT state) then the connection was refused, signal the user "connection refused". In either case, all segments on the retransmission queue should be removed. And in the active OPEN case, enter the CLOSED state and delete the TCB, and return.

ESTABLISHED
FIN-WAIT-1
FIN-WAIT-2
CLOSE-WAIT

If the RST bit is set then, any outstanding RECEIVEs and SEND should receive "reset" responses. All segment queues should be flushed. Users should also receive an unsolicited general "connection reset" signal. Enter the CLOSED state, delete the TCB, and return.

CLOSING STATE
LAST-ACK STATE
TIME-WAIT

If the RST bit is set then, enter the CLOSED state, delete the TCB, and return.

third check security and precedence

SYN-RECEIVED

If the security/compartment and precedence in the segment do not exactly match the security/compartment and precedence in the TCB then send a reset, and return.
ESTABLISHED
FIN-WAIT-1
FIN-WAIT-2
CLOSE-WAIT
CLOSING
LAST-ACK
TIME-WAIT

If the security/compartment and precedence in the segment do not exactly match the security/compartment and precedence in the TCB then send a reset, any outstanding RECEIVEs and SEND should receive "reset" responses. All segment queues should be flushed. Users should also receive an unsolicited general "connection reset" signal. Enter the CLOSED state, delete the TCB, and return.

Note this check is placed following the sequence check to prevent a segment from an old connection between these ports with a different security or precedence from causing an abort of the current connection.

fourth, check the SYN bit,

SYN-RECEIVED
ESTABLISHED STATE
FIN-WAIT STATE-1
FIN-WAIT STATE-2
CLOSE-WAIT STATE
CLOSING STATE
LAST-ACK STATE
TIME-WAIT STATE

TODO: need to incorporate RFC 1122 4.2.2.20(e) here

If the SYN is in the window it is an error, send a reset, any outstanding RECEIVEs and SEND should receive "reset" responses, all segment queues should be flushed, the user should also receive an unsolicited general "connection reset" signal, enter the CLOSED state, delete the TCB, and return.

If the SYN is not in the window this step would not be reached and an ack would have been sent in the first step (sequence number check).

fifth check the ACK field,

if the ACK bit is off drop the segment and return
if the ACK bit is on

SYN-RECEIVED STATE

If SND.UNA < SEG.ACK <= SND.NXT then enter ESTABLISHED state and continue processing with variables below set to:

SND.WND <- SEG.WND
SND.WL1 <- SEG.SEQ
SND.WL2 <- SEG.ACK

If the segment acknowledgment is not acceptable, form a reset segment,

<SEQ=SEG.ACK><CTL=RST>

and send it.

ESTABLISHED STATE

If SND.UNA < SEG.ACK <= SND.NXT then, set SND.UNA <- SEG.ACK. Any segments on the retransmission queue which are thereby entirely acknowledged are removed. Users should receive positive acknowledgments for buffers which have been SENT and fully acknowledged (i.e., SEND buffer should be returned with "ok" response). If the ACK is a duplicate (SEG.ACK <= SND.UNA), it can be ignored. If the ACK acks something not yet sent (SEG.ACK > SND.NXT) then send an ACK, drop the segment, and return.

If SND.UNA <= SEG.ACK <= SND.NXT, the send window should be updated. If (SND.WL1 < SEG.SEQ or (SND.WL1 = SEG.SEQ and SND.WL2 <= SEG.ACK)), set SND.WND <- SEG.WND, set SND.WL1 <- SEG.SEQ, and set SND.WL2 <- SEG.ACK.

Note that SND.WND is an offset from SND.UNA, that SND.WL1 records the sequence number of the last segment used to update SND.WND, and that SND.WL2 records the acknowledgment number of the last segment used to update SND.WND. The check here prevents using old segments to update the window.

FIN-WAIT-1 STATE
In addition to the processing for the ESTABLISHED state, if our FIN is now acknowledged then enter FIN-WAIT-2 and continue processing in that state.

FIN-WAIT-2 STATE

In addition to the processing for the ESTABLISHED state, if the retransmission queue is empty, the user’s CLOSE can be acknowledged ("ok") but do not delete the TCB.

CLOSE-WAIT STATE

Do the same processing as for the ESTABLISHED state.

CLOSING STATE

In addition to the processing for the ESTABLISHED state, if the ACK acknowledges our FIN then enter the TIME-WAIT state, otherwise ignore the segment.

LAST-ACK STATE

The only thing that can arrive in this state is an acknowledgment of our FIN. If our FIN is now acknowledged, delete the TCB, enter the CLOSED state, and return.

TIME-WAIT STATE

The only thing that can arrive in this state is a retransmission of the remote FIN. Acknowledge it, and restart the 2 MSL timeout.

sixth, check the URG bit,

ESTABLISHED STATE
FIN-WAIT-1 STATE
FIN-WAIT-2 STATE

If the URG bit is set, $\text{RCV.UP} \leftarrow \max(\text{RCV.UP}, \text{SEG.UP})$, and signal the user that the remote side has urgent data if the urgent pointer (RCV.UP) is in advance of the data consumed. If the user has already been signaled (or is still in the "urgent mode") for this continuous sequence of urgent data, do not signal the user again.

CLOSE-WAIT STATE
CLOSING STATE
LAST-ACK STATE
TIME-WAIT

This should not occur, since a FIN has been received from the remote side. Ignore the URG.

seventh, process the segment text,

ESTABLISHED STATE
FIN-WAIT-1 STATE
FIN-WAIT-2 STATE

Once in the ESTABLISHED state, it is possible to deliver segment text to user RECEIVE buffers. Text from segments can be moved into buffers until either the buffer is full or the segment is empty. If the segment empties and carries an PUSH flag, then the user is informed, when the buffer is returned, that a PUSH has been received.

When the TCP takes responsibility for delivering the data to the user it must also acknowledge the receipt of the data.

Once the TCP takes responsibility for the data it advances RCV.NXT over the data accepted, and adjusts RCV.WND as appropriate to the current buffer availability. The total of RCV.NXT and RCV.WND should not be reduced.

Please note the window management suggestions in section 3.7.

Send an acknowledgment of the form:

<SEQ=SND.NXT><ACK=RCV.NXT><CTL=ACK>

This acknowledgment should be piggybacked on a segment being transmitted if possible without incurring undue delay.

CLOSE-WAIT STATE
CLOSING STATE
LAST-ACK STATE
TIME-WAIT STATE

This should not occur, since a FIN has been received from the remote side. Ignore the segment text.
eighth, check the FIN bit,

Do not process the FIN if the state is CLOSED, LISTEN or SYN-SENT since the SEG.SEQ cannot be validated; drop the segment and return.

If the FIN bit is set, signal the user "connection closing" and return any pending RECEIVEs with same message, advance RCV.NXT over the FIN, and send an acknowledgment for the FIN. Note that FIN implies PUSH for any segment text not yet delivered to the user.

SYN-RECEIVED STATE
ESTABLISHED STATE

Enter the CLOSE-WAIT state.

FIN-WAIT-1 STATE

If our FIN has been ACKed (perhaps in this segment), then enter TIME-WAIT, start the time-wait timer, turn off the other timers; otherwise enter the CLOSING state.

FIN-WAIT-2 STATE

Enter the TIME-WAIT state. Start the time-wait timer, turn off the other timers.

CLOSE-WAIT STATE

Remain in the CLOSE-WAIT state.

CLOSING STATE

Remain in the CLOSING state.

LAST-ACK STATE

Remain in the LAST-ACK state.

TIME-WAIT STATE

Remain in the TIME-WAIT state. Restart the 2 MSL time-wait timeout.

and return.
USER TIMEOUT

For any state if the user timeout expires, flush all queues, signal the user "error: connection aborted due to user timeout" in general and for any outstanding calls, delete the TCB, enter the CLOSED state and return.

RETRANSMISSION TIMEOUT

For any state if the retransmission timeout expires on a segment in the retransmission queue, send the segment at the front of the retransmission queue again, reinitialize the retransmission timer, and return.

TIME-WAIT TIMEOUT

If the time-wait timeout expires on a connection delete the TCB, enter the CLOSED state and return.
3.11. Glossary

1822  BBN Report 1822, "The Specification of the Interconnection of a Host and an IMP". The specification of interface between a host and the ARPANET.

ACK  
A control bit (acknowledge) occupying no sequence space, which indicates that the acknowledgment field of this segment specifies the next sequence number the sender of this segment is expecting to receive, hence acknowledging receipt of all previous sequence numbers.

ARPANET message  
The unit of transmission between a host and an IMP in the ARPANET. The maximum size is about 1012 octets (8096 bits).

ARPANET packet  
A unit of transmission used internally in the ARPANET between IMPs. The maximum size is about 126 octets (1008 bits).

connection  
A logical communication path identified by a pair of sockets.

datagram  
A message sent in a packet switched computer communications network.

Destination Address  
The destination address, usually the network and host identifiers.

FIN  
A control bit (finis) occupying one sequence number, which indicates that the sender will send no more data or control occupying sequence space.

fragment  
A portion of a logical unit of data, in particular an internet fragment is a portion of an internet datagram.

FTP  
A file transfer protocol.

header  
Control information at the beginning of a message, segment, fragment, packet or block of data.
host
A computer. In particular a source or destination of
messages from the point of view of the communication network.

Identification
An Internet Protocol field. This identifying value assigned
by the sender aids in assembling the fragments of a datagram.

IMP
The Interface Message Processor, the packet switch of the
ARPANET.

internet address
A source or destination address specific to the host level.

internet datagram
The unit of data exchanged between an internet module and the
higher level protocol together with the internet header.

internet fragment
A portion of the data of an internet datagram with an
internet header.

IP
Internet Protocol.

IRS
The Initial Receive Sequence number. The first sequence
number used by the sender on a connection.

ISN
The Initial Sequence Number. The first sequence number used
on a connection, (either ISS or IRS). Selected in a way that
is unique within a given period of time and is unpredictable
to attackers.

ISS
The Initial Send Sequence number. The first sequence number
used by the sender on a connection.

leader
Control information at the beginning of a message or block of
data. In particular, in the ARPANET, the control information
on an ARPANET message at the host-IMP interface.

left sequence
This is the next sequence number to be acknowledged by the
data receiving TCP (or the lowest currently unacknowledged
sequence number) and is sometimes referred to as the left edge of the send window.

local packet
The unit of transmission within a local network.

module
An implementation, usually in software, of a protocol or other procedure.

MSL
Maximum Segment Lifetime, the time a TCP segment can exist in the internetwork system. Arbitrarily defined to be 2 minutes.

octet
An eight bit byte.

Options
An Option field may contain several options, and each option may be several octets in length. The options are used primarily in testing situations; for example, to carry timestamps. Both the Internet Protocol and TCP provide for options fields.

packet
A package of data with a header which may or may not be logically complete. More often a physical packaging than a logical packaging of data.

port
The portion of a socket that specifies which logical input or output channel of a process is associated with the data.

process
A program in execution. A source or destination of data from the point of view of the TCP or other host-to-host protocol.

PUSH
A control bit occupying no sequence space, indicating that this segment contains data that must be pushed through to the receiving user.

RCV.NXT
receive next sequence number

RCV.UP
receive urgent pointer
RCV.WND
receive window
receive next sequence number
This is the next sequence number the local TCP is expecting to receive.

receive window
This represents the sequence numbers the local (receiving) TCP is willing to receive. Thus, the local TCP considers that segments overlapping the range RCV.NXT to RCV.NXT + RCV.WND - 1 carry acceptable data or control. Segments containing sequence numbers entirely outside of this range are considered duplicates and discarded.

RST
A control bit (reset), occupying no sequence space, indicating that the receiver should delete the connection without further interaction. The receiver can determine, based on the sequence number and acknowledgment fields of the incoming segment, whether it should honor the reset command or ignore it. In no case does receipt of a segment containing RST give rise to a RST in response.

RTP
Real Time Protocol: A host-to-host protocol for communication of time critical information.

SEG.ACK
segment acknowledgment

SEG.LEN
segment length

SEG.PRC
segment precedence value

SEGSEQ
segment sequence

SEG.UP
segment urgent pointer field

SEG.WND
segment window field

segment
A logical unit of data, in particular a TCP segment is the unit of data transfered between a pair of TCP modules.

segment acknowledgment
The sequence number in the acknowledgment field of the arriving segment.

segment length
The amount of sequence number space occupied by a segment, including any controls which occupy sequence space.

segment sequence
The number in the sequence field of the arriving segment.

send sequence
This is the next sequence number the local (sending) TCP will use on the connection. It is initially selected from an initial sequence number curve (ISN) and is incremented for each octet of data or sequenced control transmitted.

send window
This represents the sequence numbers which the remote (receiving) TCP is willing to receive. It is the value of the window field specified in segments from the remote (data receiving) TCP. The range of new sequence numbers which may be emitted by a TCP lies between SND.NXT and SND.UNA + SND.WND - 1. (Retransmissions of sequence numbers between SND.UNA and SND.NXT are expected, of course.)

SND.NXT
send sequence
SND.UNA
left sequence
SND.UP
send urgent pointer
SND.WL1
segment sequence number at last window update
SND.WL2
segment acknowledgment number at last window update
SND.WND
send window

socket
An address which specifically includes a port identifier, that is, the concatenation of an Internet Address with a TCP port.

Source Address
The source address, usually the network and host identifiers.

SYN
A control bit in the incoming segment, occupying one sequence number, used at the initiation of a connection, to indicate where the sequence numbering will start.

TCB
Transmission control block, the data structure that records the state of a connection.

TCB.PRC
The precedence of the connection.

TCP
Transmission Control Protocol: A host-to-host protocol for reliable communication in internetwork environments.

TOS
Type of Service, an Internet Protocol field.

Type of Service
An Internet Protocol field which indicates the type of service for this internet fragment.

URG
A control bit (urgent), occupying no sequence space, used to indicate that the receiving user should be notified to do urgent processing as long as there is data to be consumed with sequence numbers less than the value indicated in the urgent pointer.

urgent pointer
A control field meaningful only when the URG bit is on. This field communicates the value of the urgent pointer which indicates the data octet associated with the sending user’s urgent call.

4. Changes from RFC 793

This document obsoletes RFC 793 as well as RFC 6093 and 6528, which updated 793. In all cases, only the normative protocol specification and requirements have been incorporated into this document, and the
informational text with background and rationale has not been carried in. The informational content of those documents is still valuable in learning about and understanding TCP, and they are valid Informational references, even though their normative content has been incorporated into this document.

The main body of this document was adapted from RFC 793’s Section 3, titled "FUNCTIONAL SPECIFICATION", with an attempt to keep formatting and layout as close as possible.

The collection of applicable RFC Errata that have been reported and either accepted or held for an update to RFC 793 were incorporated (Errata IDs: 573, 574, 700, 701, 1283, 1561, 1562, 1564, 1565, 1571, 1572, 2296, 2297, 2298, 2748, 2749, 2934, 3213, 3300, 3301). Some errata were not applicable due to other changes (Errata IDs: 572, 575, 1569, 3602). TODO: 3305

Changes to the specification of the Urgent Pointer described in RFC 1122 and 6093 were incorporated. See RFC 6093 for detailed discussion of why these changes were necessary.

The more secure Initial Sequence Number generation algorithm from RFC 6528 was incorporated. See RFC 6528 for discussion of the attacks that this mitigates, as well as advice on selecting PRF algorithms and managing secret key data.

RFC EDITOR’S NOTE: the content below is for detailed change tracking and planning, and not to be included with the final revision of the document.

This document started as draft-eddy-rfc793bis-00, that was merely a proposal and rough plan for updating RFC 793.

The -01 revision of this document incorporates the content of RFC 793 Section 3 titled "FUNCTIONAL SPECIFICATION". Other content from RFC 793 has not been incorporated. The -01 revision of this document makes some minor formatting changes to the RFC 793 content in order to convert the content into XML2RFC format and account for left-out parts of RFC 793. For instance, figure numbering differs and some indentation is not exactly the same.

The -02 revision of draft-eddy-rfc793bis incorporates errata that have been verified:

    Errata ID 573: Reported by Bob Braden (note: This errata basically is just a reminder that RFC 1122 updates 793. Some of the associated changes are left pending to a separate revision that incorporates 1122. Bob’s mention of PUSH in 793 section 2.8 was
not applicable here because that section was not part of the "functional specification". Also the 1122 text on the retransmission timeout also has been updated by subsequent RFCs, so the change here deviates from Bob’s suggestion to apply the 1122 text.)

Errata ID 574: Reported by Yin Shuming
Errata ID 700: Reported by Yin Shuming
Errata ID 701: Reported by Yin Shuming
Errata ID 1283: Reported by Pei-chun Cheng
Errata ID 1561: Reported by Constantin Hagemeier
Errata ID 1562: Reported by Constantin Hagemeier
Errata ID 1564: Reported by Constantin Hagemeier
Errata ID 1565: Reported by Constantin Hagemeier
Errata ID 1571: Reported by Constantin Hagemeier
Errata ID 1572: Reported by Constantin Hagemeier
Errata ID 2296: Reported by Vishwas Manral
Errata ID 2297: Reported by Vishwas Manral
Errata ID 2298: Reported by Vishwas Manral
Errata ID 2748: Reported by Mykyta Yevstifeyev
Errata ID 2749: Reported by Mykyta Yevstifeyev
Errata ID 2934: Reported by Constantin Hagemeier
Errata ID 3213: Reported by EugnJun Yi
Errata ID 3300: Reported by Botong Huang
Errata ID 3301: Reported by Botong Huang

Note: Some verified errata were not used in this update, as they relate to sections of RFC 793 elided from this document. These include Errata ID 572, 575, and 1569.

Note: Errata ID 3602 was not applied in this revision as it is duplicative of the 1122 corrections.

There is an errata 3305 currently reported that need to be verified, held, or rejected by the ADs; it is addressing the same issue as draft-gont-tcpm-tcp-seq-validation and was not attempted to be applied to this document.

Not related to RFC 793 content, this revision also makes small tweaks to the introductory text, fixes indentation of the pseudoheader diagram, and notes that the Security Considerations should also include privacy, when this section is written.

The -03 revision of draft-eddy-rfc793bis revises all discussion of the urgent pointer in order to comply with RFC 6093, 1122, and 1011. Since 1122 held requirements on the urgent pointer, the full list of requirements was brought into an appendix of this document, so that it can be updated as-needed.

The -04 revision of draft-eddy-rfc793bis includes the ISN generation changes from RFC 6528.
The -05 revision of draft-eddy-rfc793bis incorporates MSS requirements and definitions from RFC 879, 1122, and 6691, as well as option-handling requirements from RFC 1122.

The -00 revision of draft-ietf-tcpm-rfc793bis incorporates several additional clarifications and updates to the section on segmentation, many of which are based on feedback from Joe Touch improving from the initial text on this in the previous revision.

The -01 revision incorporates the change to Reserved bits due to ECN, as well as many other changes that come from RFC 1122.

TODO: Incomplete list of other planned changes - these can be added to and made more specific, as the document proceeds:

1. incorporate all other 1122 additions (sections on Data Communication, Retransmission Timeout, Managing the Window, Probing Zero Windows, Passive OPEN Calls, Time to Live, Event Processing, Acknowledging Queued Segments, Retransmission Timeout Calculation, When to Send an ACK Segment, When to Send a Window Update, When to Send Data, TCP Connection Failures, TCP Keep- Alives, TCP Multihoming, IP options, ICMP messages, remote address validation)
2. point to major additional docs like 1323bis and 5681
3. incorporate relevant parts of 3168 (ECN) - beyond just indicating the names of the 2 bits already done
4. incorporate Fernando’s new number-checking fixes (if past the IESG in time)
5. point to 5461 (soft errors)
6. mention 5961 state machine option
7. mention 6161 (reducing TIME-WAIT)
8. incorporate 6429 (ZWP/persist)
9. look at Tony Sabatini suggestion for describing DO field
10. clearly specify treatment of reserved bits (see TCPM thread on EDO draft April 25, 2014)
11. look at possible mention of draft-minshall-nagle (e.g. as in Linux)
12. make sure that clarifications in RFC 1011 are captured
13. per TCPM discussion, discussion of checking reserved bits may need to be altered from 793
14. per discussion with Joe Touch (TAPS list, 6/20/2015), the description of the API could be revisited
15. there is inconsistency between use of SYN_RCVD and SYNC-RECEIVED in diagrams and text in various places
16. TOS material does not take DSCP changes into account
17. discuss with working group whether to include anything like section 4.2.3.12 of 1122 (on "efficiency" ... basically
implementation advice), maybe similar to 2525 in handling for this document. also 4.2.3.11 on "TCP Traffic Patterns"

5.  IANA Considerations

This memo includes no request to IANA. Existing IANA registries for TCP parameters are sufficient.

TODO: check whether entries pointing to 793 and other documents obsoleted by this one should be updated to point to this one instead.

6.  Security and Privacy Considerations

TODO


Editor’s Note: Scott Brim mentioned that this should include a PERPASS/privacy review.

7.  Acknowledgements

This document is largely a revision of RFC 793, which Jon Postel was the editor of. Due to his excellent work, it was able to last for three decades before we felt the need to revise it.

Andre Oppermann was a contributor and helped to edit the first revision of this document.

We are thankful for the assistance of the IETF TCPM working group chairs:

Michael Scharf
Yoshifumi Nishida
Pasi Sarolahti

During early discussion of this work on the TCPM mailing list, and at the IETF 88 meeting in Vancouver, helpful comments, critiques, and reviews were received from (listed alphabetically): David Borman, Yuchung Cheng, Martin Duke, Kevin Lahey, Kevin Mason, Matt Mathis, Hagen Paul Pfeifer, Anthony Sabatini, Joe Touch, Reji Varghese, Lloyd Wood, and Alex Zimmermann.

This document includes content from errata that were reported by (listed chronologically): Yin Shuming, Bob Braden, Morris M. Keesan, Pei-chun Cheng, Constantin Hagemeier, Vishwas Manral, Mykyta Yevstifeyev, EungJun Yi, Botong Huang.
8. References

8.1. Normative References


8.2. Informative References


Appendix A. TCP Requirement Summary

This section is adapted from RFC 1122.

TODO: this needs to be seriously redone, to use 793bis section numbers instead of 1122 ones, the RFC1122 heading should be removed, and all 1122 requirements need to be reflected in 793bis text.

TODO: NOTE that PMTUD+PLPMTUD is not included in this table of recommendations.
<table>
<thead>
<tr>
<th>FEATURE</th>
<th>SECTION</th>
<th>T</th>
<th>T</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Push flag</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggregate or queue un-pushed data</td>
<td>4.2.2.2</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sender collapse successive PSH flags</td>
<td>4.2.2.2</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEND call can specify PUSH</td>
<td>4.2.2.2</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If cannot: sender buffer indefinitely</td>
<td>4.2.2.2</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If cannot: PSH last segment</td>
<td>4.2.2.2</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notify receiving ALP of PSH</td>
<td>4.2.2.2</td>
<td>x</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Send max size segment when possible</td>
<td>4.2.2.2</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Window</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treat as unsigned number</td>
<td>4.2.2.3</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handle as 32-bit number</td>
<td>4.2.2.3</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shrink window from right</td>
<td>4.2.2.16</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robust against shrinking window</td>
<td>4.2.2.16</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Receiver’s window closed indefinitely</td>
<td>4.2.2.17</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sender probe zero window</td>
<td>4.2.2.17</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First probe after RTO</td>
<td>4.2.2.17</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential backoff</td>
<td>4.2.2.17</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allow window stay zero indefinitely</td>
<td>4.2.2.17</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sender timeout OK conn with zero wind</td>
<td>4.2.2.17</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urgent Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pointer indicates first non-urgent octet</td>
<td>4.2.2.4</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arbitrary length urgent data sequence</td>
<td>4.2.2.4</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inform ALP asynchronously of urgent data</td>
<td>4.2.2.4</td>
<td>x</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ALP can learn if/how much urgent data Q’d</td>
<td>4.2.2.4</td>
<td>x</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>TCP Options</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Receive TCP option in any segment</td>
<td>4.2.2.5</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ignore unsupported options</td>
<td>4.2.2.5</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cope with illegal option length</td>
<td>4.2.2.5</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implement sending &amp; receiving MSS option</td>
<td>4.2.2.6</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPv4 Send MSS option unless 536</td>
<td>4.2.2.6</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPv6 Send MSS option unless 1220</td>
<td>N/A</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Send MSS option always</td>
<td>4.2.2.6</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPv4 Send-MSS default is 536</td>
<td>4.2.2.6</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPv6 Send-MSS default is 1220</td>
<td>N/A</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calculate effective send seg size</td>
<td>4.2.2.6</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSS accounts for varying MTU</td>
<td>N/A</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCP Checksums</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sender compute checksum</td>
<td>4.2.2.7</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Receiver check checksum</td>
<td>4.2.2.7</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISN Selection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Eddy Expires March 24, 2016 [Page 83]
| **Include a clock-driven ISN generator component** | 4.2.2.9 | x |   |
| **Secure ISN generator with a PRF component** | N/A | x |   |

**Opening Connections**

- Support simultaneous open attempts | 4.2.2.10 | x |   |
- SYN-RCVD remembers last state | 4.2.2.11 | x |   |
- Passive Open call interfere with others | 4.2.2.18 | x |   |
- Function: simultan. LISTENs for same port | 4.2.2.18 | x |   |
- Ask IP for src address for SYN if necc. | 4.2.3.7 | x |   |
- Otherwise, use local addr of conn. | 4.2.3.7 | x |   |
- OPEN to broadcast/multicast IP Address | 4.2.3.14 | x |   |
- Silently discard seg to bcast/mcast addr | 4.2.3.14 | x |   |

**Closing Connections**

- RST can contain data | 4.2.2.12 | x |   |
- Inform application of aborted conn | 4.2.2.13 | x |   |
- Half-duplex close connections | 4.2.2.13 | x |   |
- Send RST to indicate data lost | 4.2.2.13 | x |   |
- In TIME-WAIT state for 2MSL seconds | 4.2.2.13 | x |   |
- Accept SYN from TIME-WAIT state | 4.2.2.13 | x |   |

**Retransmissions**

- Jacobson Slow Start algorithm | 4.2.2.15 | x |   |
- Jacobson Congestion-Avoidance algorithm | 4.2.2.15 | x |   |
- Retransmit with same IP ident | 4.2.2.15 | x |   |
- Karn’s algorithm | 4.2.3.1 | x |   |
- Jacobson’s RTO estimation alg. | 4.2.3.1 | x |   |
- Exponential backoff | 4.2.3.1 | x |   |
- SYN RTO calc same as data | 4.2.3.1 | x |   |
- Recommended initial values and bounds | 4.2.3.1 | x |   |

**Generating ACK’s:**

- Queue out-of-order segments | 4.2.2.20 | x |   |
- Process all Q’d before send ACK | 4.2.2.20 | x |   |
- Send ACK for out-of-order segment | 4.2.2.21 | x |   |
- Delayed ACK’s | 4.2.3.2 | x |   |
- Delay < 0.5 seconds | 4.2.3.2 | x |   |
- Every 2nd full-sized segment ACK’d | 4.2.3.2 | x |   |
- Receiver SWS-Avoidance Algorithm | 4.2.3.3 | x |   |

**Sending data**

- Configurable TTL | 4.2.2.19 | x |   |
- Sender SWS-Avoidance Algorithm | 4.2.3.4 | x |   |
- Nagle algorithm | 4.2.3.4 | x |   |
- Application can disable Nagle algorithm | 4.2.3.4 | x |   |

**Connection Failures:**

- Negative advice to IP on R1 retxs | 4.2.3.5 | x |   |
Close connection on R2 retxs | 4.2.3.5 | x |
ALP can set R2 | 4.2.3.5 | x |
Inform ALP of R1<=retxs<R2 | 4.2.3.5 | x |
Recommended values for R1, R2 | 4.2.3.5 | x |
Same mechanism for SYN | 4.2.3.5 | x |
R2 at least 3 minutes for SYN | 4.2.3.5 | x |

Send Keep-alive Packets: | 4.2.3.6 | x |
- Application can request | 4.2.3.6 | x |
- Default is "off" | 4.2.3.6 | x |
- Only send if idle for interval | 4.2.3.6 | x |
- Interval configurable | 4.2.3.6 | x |
- Default at least 2 hrs. | 4.2.3.6 | x |
- Tolerant of lost ACK’s | 4.2.3.6 | x |

IP Options | 4.2.3.8 | x |
- Ignore options TCP doesn’t understand | 4.2.3.8 | x |
- Time Stamp support | 4.2.3.8 | x |
- Record Route support | 4.2.3.8 | x |
- Source Route: | 4.2.3.8 | x |
  ALP can specify | 4.2.3.8 | x |
  Overrides src rt in datagram | 4.2.3.8 | x |
  Build return route from src rt | 4.2.3.8 | x |
  Later src route overrides | 4.2.3.8 | x |

Receiving ICMP Messages from IP | 4.2.3.9 | x |
- Dest. Unreach (0,1,5) => inform ALP | 4.2.3.9 | x |
- Dest. Unreach (0,1,5) => abort conn | 4.2.3.9 | x |
- Dest. Unreach (2-4) => abort conn | 4.2.3.9 | x |
- Source Quench => slow start | 4.2.3.9 | x |
- Time Exceeded => tell ALP, don’t abort | 4.2.3.9 | x |
- Param Problem => tell ALP, don’t abort | 4.2.3.9 | x |

Address Validation | 4.2.3.10 | x |
- Reject OPEN call to invalid IP address | 4.2.3.10 | x |
- Reject SYN from invalid IP address | 4.2.3.10 | x |
- Silently discard SYN to bcast/mcast addr | 4.2.3.10 | x |

TCP/ALP Interface Services | 4.2.4.1 | x |
- Error Report mechanism | 4.2.4.1 | x |
- ALP can disable Error Report Routine | 4.2.4.2 | x |
- ALP can specify TOS for sending | 4.2.4.2 | x |
  Passed unchanged to IP | 4.2.4.2 | x |
  ALP can change TOS during connection | 4.2.4.2 | x |
  Pass received TOS up to ALP | 4.2.4.2 | x |
  FLUSH call | 4.2.4.3 | x |
- Optional local IP addr parm. in OPEN | 4.2.4.4 | x |
FOOTNOTES: (1) "ALP" means Application-Layer program.

Author’s Address

Wesley M. Eddy (editor)
MTI Systems
US

Email: wes@mti-systems.com
TCPM WG
Internet Draft
Updates: 793
Intended status: Standards Track
Expires: May 2016

TCP Extended Data Offset Option
draft-ietf-tcpm-tcp-edo-04.txt

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), its areas, and its working groups. Note that
other groups may also distribute working documents as Internet-
Drafts.

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."

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html

This Internet-Draft will expire on April 2, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the
document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal
Provisions Relating to IETF Documents
(http://trustee.ietf.org/license-info) in effect on the date of
publication of this document. Please review these documents
carefully, as they describe your rights and restrictions with
respect to this document. Code Components extracted from this
document must include Simplified BSD License text as described in
Abstract

TCP segments include a Data Offset field to indicate space for TCP options but the size of the field can limit the space available for complex options such as SACK and Multipath TCP and can limit the combination of such options supported in a single connection. This document updates RFC 793 with an optional TCP extension to that space to support the use of multiple large options. It also explains why the initial SYN of a connection cannot be extending a single segment.

Table of Contents

1. Introduction...................................................3
2. Conventions used in this document..............................3
3. Motivation.....................................................3
4. Requirements for Extending TCP’s Data Offset...................4
5. The TCP EDO Option.............................................4
   5.1. EDO Supported.............................................5
   5.2. EDO Extension.............................................5
   5.3. The two EDO Extension variants............................8
6. TCP EDO Interaction with TCP...................................9
   6.1. TCP User Interface........................................9
   6.2. TCP States and Transitions................................9
   6.3. TCP Segment Processing...................................10
   6.4. Impact on TCP Header Size................................10
   6.5. Connectionless Resets....................................11
   6.6. ICMP Handling............................................11
7. Interactions with Middleboxes.................................12
   7.1. Middlebox Coexistence with EDO...........................12
   7.2. Middlebox Interference with EDO..........................13
8. Comparison to Previous Proposals..............................14
   8.1. EDO Criteria.............................................14
   8.2. Summary of Approaches....................................15
   8.3. Extended Segments.......................................16
   8.4. TCPx2....................................................16
   8.5. LO/SLO...................................................17
   8.6. LOIC.....................................................17
   8.7. Problems with Extending the Initial SYN.................18
9. Implementation Issues..........................................19
10. Security Considerations......................................20
11. IANA Considerations.........................................20
12. References...................................................20
1. Introduction

TCP’s Data Offset (DO) is a 4-bit field, which indicates the number of 32-bit words of the entire TCP header [RFC793]. This limits the current total header size to 60 bytes, of which the basic header occupies 20, leaving 40 bytes for options. These 40 bytes are increasingly becoming a limitation to the development of advanced capabilities, such as when SACK [RFC2018][RFC6675] is combined with either Multipath TCP [RFC6824], TCP-AO [RFC5925], or TCP Fast Open [RFC7413].

This document specifies the TCP Extended Data Offset (EDO) option, and is independent of (and thus compatible with) IPv4 and IPv6. EDO extends the space available for TCP options, except for the initial SYN and SYN/ACK. This document also explains why the option space of the initial SYN segments cannot be extended as individual segments without severe impact on TCP’s initial handshake and the SYN/ACK limitation that results from potential middlebox misbehavior. Multiple other TCP extensions are being considered in the TCPM working group in order to address the case of SYN and SYN/ACK segments [Bo14][Br14][To15]. Some of these other extensions can work in conjunction with EDO (e.g., [To15]).

2. Conventions used in this document

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 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

3. Motivation

TCP supports headers with a total length of up to 15 32-bit words, as indicated in the 4-bit Data Offset field [RFC793]. This accounts....
for a total of 60 bytes, of which the default TCP header fields occupy 20 bytes, leaving 40 bytes for options.

TCP connections already use this option space for a variety of capabilities. These include Maximum Segment Size (MSS) [RFC793], Window Scale (WS) [RFC7323], Timestamp (TS) [RFC7323], Selective Acknowledgement (SACK) [RFC2018][RFC6675], TCP Authentication Option (TCP-AO) [RFC5925], Multipath TCP (MP-TCP) [RFC6824], and TCP User Timeout [RFC5482]. Some options occur only in a SYN or SYN/ACK (MSS, WS), and others vary in size when used in SYN vs. non-SYN segments.

Each of these options consumes space, where some options consuming as much space as available (SACK) and other desired combinations can easily exceed the currently available space. For example, it is not currently possible to use TCP-AO with both TS and MP-TCP in the same non-SYN segment, i.e., to combine accurate round-trip estimation, authentication, and multipath support in the same connection - even though these options can be negotiated during a SYN exchange (10 for TS, 16 for TCP-AO, and 12 for MP-TCP).

TCP EDO is intended to overcome this limitation for non-SYN segments, as well as to increase the space available for SACK blocks. Further discussion of the impact of EDO and existing options is discussed in Section 6.4. Extending SYN segments is much more complicated, as discussed in Section 8.7.

4. Requirements for Extending TCP’s Data Offset

The primary goal of extending the TCP Data Offset field is to increase the space available for TCP options in all segments except the initial SYN.

An important requirement of any such extension is that it not impact legacy endpoints. Endpoints seeking to use this new option should not incur additional delay or segment exchanges to connect to either new endpoints supporting this option or legacy endpoints without this option. We call this a "backward downgrade" capability.

An additional consideration of this extension is avoiding user data corruption in the presence of popular network devices, including middleboxes. Consideration of middlebox misbehavior can also interfere with extension in the SYN/ACK.

5. The TCP EDO Option

TCP EDO extends the option space for all segments except the initial SYN (i.e., SYN set and ACK not set) and SYN/ACK response. EDO is
indicated by the TCP option codepoint of EDO-OPT and has two types: EDO Supported and EDO Extension, as discussed in the following subsections.

5.1. EDO Supported

EDO capability is determined in both directions using a single exchange of the EDO Supported option (Figure 1). When EDO is desired on a given connection, the SYN and SYN/ACK segments include the EDO Supported option, which consists of the two required TCP option fields: Kind and Length. The EDO Supported option is used only in the SYN and SYN/ACK segments and only to confirm support for EDO in subsequent segments.

```
+--------+--------+
|  Kind  | Length |
+--------+--------+
```

Figure 1 TCP EDO Supported option

An endpoint seeking to enable EDO includes the EDO Supported option in the initial SYN. If receiver of that SYN agrees to use EDO, it responds with the EDO Supported option in the SYN/ACK. The EDO Supported option does not extend the TCP option space.

>> Connections using EDO MUST negotiate its availability during the SYN exchange of the initial three-way handshake.

>> An endpoint confirming and agreeing to EDO use MUST respond with the EDO Supported option in its SYN/ACK.

The SYN/ACK uses only the EDO Supported option (and not the EDO Extension option, below) because it may not yet be safe to extend the option space in the reverse direction due to potential middlebox misbehavior (see Section 7.2). Extension of the SYN and SYN/ACK space is addressed as a separate option (see Section 8.7).

5.2. EDO Extension

When EDO is successfully negotiated, all other segments use the EDO Extension option, of which there are two variants (Figure 2 and Figure 3). Both variants are considered equivalent and either variant can be used in any segment where the EDO Extension option is required. Both variants add a Header_Length field (in network-standard byte order), indicating the length of the entire TCP header in 32-bit words. Figure 3 depicts the longer variant, which includes an additional Segment_Length field, which is identical to the TCP
pseudoheader TCP Length field and used to detect when segments have been altered in ways that would interfere with EDO (discussed further in Section 5.3).

```
+--------+--------+--------+--------+
|  Kind  | Length |  Header_Length |
+--------+--------+--------+--------+
```

Figure 2 TCP EDO Extension option - simple variant

```
+--------+--------+--------+--------+
|  Kind  | Length |  Header_Length |
+--------+--------+--------+--------+
|        |        | Segment_Length |
+--------+--------+
```

Figure 3 TCP EDO Extension option - with segment length verification

>> Once enabled on a connection, all segments in both directions MUST include the EDO Extension option. Segments not needing extension MUST set the EDO Extension option Header Length field equal to the Data Offset length.

>> The EDO Extension option MAY be used only if confirmed when the connection transitions to the ESTABLISHED state, e.g., a client is enabled after receiving the EDO Supported option in the SYN/ACK and the server is enabled after seeing the EDO Extension option in the final ACK of the three-way handshake. If either of those segments lacks the appropriate EDO option, the connection MUST NOT use any EDO options on any other segments.

Internet paths may vary after connection establishment, introducing misbehaving middleboxes (see Section 7.2). Using EDO on all segments in both directions allows this condition to be detected.

>> The EDO Supported option MAY occur in an initial SYN as desired (e.g., as expressed by the user/application) and in the SYN/ACK as confirmation, but MUST NOT be inserted in other segments. If the EDO Supported option is received in other segments, it MUST be silently ignored.

>> If EDO has not been negotiated and agreed, the EDO Extension option MUST be silently ignored on subsequent segments. The EDO Extension option MUST NOT be sent in an initial SYN segment or SYN/ACK, and MUST be silently ignored and not acknowledged if so received.
If EDO has been negotiated, any subsequent segments arriving without the EDO Extension option MUST be silently ignored. Such events MAY be logged as warning errors and logging MUST be rate limited.

When processing a segment, EDO needs to be visible within the area indicated by the Data Offset field, so that processing can use the EDO Header_length to override the field for that segment.

The EDO Extension option MUST occur within the space indicated by the TCP Data Offset.

The EDO Extension option indicates the total length of the header. The EDO Header_length field MUST NOT exceed that of the total segment size (i.e., TCP Length).

The EDO Header Length MUST be at least as large as the TCP Data Offset field of the segment in which they both appear. When the EDO Header Length equals the Data Offset length, the EDO Extension option is present but it does not extend the option space. When the EDO Header Length is invalid, the TCP segment MUST be silently dropped.

The EDO Supported option SHOULD be aligned on a 16-bit boundary and the EDO Extension option SHOULD be aligned on a 32-bit boundary, in both cases for simpler processing.

For example, a segment with only EDO would have a Data Offset of 6 or 7 (depending on the EDO Extension variant used), where EDO would be the first option processed, at which point the EDO Extension option would override the Data Offset and processing would continue until the end of the TCP header as indicated by the EDO Header_length field.

There are cases where it might be useful to process other options before EDO, notably those that determine whether the TCP header is valid, such as authentication, encryption, or alternate checksums. In those cases, the EDO Extension option is preferably the first option after a validation option, and the payload after the Data Offset is treated as user data for the purposes of validation.

The EDO Extension option SHOULD occur as early as possible, either first or just after any authentication or encryption, and SHOULD be the last option covered by the Data Offset value.

Other options are generally handled in the same manner as when the EDO option is not active, unless they interact with other options.
One such example is TCP-AO [RFC5925], which optionally ignores the contents of TCP options, so it would need to be aware of EDO to operate correctly when options are excluded from the HMAC calculation.

Options that depend on other options, such as TCP-AO [RFC5925] (which may include or exclude options in MAC calculations) MUST also be augmented to interpret the EDO Extension option to operate correctly.

5.3. The two EDO Extension variants

There are two variants of the EDO Extension option; one includes a copy of the TCP segment length, copied from the TCP pseudoheader [RFC793]. The Segment_Length field is added to the longer variant to detect when segments are merged by middleboxes or TCP offload processing but without consideration for the additional option space indicated by the EDO Header_Length field. Such effects are described in further detail in Section 7.2.

An endpoint MAY use either variant of the EDO Extension option interchangeably.

When the longer, 6-byte variant is used, the Segment_Length field is used to check whether modification of the segment was performed consistent with knowledge of the EDO option. The Segment_Length field will detect any modification of the length of the segment, such as might occur when segments are split or merged, that occurs without also updating the Segment Length field as well. The Segment Length field thus helps endpoints detects devices that merge or split TCP segments without support for EDO. Devices that merge or split TCP segments that support EDO would update the Segment Length field as needed, but would also ensure that the user data is handled separately from the extended option space indicate by EDO.

When an endpoint creates a new segment using the 6-byte EDO Extension option, the Segment_Length field is initialized with a copy of the segment length from the TCP pseudoheader.

When an endpoint receives a segment using the 6-byte EDO Extension option, it MUST validate the Segment_Length field with the length of the segment as indicated in the TCP pseudoheader. If the segment lengths do not match, the segment MUST be discarded and an error SHOULD be logged in a rate-limited manner.

The 6-byte EDO Extension variant SHOULD be used where middlebox or TCP offload support could merge or split TCP segments without
consideration for the EDO option. Because these conditions could occur at either endpoint or along the network path, the 6-byte variant SHOULD be preferred until sufficient evidence for safe use of the 4-byte variant is determined by the community.

The field will not detect other modification of the TCP user data; such modifications would need more complex detection mechanisms, such as checksums or hashes. When these are used, as with IPsec or TCP-AO, the 4-byte variant is sufficient.

The 4-byte EDO Extension variant is sufficient when EDO is used in conjunction with other mechanisms that provide integrity protection, such as IPsec or TCP-AO.

6. TCP EDO Interaction with TCP

The following subsections describe how EDO interacts with the TCP specification [RFC793].

6.1. TCP User Interface

The TCP EDO option is enabled on a connection using a mechanism similar to any other per-connection option. In Unix systems, this is typically performed using the ‘setsockopt’ system call.

Implementations can also employ system-wide defaults, however systems SHOULD NOT activate this extension by default to avoid interfering with legacy applications.

Due to the potential impacts of legacy middleboxes (discussed in Section 7), a TCP implementation supporting EDO SHOULD log any events within an EDO connection when options that are malformed or show other evidence of tampering arrive. An operating system MAY choose to cache the list of destination endpoints where this has occurred with and block use of EDO on future connections to those endpoints, but this cache MUST be accessible to users/applications on the host. Note that such endpoint assumptions can vary in the presence of load balancers where server implementations vary behind such balancers.

6.2. TCP States and Transitions

TCP EDO does not alter the existing TCP state or state transition mechanisms.
6.3. TCP Segment Processing

TCP EDO alters segment processing during the TCP option processing step. Once detected, the TCP EDO Extension option overrides the TCP Data Offset field for all subsequent option processing. Option processing continues at the next option (if present) after the EDO Extension option.

6.4. Impact on TCP Header Size

The TCP EDO Supported option increases SYN header length by a minimum of 2 bytes, but could increase it by more depending on 32-bit word alignment. Currently popular SYN options total 19 bytes, which leaves more than enough room for the EDO Supported option:

- SACK permitted (2 bytes in SYN, optionally 2 + 8N bytes after) [RFC2018][RFC6675]
- Timestamp (10 bytes) [RFC7323]
- Window scale (3 bytes) [RFC7323]
- MSS option (4 bytes) [RFC793]

Adding the EDO Supported option would result in a total of 21 bytes of SYN option space.

Subsequent segments would use 10 bytes of option space without any SACK blocks (TS only; WS and MSS are used only in SYN and SYN/ACK) or allow up to 3 SACK blocks before needing to use EDO; with EDO, the number of SACK blocks or additional options would be substantially increased. There are also other options that are emerging in the SYN, including TCP Fast Open, which uses another 6-18 (typically 10) bytes in the SYN/ACK of the first connection and in the SYN of subsequent connections [RFC7413].

TCP EDO can also be negotiated in SYNs with either of the following large options:

- TCP-AO (authentication) (16 bytes) [RFC5925]
- Multipath TCP (12 bytes in SYN and SYN/ACK, 20 after) [RFC6824]

Including TCP-AO with TS, WS, SACK increases the SYN option space use to 35 bytes; with Multipath TCP the use is 31 bytes. When Multipath TCP is enabled with the typical options, later segments would require 30 bytes without SACK, thus limiting the SACK option
to one block unless EDO is also supported on at least non-SYN segments.

The full combination of the above options (47 bytes for TS, WS, MSS, SACK, TCP-AO, and MPTCP) does not fit in the existing SYN option space and (as noted) that space cannot be extended within a single SYN segment. There has been a proposal to change TS to a 2 byte "TS permitted" signal in the initial SYN, provided it can be safely enabled during the connection later or might be avoided completely [Ni15]. Even using "TS-permitted", the total space is still too large to support in the initial SYN without SYN option space extension [Bo14][Br14][To15].

The EDO Extension option has negligible impact on other headers, because it can either come first or just after security information, and in either case the additional 4 or 6 bytes are easily accommodated within the TCP Data Offset length. Once the EDO option is processed, the entirety of the remainder of the TCP segment is available for any remaining options.

6.5. Connectionless Resets

A RST may arrive during a currently active connection or may be needed to cleanup old state from an abandoned connection. The latter occurs when a new SYN is sent to an endpoint with matching existing connection state, at which point that endpoint responds with a RST and both ends remove stale information.

The EDO Extension option is mandatory on all TCP segments once negotiated, i.e., except in the SYN and SYN/ACK (which establish support) and the RST. A RST may lack the context to know that EDO is active on a connection.

>> The EDO Extension option MAY occur in a RST when the endpoint has connection state that has negotiated EDO. However, unless the RST is generated by an incoming segment that includes an EDO Extension option, the transmitted RST MUST NOT include the EDO Extension option.

6.6. ICMP Handling

ICMP responses are intended to include the IP and the port fields of TCP and UDP headers of typical TCP/IP and UDP/IP packets [RFC792]. This includes the first 8 data bytes of the original datagram, intended to include the transport port numbers used for connection demultiplexing. Later specifications encourage returning as much of the original payload as possible [RFC1812]. In either case, legacy
options or new options in the EDO extension area might or might not be included, and so options are generally not assumed to be part of ICMP processing anyway.

7. Interactions with Middleboxes

Middleboxes are on-path devices that typically examine or modify packets in ways that Internet routers do not [RFC3234]. This includes parsing transport headers and/or rewriting transport segments in ways that may affect EDO.

There are several cases to consider:

- Typical NAT/NAPT devices, which modify only IP address and/or TCP port number fields (with associated TCP checksum updates)
- Middleboxes that try to reconstitute TCP data streams, such as for deep-packet inspection for virus scanning
- Middleboxes that modify known TCP header fields
- Middleboxes that rewrite TCP segments

7.1. Middlebox Coexistence with EDO

Middleboxes can coexist with EDO when they either support EDO or when they ignore its impact on segment structure.

NATs and NAPTs, which rewrite IP address and/or transport port fields, are the most common form of middlebox and are not affected by the EDO option.

Middleboxes that support EDO would be those that correctly parse the EDO option. Such boxes can reconstitute the TCP data stream correctly or can modify header fields and/or rewrite segments without impact to EDO.

Conventional TCP proxies terminate the TCP connection in both directions and thus operate as TCP endpoints, such as when a client-middlebox and middlebox-server each have separate TCP connections. They would support EDO by following the host requirements herein on both connections. The use of EDO on one connection is independent of its use on the other in this case.
7.2. Middlebox Interference with EDO

Middleboxes that do not support EDO cannot coexist with its use when they modify segment boundaries or do not forward unknown (e.g., the EDO) options.

So-called "transparent" rewriting proxies, which modify TCP segment boundaries, might mix option information with user data if they did not support EDO. Such devices might also interfere with other TCP options such as TCP-AO. There are three types of such boxes:

- Those that process received options and transmit sent options separately, i.e., although they rewrite segments, they behave as TCP endpoints in both directions.

- Those that split segments, taking a received segment and emitting two or more segments with revised headers.

- Those that join segments, receiving multiple segments and emitting a single segment whose data is the concatenation of the components.

In all three cases, EDO is either treated as independent on different sides of such boxes or not. If independent, EDO would either be correctly terminated in either or both directions or disabled due to lack of SYN/ACK confirmation in either or both directions. Problems would occur only when TCP segments with EDO are combined or split while ignoring the EDO option. In the split case, the key concern is if the split happens within the option extension space or if EDO is silently copied to both segments without copying the corresponding extended option space contents. However, the most comprehensive study of these cases indicates that "although middleboxes do split and coalesce segments, none did so while passing unknown options" [Ho11].

Note that the second and third types of middlebox behaviors listed above may create syndromes similar to TCP transmit and receive hardware offload engines that incorrectly modify segments with unknown options.

Middleboxes that silently remove options they do not implement have been observed [Ho11]. Such boxes interfere with the use of the EDO Extension option in the SYN and SYN/ACK segments because extended option space would be misinterpreted as user data if the EDO Extension option were removed, and this cannot be avoided. This is one reason that SYN and SYN/ACK extension requires alternate mechanisms (see Section 8.7). It is also the reason for the 6-byte
EDO Extension variant (see Section 5.3), which can detect such merging or splitting of segments. Further, if such middleboxes become present on a path they could cause similar misinterpretation on segments exchanged in the ESTABLISHED and subsequent states. As a result, this document requires that the EDO Extension option be avoided on the SYN/ACK and that this option needs to be used on all segments once successfully negotiated and encourages use of the 6-byte EDO Extension variant.

Deep-packet inspection systems that inspect TCP segment payloads or attempt to reconstitute the data stream would incorrectly include option data in the reconstituted user data stream, which might interfere with their operation.

>> It can be important to detect misbehavior that could cause EDO space to be misinterpreted as user data. In such cases, EDO SHOULD be used in conjunction with an integrity protection mechanism. This includes the 6-byte EDO Extension variant or stronger mechanisms such as IPsec, TCP-AO, etc. It is useful to note that such protection only helps non-compliant components and enable avoidance (e.g., disabling EDO), but integrity protection alone cannot correct the misinterpretation of EDO space as user data.

This situation is similar to that of ECN and ICMP support in the Internet. In both cases, endpoints have evolved mechanisms for detecting and robustly operating around "black holes". Very similar algorithms are expected to be applicable for EDO.

8. Comparison to Previous Proposals

EDO is the latest in a long line of attempts to increase TCP option space [Al06][Ed08][Ko04][Ra12][Yo11]. The following is a comparison of these approaches to EDO, based partly on a previous summary [Ra12]. This comparison differs from that summary by using a different set of success criteria.

8.1. EDO Criteria

Our criteria for a successful solution are as follows:

- Zero-cost fallback to legacy endpoints.
- Minimal impact on middlebox compatibility.
- No additional side-effects.
Zero-cost fallback requires that upgraded hosts incur no penalty for attempting to use EDO. This disqualifies dual-stack approaches, because the client might have to delay connection establishment to wait for the preferred connection mode to complete. Note that the impact of legacy endpoints that silently reflect unknown options are not considered, as they are already non-compliant with existing TCP requirements [RFC793].

Minimal impact on middlebox compatibility requires that EDO works through simple NAT and NAPT boxes, which modify IP addresses and ports and recompute IPv4 header and TCP segment checksums. Middleboxes that reject unknown options or that process segments in detail without regard for unknown options are not considered; they process segments as if they were an endpoint but do so in ways that are not compliant with existing TCP requirements (e.g., they should have rejected the initial SYN because of its unknown options rather than silently relaying it).

EDO also attempts to avoid creating side-effects, such as might happen if options were split across multiple TCP segments (which could arrive out of order or be lost) or across different TCP connections (which could fail to share fate through firewalls or NAT/NAPTs).

These requirements are similar to those noted in [Ra12], but EDO groups cases of segment modification beyond address and port - such as rewriting, segment drop, sequence number modification, and option stripping - as already in violation of existing TCP requirements regarding unknown options, and so we do not consider their impact on this new option.

8.2. Summary of Approaches

There are three basic ways in which TCP option space extension has been attempted:

1. Use of a TCP option.
2. Redefinition of the existing TCP header fields.
3. Use of option space in multiple TCP segments (split across multiple segments).

A TCP option is the most direct way to extend the option space and is the basis of EDO. This approach cannot extend the option space of the initial SYN.
Redefining existing TCP header fields can be used to either contain additional options or as a pointer indicating alternate ways to interpret the segment payload. All such redefinitions make it difficult to achieve zero-impact backward compatibility, both with legacy endpoints and middleboxes.

Splitting option space across separate segments can create unintended side-effects, such as increased delay to deal with path latency or loss differences.

The following discusses three of the most notable past attempts to extend the TCP option space: Extended Segments, TCPx2, LO/SLO, and LOIC. [Ra12] suggests a few other approaches, including use of TCP option cookies, reuse/overload of other TCP fields (e.g., the URG pointer), or compressing TCP options. None of these is compatible with legacy endpoints or middleboxes.

8.3. Extended Segments

TCP Extended Segments redefined the meaning of currently unused values of the Data Offset (DO) field [Ko04]. TCP defines DO as indicating the length of the TCP header, including options, in 32-bit words. The default TCP header with no options is 5 such words, so the minimum currently valid DO value is 5 (meaning 40 bytes of option space). This document defines interpretations of values 0-4: DO=0 means 48 bytes of option space, DO=1 means 64, DO=2 means 128, DO=3 means 256, and DO=4 means unlimited (e.g., the entire payload is option space). This variant negotiates the use of this capability by using one of these invalid DO values in the initial SYN.

Use of this variant is not backward-compatible with legacy TCP implementations, whether at the desired endpoint or on middleboxes. The variant also defines a way to initiate the feature on the passive side, e.g., using an invalid DO during the SYN/ACK when the initial SYN had a valid DO. This capability allows either side to initiate use of the feature but is also not backward compatible.

8.4. TCPx2

TCPx2 redefines legacy TCP headers by basically doubling all TCP header fields [Al06]. It relies on a new transport protocol number to indicate its use, defeating backward compatibility with all existing TCP capabilities, including firewalls, NATs/NAFTs, and legacy endpoints and applications.
8.5. LO/SLO

The TCP Long Option (LO, [Ed08]) is very similar to EDO, except that presence of LO results in ignoring the existing Data Offset (DO) field and that LO is required to be the first option. EDO considers the need for other fields to be first and declares that the EDO is the last option as indicated by the DO field value. Like LO, EDO is required in every segment once negotiated.

The TCP Long Option draft also specified the SYN Long Option (SLO) [Ed08]. If SLO is used in the initial SYN and successfully negotiated, it is used in each subsequent segment until all of the initial SYN options are transmitted.

LO is backward compatible, as is SLO; in both cases, endpoints not supporting the option would not respond with the option, and in both cases the initial SYN is not itself extended.

SLO does modify the three-way handshake because the connection isn’t considered completely established until the first data byte is acknowledged. Legacy TCP can establish a connection even in the absence of data. SLO also changes the semantics of the SYN/ACK; for legacy TCP, this completes the active side connection establishment, where in SLO an additional data ACK is required. A connection whose initial SYN options have been confirmed in the SYN/ACK might still fail upon receipt of additional options sent in later SLO segments. This case – of late negotiation fail – is not addressed in the specification.

8.6. LOIC

TCP Long Options by Invalid Checksum is a dual-stack approach that uses two initial SYNS to initiate all updated connections [Yo11]. One SYN negotiates the new option and the other SYN payload contains only the entire options. The negotiation SYN is compliant with existing procedures, but the option SYN has a deliberately incorrect TCP checksum (decremented by 2). A legacy endpoint would discard the segment with the incorrect checksum and respond to the negotiation SYN without the LO option.

Use of the option SYN and its incorrect checksum both interfere with other legacy components. Segments with incorrect checksums will be silently dropped by most middleboxes, including NATs/NAPT. Use of two SYNs creates side-effects that can delay connections to upgraded endpoints, notably when the option SYN is lost or the SYNs arrive out of order. Finally, by not allowing other options in the negotiation SYN, all connections to legacy endpoints either use no
options or require a separate connection attempt (either concurrent or subsequent).

8.7. Problems with Extending the Initial SYN

The key difficulty with most previous proposals is the desire to extend the option space in all TCP segments, including the initial SYN, i.e., SYN with no ACK, typically the first segment of a connection, as well as possibly the SYN/ACK. It has proven difficult to extend space within the segment of the initial SYN in the absence of prior negotiation while maintaining current TCP three-way handshake properties, and it may be similarly challenging to extend the SYN/ACK (depending on asymmetric middlebox assumptions).

A new TCP option cannot extend the Data Offset of a single TCP initial SYN segment, and cannot extend a SYN/ACK in a single segment when considering misbehaving middleboxes. All TCP segments, including the initial SYN and SYN/ACK, may include user data in the payload data [RFC793], and this can be useful for some proposed features such as TCP Fast Open [RFC7413]. Legacy endpoints that ignore the new option would process the payload contents as user data and send an ACK. Once ACK’ed, this data cannot be removed from the user stream.

The Reserved TCP header bits cannot be redefined easily, even though three of the six total bits have already been redefined (ECE/CWR [RFC3168] and NS [RFC3540]). Legacy endpoints have been known to reflect received values in these fields; this was safely dealt with for ECN but would be difficult here [RFC3168].

TCP initial SYN (SYN and not ACK) segments can use every other TCP header field except the Acknowledgement number, which is not used because the ACK field is not set. In all other segments, all fields except the three remaining Reserved header bits are actively used. The total amount of available header fields, in either case, is insufficient to be useful in extending the option space.

The representation of TCP options can be optimized to minimize the space needed. In such cases, multiple Kind and Length fields are combined, so that a new Kind would indicate a specific combination of options, whose order is fixed and whose length is indicated by one Length field. Most TCP options use fields whose size is much larger than the required Kind and Length components, so the resulting efficiency is typically insufficient for additional options.
The option space of an initial SYN segment might be extended by using multiple initial segments (e.g., multiple SYNs or a SYN and non-SYN) or based on the context of previous or parallel connections. This method may also be needed to extend space in the SYN/ACK in the presence of misbehaving middleboxes. Because of their potential complexity, these approaches are addressed in separate documents [Bo14][Br14][To15].

Option space cannot be extended in outer layer headers, e.g., IPv4 or IPv6. These layers typically try to avoid extensions altogether, to simplify forwarding processing at routers. Introducing new shim layers to accommodate additional option space would interfere with deep-packet inspection mechanisms that are in widespread use.

As a result, EDO does not attempt to extend the space available for options in TCP initial SYNs. It does extend that space in all other segments (including SYN/ACK), which has always been trivially possible once an option is defined.

9. Implementation Issues

TCP segment processing can involve accessing nonlinear data structures, such as chains of buffers. Such chains are often designed so that the maximum default TCP header (60 bytes) fits in the first buffer. Extending the TCP header across multiple buffers may necessitate buffer traversal functions that span boundaries between buffers. Such traversal can also have a significant performance impact, which is additional rationale for using TCP option space - even extended option space - sparingly.

Although EDO can be large enough to consume the entire segment, it is important to leave space for data so that the TCP connection can make forward progress. It would be wise to limit EDO to consuming no more than MSS-4 bytes of the IP segment, preferably even less (e.g., MSS-128 bytes).

When using the ExID variant for testing and experimentation, either TCP option codepoint (253, 254) is valid in sent or received segments.

Implementers need to be careful about the potential for offload support interfering with this option. The EDO data needs to be passed to the protocol stack as part of the option space, not integrated with the user segment, to allow the offload to independently determine user data segment boundaries and combine them correctly with the extended option data. Some legacy hardware receive offload engines may present challenges in this regard, and
may be incompatible with EDO where they incorrectly process segments with unknown options. Such offload engines should be considered part of the protocol stack and updated accordingly. Issues with incorrect resegmentation by an offload engine can be detected in the same way as middlebox tampering.

10. Security Considerations

It is meaningless to have the Data Offset further exceed the position of the EDO data offset option.

>> When the EDO Extension option is present, the EDO Extension option SHOULD be the last non-null option covered by the TCP Data Offset, because it would be the last option affected by Data Offset.

This also makes it more difficult to use the Data Offset field as a covert channel.

11. IANA Considerations

We request that, upon publication, this option be assigned a TCP Option codepoint by IANA, which the RFC Editor will replace EDO-OPT in this document with codepoint value.

The TCP Experimental ID (ExID) with a 16-bit value of 0x0E0D0 (in network standard byte order) has been assigned for use during testing and preliminary experiments.

12. References

12.1. Normative References


12.2. Informative References


[Ra12]  Ramaiah, A., "TCP option space extension", draft-ananth-tcpm-tcpopstext-00 (work in progress), March 2012.


13. Acknowledgments

The authors would like to thank the IETF TCPM WG for their feedback, in particular: Oliver Bonaventure, Bob Briscoe, Ted Faber, John Leslie, Pasi Sarolahti, Richard Scheffenegger, and Alexander Zimmerman.

This document was prepared using 2-Word-v2.0.template.dot.

Authors’ Addresses

Joe Touch
USC/ISI
4676 Admiralty Way
Marina del Rey, CA 90292-6695 USA

Phone: +1 (310) 448-9151
Email: touch@isi.edu
TCP Alternative Backoff with ECN (ABE)
draft-khademi-alternativebackoff-ecn-01

Abstract

This memo provides an experimental update to RFC3168. It updates the TCP sender-side reaction to a congestion notification received via Explicit Congestion Notification (ECN). ECN-marking can allow a network device to signal congestion at a point before a transport experiences congestion loss or additional queuing delay. The updated method is less conservative than the TCP reaction in response to loss. The intention is to achieve good throughput when the queue at the bottleneck is smaller than the bandwidth-delay-product of the connection. This is more likely when an Active Queue Management (AQM) mechanism has used ECN to CE-mark a packet, than when a packet was lost. Future versions of this document will discuss SCTP as well as other transports using ECN.

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 http://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 March 24, 2016.
1. Introduction

Explicit Congestion Notification (ECN) is specified in [RFC3168]. It allows a network device that uses Active Queue Management (AQM) to set the congestion experienced, CE, codepoint in the ECN field of the IP packet header, rather than to drop ECN-capable packets when incipient congestion is detected. When an ECN-capable transport is used over a path that supports ECN, it provides the opportunity for flows to improve their performance in the presence of incipient congestion [I-D.AQM-ECN-benefits].

[RFC3168] not only specifies the router use of the ECN field, it also specifies a TCP procedure for using ECN. This states that a TCP sender should treat the ECN indication of congestion in the same way as that of a non-ECN-Capable TCP flow experiencing loss, by halving
the congestion window "cwnd" and by reducing the slow start threshold "ssthresh". [RFC5681] stipulates that TCP congestion control sets "ssthresh" to max(FlightSize / 2, 2*SMSS) in response to packet loss. Consequently, a non-ECN enabled standard TCP flow using this reaction needs significant network queue space: it can only fully utilize a bottleneck when the length of the link queue (or the AQM dropping threshold) is at least the bandwidth-delay product (BDP) of the flow.

A backoff multiplier of 0.5 (halving cwnd and ssthresh after packet loss) is not the only available strategy. As defined in [ID.CUBIC], CUBIC multiplies the current cwnd by 0.8 in response to loss (although the Linux implementation of CUBIC has used a multiplier of 0.7 since kernel version 2.6.25 released in 2008). Consequently, CUBIC flows more fully utilize paths even when the bottleneck queue is slightly shorter than the bandwidth-delay product of the flow. However, in the case of a DropTail (FIFO) queue without AQM, such less-aggressive backoff increases the risk of creating a standing queue [CODEL2012].

Devices implementing AQM are likely to be the dominant (and possibly only) source of ECN CE-marking for packets from ECN-capable senders. AQM mechanisms typically strive to maintain a small queue length, regardless of the bandwidth-delay product of flows passing through them. Receipt of an ECN CE-mark might therefore reasonably be taken to indicate that a small bottleneck queue exists in the path, and hence the TCP flow would benefit from using a less aggressive backoff multiplier.

Results reported in [ABE2015] show significant benefits (improved throughput, resulting in reduced completion times for short flows) when reacting to ECN-Echo by multiplying cwnd and ssthresh with a value in the range [0.7..0.85]. Section 2 describes the rationale for this change. Section 3 specifies a change to the TCP sender backoff behaviour in response to an indication that CE-marks have been received by the receiver.

2. Discussion

Much of the background to this proposal can be found in [ABE2015]. Using a mix of experiments, theory and simulations with standard NewReno and CUBIC, [ABE2015] recommends enabling ECN and "...letting individual TCP senders use a larger multiplicative decrease factor in reaction to ECN CE-marks from AQM-enabled bottlenecks." Such a change is noted to result in "...significant performance gains in lightly-multiplexed scenarios, without losing the delay-reduction benefits of deploying CoDel or PIE."
2.1. Why use ECN to vary the degree of backoff?

The classic rule-of-thumb dictates a BDP of bottleneck buffering if a TCP connection wishes to optimise path utilisation. A single TCP connection running through such a bottleneck will have opened cwnd up to 2*BDP by the time packet loss occurs. [RFC5681]’s halving of cwnd and ssthresh pushes the TCP connection back to allowing only a BDP of packets in flight -- just enough to maintain 100% utilisation of the network path.

AQM schemes like CoDel and PIE use congestion notifications to constrain the queuing delays experienced by packets, rather than in response to impending or actual bottleneck buffer exhaustion. With current default delay targets, CoDel and PIE both effectively emulate a shallow buffered bottleneck (section II, [ABE2015]). This interacts acceptably for TCP connections over low BDP paths, or highly multiplexed scenarios (many concurrent TCP connections). However, it interacts badly with lightly-multiplexed cases (few concurrent connections) over high BDP paths. Conventional TCP backoff in such cases leads to gaps in packet transmission and underutilisation of the path.

In an ideal world, the TCP sender would adapt its backoff strategy to match the effective depth at which a bottleneck begins indicating congestion. In the practical world, [ABE2015] proposes using the existence of ECN CE-marks to infer whether a path’s bottleneck is AQM-enabled (shallow queue) or classic DropTail (deep queue), and adjust backoff accordingly. This results in a change to the requirements of [RFC3168], which required TCP senders to respond the same following indication of a received ECN CE-mark and a packet loss, making these equivalent signals of congestion. (The idea to change this behaviour pre-dates ABE. [ICC2002] also proposed using ECN CE-marks to modify TCP congestion control behaviour, using a larger multiplicative decrease factor in conjunction with a smaller additive increase factor to deal with RED-based bottlenecks that were not necessarily configured to emulate a shallow queue.)

[RFC7567] states that "deployed AQM algorithms SHOULD support Explicit Congestion Notification (ECN) as well as loss to signal congestion to endpoints" and [I-D.AQM-ECN-benefits] encourages this deployment. Apple recently announced their intention to enable ECN in iOS 9 and OS X 10.11 devices [WWDC2015]. By 2014, server-side ECN negotiation was observed to be provided by the majority of the top million web servers [PAM2015], and only 0.5% of websites incurred additional connection setup latency using RFC3168-compliant ECN-fallback mechanisms.
2.2. Choice of ABE multiplier

ABE decouples a TCP sender’s reaction to loss and ECN CE-marks. The description respectively uses beta_{loss} and beta_{ecn} to refer to the multiplicative decrease factors applied in response to packet loss and in response to an indication of a received CN CE-mark on an ECN-enabled TCP connection (based on the terms used in [ABE2015]). For non-ECN-enabled TCP connections, no ECN CE-marks are received and only beta_{loss} applies.

In other words, in response to detected loss:
\[ cwnd_{(n+1)} = cwnd_n \times beta_{(loss)} \]
and in response to an indication of a received ECN CE-mark:
\[ cwnd_{(n+1)} = cwnd_n \times beta_{(ecn)} \]

The higher the values of beta_{*}, the less aggressive the response of any individual backoff event.

The appropriate choice for beta_{loss} and beta_{ecn} values is a balancing act between path utilisation and draining the bottleneck queue. More aggressive backoff (smaller beta_{*}) risks underutilising the path, while less aggressive backoff (larger beta_{*}) can result in slower draining of the bottleneck queue.

The Internet is already running with at least two different beta_{loss} values, [RFC5681]’s 0.5, and Linux CUBIC’s 0.7. ABE proposes no change to beta_{loss} used by any current TCP implementations.

beta_{ecn} depends on how we want to optimise the response of a TCP connection to shallow AQM marking thresholds. beta_{loss} reflects the preferred response of each TCP algorithm when faced with exhaustion of buffers (of unknown depth) signalled by packet loss. Consequently, for any given TCP algorithm the choice of beta_{ecn} is likely to be algorithm-specific, rather than a constant multiple of the algorithm’s existing beta_{loss}.

A range of experiments (section IV, [ABE2015]) with NewReno and CUBIC over CoDel and PIE in lightly multiplexed scenarios have explored this choice of parameter. These experiments indicate that CUBIC connections benefit from beta_{ecn} of 0.85 (cf. beta_{loss} = 0.7), and NewReno connections see improvements with beta_{ecn} in the range 0.7 to 0.85 (c.f., beta_{loss} = 0.5).
3. Updating the Sender-side ECN Reaction

This section specifies an experimental update to [RFC3168].

3.1. RFC 2119

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 [RFC2119].

3.2. Update to RFC 3168

This document specifies an update to the TCP sender reaction that follows when the TCP receiver signals that ECN CE-marked packets have been received.

The first paragraph of Section 6.1.2, "The TCP Sender", in [RFC3168] contains the following text:

"If the sender receives an ECN-Echo (ECE) ACK packet (that is, an ACK packet with the ECN-Echo flag set in the TCP header), then the sender knows that congestion was encountered in the network on the path from the sender to the receiver. The indication of congestion should be treated just as a congestion loss in non-ECN-Capable TCP. That is, the TCP source halves the congestion window "cwnd" and reduces the slow start threshold "ssthresh"."

This memo updates this by replacing this with the following text:

"If the sender receives an ECN-Echo (ECE) ACK packet (that is, an ACK packet with the ECN-Echo flag set in the TCP header), then the sender knows that congestion was encountered in the network on the path from the sender to the receiver. The indication of congestion SHOULD induce a less conservative reaction than loss: the TCP source multiplies the congestion window ‘cwnd’ with 0.8 and reduces the slow start threshold ‘ssthresh’."  

3.3. Status of the Update

XXX Author’s note: Once ICCRG evaluation has been completed an appropriate outcome may be inserted here XXX

The congestion control behaviour specified in this update will be evaluated by the IRTF Internet Congestion Control Research Group (ICCRG), to determine whether it is thought safe for deployment in the general Internet.
XXX Author’s note: If this is adopted for publication as an Experimental RFC we need to explain why this is not PS XXX

The present specification has been assigned an Experimental status, because this is common practice for first introduction of changes to the TCP protocol specification, where deployment experience is usually required prior to publishing a Standards-Track document.

This update is a sender-side only change. Like other changes to congestion-control algorithms it does not require any change to the TCP receiver or to network devices (except to enable an ECN-marking algorithm [RFC3168] [RFC7567]). If the method is only deployed by some TCP senders, and not by others, the senders that use this method can gain advantage, possibly at the expense of other flows that do not use this updated method. This advantage applies only to ECN-marked packets and not to loss indications. Hence, the new method can not lead to congestion collapse.

4. Acknowledgements

Authors N. Khademi, M. Welzl and G. Fairhurst were part-funded by the European Community under its Seventh Framework Programme through the Reducing Internet Transport Latency (RITE) project (ICT-317700). The views expressed are solely those of the authors.

The authors would like to thank the following people for their contributions to [ABE2015]: Chamil Kulatunga, David Ros, Stein Gjessing, Sebastian Zander.

5. IANA Considerations

XX RFC ED - PLEASE REMOVE THIS SECTION XXX

This memo includes no request to IANA.

6. Security Considerations

The described method is a sender-side only transport change, and does not change the protocol messages exchanged. The security considerations of RFC 3819 therefore still apply.

This document describes a change to TCP congestion control that can make TCP senders more aggressive than flows using TCP as specified in RFC 3819. This could lead to a change in the capacity achieved by flows sharing a network bottleneck. If some flows use this method and share capacity with other flows using previous methods this could reduce fairness in the capacity allocation. Similar unfairness is also exhibited by other congestion control mechanisms that have been
in use in the Internet for many years (e.g., CUBIC [ID.CUBIC]). Unfairness may also be a result of other factors, including the round trip time experienced by a flow. This advantage applies only to ECN-marked packets and not to loss indications, and will therefore can not lead to congestion collapse.

7. References

7.1. Normative References


7.2. Informative References


Internet-Draft                     ABE                    September 2015


Authors’ Addresses

Naeem Khademi
University of Oslo
PO Box 1080 Blindern
Oslo N-0316
Norway

Email: naeemk@ifi.uio.no

Michael Welzl
University of Oslo
PO Box 1080 Blindern
Oslo N-0316
Norway

Email: michawe@ifi.uio.no
Grenville Armitage
Centre for Advanced Internet Architectures
Swinburne University of Technology
PO Box 218
John Street, Hawthorn
Victoria 3122
Australia
Email: garmitage@swin.edu.au

Godred Fairhurst
University of Aberdeen
School of Engineering, Fraser Noble Building
Aberdeen AB24 3UE
UK
Email: gorry@erg.abdn.ac.uk
Abstract

This document describes an idea "Sharp Close" that eliminates or minimizes TIME-WAIT state of TCP connections.

In the current TCP specification ([RFC0793]), there are some inappropriate or not up-to-date functions. Here we focus and discuss on TCP TIME-WAIT state function.

TIME-WAIT is the last state of TCP connections of Active Close side nodes. After TCP connections are effectively closed, state of them move to TIME-WAIT state. After TIME-WAIT state is finished, resources of connections are released. This means that even if connections are effectively finished, resources of connections are NOT released. The TIME-WAIT state prevents from releasing them.

From the viewpoints of current high-speed and high-multiplicity communication styles, it is thought that TIME-WAIT state is one of evil functions.

In order to provide efficient communications that match current styles, an idea "Sharp Close" that eliminates or minimizes TIME-WAIT state of TCP connections is proposed.
1. Introduction

This document describes an idea "Sharp Close" that eliminates or minimizes TIME-WAIT state of TCP connections.

In the current TCP specification ([RFC0793]), there are some inappropriate or not up-to-date functions. Here we focus and discuss on TCP TIME-WAIT state function.

TIME-WAIT is the last state of TCP connections of Active Close side nodes. After TCP connections are effectively closed, state of them move to TIME-WAIT state. [RFC0793] defines that the connections stay there 2MSL(Maximum Segment Lifetime) seconds. (2MSL = 240 sec.)
After TIME-WAIT state is finished, resources of connections are released. This means that even if connections are effectively finished, resources of connections are NOT released. The TIME-WAIT state prevents from releasing them.

From the viewpoints of current high-speed and high-multiplicity communication styles that require highly resource recycling, it is thought that TIME-WAIT state is one of evil functions.

In order to provide efficient communications that match current styles, an idea "Sharp Close" that eliminates or minimizes TIME-WAIT state of TCP connections is proposed.

In the following sections, analysis of current TIME-WAIT state and design of "Sharp Close" etc. are described.

2. Analysis of current TIME-WAIT state

![Current ACTIVE-PASSIVE Close Sequence](image)

Fig. 1 Current ACTIVE-PASSIVE Close Sequence
Fig. 1 and Fig. 2 show Close Sequence that is defined by current specification [RFC0793]. TCP connections on ACTIVE CLOSE node (that initiates sending FIN) side reach TIME-WAIT as a last state. They stay there 2MSL seconds.

Table 1  Actual 2MSL values used by major OS implementation.

<table>
<thead>
<tr>
<th>RFC/OS</th>
<th>2MSL value</th>
</tr>
</thead>
<tbody>
<tr>
<td>[RFC0793]</td>
<td>240 sec.</td>
</tr>
<tr>
<td>Windows2000</td>
<td>240 sec.</td>
</tr>
<tr>
<td>Windows (after Win2K)</td>
<td>120 sec.</td>
</tr>
<tr>
<td>Unix/Linux</td>
<td>60 sec.</td>
</tr>
</tbody>
</table>
Table 1 shows actual 2MSL values that are surveyed by authors. [RFC0793] says "For this specification the MSL is taken to be 2 minutes."

Since 240 sec. ([RFC0793]) is long time, recent major OSes adopt rather shorter time.

However, from the viewpoints of current communication styles that require highly resource recycling, TIME-WAIT time is still too long.

Now, it is almost thought that staying at TIME-WAIT state is waste of time.

3. Why TIME-WAIT state is needed?

Basically, TIME-WAIT state is designed for fail-safe! purpose.

If it is assumed that packets transferring order is not changed, all of !data! packets from a corresponding node are received when FIN-WAIT-2 state is finished (responding FIN packet is received) and no !data! packets will not be received after that.

At TIME-WAIT state, an ACTIVE CLOSE node waits for a 'resending' !control! packet FIN only from the corresponding node for the case of the sent ACK (for the FIN) is lost. (No !data! packets are waited for.)

Only when the last sent ACK from the ACTIVE CLOSE node is lost, 'resending' control packet FIN from the corresponding node is issued.

It is rare case to happen this event at current stable network environment.

Since all data from the corresponding node is received by the ACTIVE CLOSE node, it is less significant issue to wait for 'resending' FIN packet.

If 'resending' FIN is NOT waited at ACTIVE CLOSE node and 'resending' FIN is issued from the corresponding node, significant problem will NOT be happened, only RST packet (to notify receiving unexpected packet) will be issued from the ACTIVE CLOSE node.
4. Design of "Sharp Close" (elimination of TIME-WAIT state)

![Diagram of Sharp ACTIVE-PASSIVE Close Sequence](image)

Fig. 3 (Proposed) Sharp ACTIVE-PASSIVE Close Sequence

![Diagram of Sharp ACTIVE-ACTIVE Close Sequence](image)

Fig. 4 (Proposed) Sharp ACTIVE-ACTIVE Close Sequence
It is easy to design "Sharp Close" function. "Sharp Close" function is achieved by eliminating or minimizing TIME-WAIT state of TCP connections.

Fig. 3 and Fig. 4. show Close Sequence that is defined by "Sharp Close" function.

5. Eliminate TIME-WAIT state by setsockopt()

Under current implementation, TIME-WAIT (close()) action can be controlled by setsockopt() function.

SO_LINGER option of setsockopt() can eliminate TIME-WAIT state and close connections immediately.

Concrete procedures how to eliminate TIME-WAIT:

Fig. 5 shows struct socket in <sys/socket.h>

```c
struct linger {
    int l_onoff;    /* linger active */
    int l_linger;   /* how many seconds to linger for */
};
```

By using the following shown procedures, TIME-WAIT state is eliminated and connections are closed immediately.

1: makes linger active(on)
   l_onoff = on;

2: sets linger time to 0
   l_linger = 0 ;

It is possible to eliminate TIME-WAIT state by these procedures. However, this behavior is "NOT default" operation. In order to utilize this feature, it is necessary to modify huge number of communication applications.

Furthermore, this feature is not implemented on every existing OSes and it is not always possible to eliminate TIME-WAIT state on every OSes.
6. Security Considerations

Goals of the proposed idea ("Sharp Close") are to eliminate or minimize TIME-WAIT state by default on OS kernel level. From functional viewpoints, the same concept to eliminate TIME-WAIT state is already implemented by using LINGER option of setsockopt() function. It is not default operation, however it has already implemented and worked.

So, there are no new Security Consideration issues that should be discussed here.

7. IANA Considerations

This document does not require any resource assignments to IANA.

Acknowledgment

A part of this work is supported by the program: SCOPE (Strategic Information and Communications R&D Promotion Programme) operated by Ministry of Internal Affairs and Communications of JAPAN.

Appendix A. Implementations

Currently, above described "Sharp Close" functions have been implemented and verified under the following OS.

Ubuntu 13.04 (kernel 3.8.13.8)

References

Normative References


Informative References

Authors’ Addresses

Hiroshi Kitamura
Cyber Security Strategy Division / Cloud System Research Laboratories,
NEC Corporation
7-1, Shiba 5-chome, Minato-ku, Tokyo 108-8001, JAPAN
Phone: +81 3 3798 0563
Email: kitamura@da.jp.nec.com

Shingo Ata
Graduate School of Engineering, Osaka City University
3-3-138, Sugimoto, Sumiyoshi-Ku, Osaka 558-8585, JAPAN
Phone: +81 6 6605 2191
Fax: +81 6 6605 2191
Email: ata@info.eng.osaka-cu.ac.jp

Masayuki Murata
Graduate School of Information Science and Technology, Osaka Univ.
1-5 Yamadaoka, Suita, Osaka 565-0871, JAPAN
Phone: +81 6 6879 4542
Fax: +81 6 6879 4544
Email: murata@ist.osaka-u.ac.jp
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 is specified for TCP in such a way that only one feedback signal can be transmitted per Round-Trip Time (RTT). Recently, new TCP mechanisms like Congestion Exposure (ConEx) or Data Center TCP (DCTCP) need more accurate ECN feedback information whenever more than one marking is received in one RTT. This document specifies an experimental scheme to provide more than one feedback signal per RTT in the TCP header. Given TCP header space is scarce, it overloads the three existing ECN-related flags in the TCP header and provides additional information in a new TCP option.
Internet-Draft          Accurate TCP-ECN Feedback           October 2015

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the
document authors.  All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal
Provisions Relating to IETF Documents
(http://trustee.ietf.org/license-info) in effect on the date of
publication of this document.  Please review these documents
carefully, as they describe your rights and restrictions with respect
to this document.  Code Components extracted from this document must
include Simplified BSD License text as described in Section 4.e of
the Trust Legal Provisions and are provided without warranty as
described in the Simplified BSD License.

Table of Contents

1.  Introduction ................................................. 3
  1.1.  Document Roadmap ....................................... 4
  1.2.  Goals .................................................. 4
  1.3.  Experiment Goals ....................................... 5
  1.4.  Terminology ............................................ 5
  1.5.  Recap of Existing ECN feedback in IP/TCP .............. 6

2.  AccECN Protocol Overview and Rationale ....................... 7
  2.1.  Capability Negotiation .................................. 8
  2.2.  Feedback Mechanism ...................................... 8
  2.3.  Delayed ACKs and Resilience Against ACK Loss .......... 9
  2.4.  Feedback Metrics ........................................ 10
  2.5.  Generic (Dumb) Reflector ............................... 10

3.  AccECN Protocol Specification ............................... 11
  3.1.  Negotiation during the TCP handshake ................... 11
  3.2.  AccECN Feedback ......................................... 14
    3.2.1.  The ACE Field ..................................... 14
    3.2.2.  Safety against Ambiguity of the ACE Field ........ 16
    3.2.3.  The AccECN Option .................................. 16
    3.2.4.  Path Traversal of the AccECN Option ................ 17
    3.2.5.  Usage of the AccECN TCP Option ...................... 19
  3.3.  AccECN Compliance by TCP Proxies, Offload Engines and
        other Middleboxes ...................................... 20

4.  Interaction with Other TCP Variants ........................ 21
  4.1.  Compatibility with SYN Cookies ........................ 21
  4.2.  Compatibility with Other TCP Options and Experiments .. 21
  4.3.  Compatibility with Feedback Integrity Mechanisms ....... 21

5.  Protocol Properties .......................................... 23

6.  IANA Considerations ......................................... 25

7.  Security Considerations ..................................... 25

8.  Acknowledgements ............................................ 26

9.  Comments Solicited .......................................... 26
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. ECN is specified for TCP in such a way that only one feedback signal can be transmitted per Round-Trip Time (RTT). Recently, proposed mechanisms like Congestion Exposure (ConEx [I-D.ietf-conex-abstract-mech]) or DCTCP [I-D.bensley-tcpm-dctcp] need more accurate ECN feedback information whenever more than one marking is received in one RTT. A fuller treatment of the motivation for this specification is given in the associated requirements document [RFC7560].

This document specifies an experimental 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. If AccECN progresses from experimental to the standards track, it is intended to be a complete replacement for classic ECN feedback, not a fork in the design of TCP. Thus, the applicability of AccECN is intended to include all public and private IP networks (and even any non-IP networks over which TCP is used today). Until the AccECN experiment succeeds, [RFC3168] will remain as the standards track specification for adding ECN to TCP. To avoid confusion, in this document we use the term 'classic ECN' for the pre-existing ECN specification [RFC3168].
AccECN is solely an (experimental) change to the TCP wire protocol. It is completely independent of how TCP might respond to congestion feedback. This specification overloads flags and fields in the main TCP header with new definitions, so both ends have to support the new wire protocol before it can be used. Therefore during the TCP handshake the two ends use the three ECN-related flags in the TCP header to negotiate the most advanced feedback protocol that they can both support.

It is likely (but not required) that the AccECN protocol will be implemented along with the following experimental additions to the TCP-ECN protocol: ECN-capable SYN/ACK [RFC5562], ECN path-probing and fall-back [I-D.kuehlewind-tcpm-ecn-fallback] and testing receiver non-compliance [I-D.moncaster-tcpm-rcv-cheat].

1.1. Document Roadmap

The following introductory sections outline the goals of AccECN (Section 1.2) and the goal of experiments with ECN (Section 1.3) so that it is clear what success would look like. Then terminology is defined (Section 1.4) and a recap of existing prerequisite technology is given (Section 1.5).

Section 2 gives an informative overview of the AccECN protocol. Then Section 3 gives the normative protocol specification. Section 4 assesses the interaction of AccECN with commonly used variants of TCP, whether standardised or not. Section 5 summarises the features and properties of AccECN.

Section 6 summarises the protocol fields and numbers that IANA will need to assign and Section 7 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.

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 recognises that a perfect scheme that fully satisfies all the requirements is unlikely and trade-offs between requirements are likely. Section 5 presents the properties of AccECN against these requirements and discusses the trade-offs made.
The requirements document recognises 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. Experiment Goals

TCP is critical to the robust functioning of the Internet, therefore any proposed modifications to TCP need to be thoroughly tested. The present specification describes an experimental protocol that adds more accurate ECN feedback to the TCP protocol. The intention is to specify the protocol sufficiently so that more than one implementation can be built in order to test its function, robustness and interoperability (with itself and with previous version of ECN and TCP).

The experimental protocol will be considered successful if it satisfies the requirements of [RFC7560] in the consensus opinion of the IETF tcpm working group. In short, this requires that it improves the accuracy and timeliness of TCP’s ECN feedback, as claimed in Section 5, while striking a balance between the conflicting requirements of resilience, integrity and minimisation of overhead. It also requires that it is not unduly complex, and that it is compatible with prevalent equipment behaviours in the current Internet, whether or not they comply with standards.

1.4. 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.

Pure ACK: A TCP acknowledgement without a data payload.

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 RFC 2119 [RFC2119].

1.5. 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.

<table>
<thead>
<tr>
<th>IP-ECN codepoint (binary)</th>
<th>Codepoint name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>Not-ECT</td>
<td>Not ECN-Capable Transport</td>
</tr>
<tr>
<td>01</td>
<td>ECT(1)</td>
<td>ECN-Capable Transport (1)</td>
</tr>
<tr>
<td>10</td>
<td>ECT(0)</td>
<td>ECN-Capable Transport (0)</td>
</tr>
<tr>
<td>11</td>
<td>CE</td>
<td>Congestion Experienced</td>
</tr>
</tbody>
</table>

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 ECN Nonce [RFC3540] is an optional experimental addition to ECN that the TCP sender can use to protect against accidental or malicious concealment of marked or dropped packets. The sender can send an ECN nonce, which is a continuous pseudo-random pattern of ECT(0) and ECT(1) codepoints in the ECN field. The receiver is required to feed back a 1-bit nonce sum that counts the occurrence of ECT(1) packets using the last bit of byte 13 in the TCP header, which is defined as the Nonce Sum (NS) flag.

![TCP header flags diagram](image)

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:

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

- 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 (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:

- this efficiently reuses scarce TCP header space, given TCP option space is approaching saturation;
- a single upgrade path for the TCP protocol is preferable to a fork in the design;
- 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;
- 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 and successful negotiation using the flags in the main header is taken as sufficient evidence that both ends also support the AccECN Option. 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 initialised at the start of the half-connection. Three count the number of arriving payload bytes marked CE, ECT(1) and ECT(0) respectively. 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. The LSBs of each of the three byte counters 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. Then, 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 required not 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. Because the 3-bit ACE field is so small, when it is the only field available the Data Sender has to interpret it conservatively assuming the worst possible wrap.

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 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 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 information about CE markings on both data and control packets. According to [RFC3168] the Data Sender is meant to set 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 [I-D.kuehlewind-tcpm-ecn-fallback] and [I-D.moncaster-tcpm-rcv-cheat]).

The initial SYN is the most critical control packet, so AccECN provides feedback on whether it is CE marked, even though it is not allowed to be ECN-capable according to RFC 3168. However, middleboxes have been known to overwrite the ECN IP field as if it is still part of the old Type of Service (ToS) field. If a TCP client has set the SYN to Not-ECT, but receives CE feedback, it can detect such middlebox interference and send Not-ECT for the rest of the connection (see [I-D.kuehlewind-tcpm-ecn-fallback] for the detailed fall-back behaviour).
Today, if a TCP server receives CE on a SYN, it cannot know whether it is invalid (or valid) because only the TCP client knows whether it originally marked the SYN as Not-ECT (or ECT). Therefore, the server’s only safe course of action is to disable ECN for the connection. Instead, the AccECN protocol allows the server to feed back the CE marking to the client, which then has all the information to decide whether the connection has to fall-back from supporting ECN (or not).

Providing feedback of CE marking on the SYN also supports future scenarios in which SYNs might be ECN-enabled (without prejudging whether they ought to be). For instance, in certain environments such as data centres, it might be appropriate to allow ECN-capable SYNs. Then, if feedback showed the SYN had been CE marked, the TCP client could reduce its initial window (IW). It could also reduce IW conservatively if feedback showed the receiver did not support ECN (because if there had been a CE marking, the receiver would not have understood it). Note that this text merely motivates dumb reflection of CE on a SYN, it does not judge whether a SYN ought to be ECN-capable.

3. AccECN Protocol Specification

3.1. Negotiation 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 NS=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 flags CWR=1 and ECE=0 on its response in the SYN/ACK segment to confirm that it supports AccECN. The TCP server MUST NOT set this combination of flags unless the preceding SYN requested support for AccECN as above.

A TCP server in AccECN mode MUST additionally set the flag NS=1 on the SYN/ACK if the SYN was CE-marked (see Section 2.5). If the received SYN was Not-ECT, ECT(0) or ECT(1), it MUST clear NS (NS=0) on the SYN/ACK.

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.

If after the normal TCP timeout the TCP client has not received a SYN/ACK to acknowledge its SYN, the SYN might just have been lost,
e.g. due to congestion, or a middlebox might be blocking segments with the AccECN flags. To expedite connection setup, the host SHOULD fall back to NS=CWR=ECE=0 on the retransmission of the SYN. It would make sense to also remove any other 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. Implementers MAY use other fall-back strategies if they are found to be more effective (e.g. attempting to retransmit a second AccECN segment before fall-back, falling back to classic ECN feedback rather than non-ECN, and/or caching the result of a previous attempt to access the same host while negotiating AccECN).

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 3.2.4.

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. Table 2 tabulates all the negotiation possibilities for ECN-related capabilities that involve at least one AccECN-capable host. To compress the width of the table, the headings of the first four columns have been severely abbreviated, as follows:

Ac: More *Ac*curate ECN Feedback
N:  ECN-*N*once [RFC3540]
E:  *E*CN [RFC3168]
I:  Not-ECN (*I*mplicit congestion notification using packet drop).
### Table 2: ECN capability negotiation between Originator (A) and Responder (B)

<table>
<thead>
<tr>
<th>Ac</th>
<th>N</th>
<th>E</th>
<th>I</th>
<th>SYN A→B</th>
<th>SYN/ACK B→A</th>
<th>Feedback Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NS CWR ECE</td>
<td>NS CWR ECE</td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AB</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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

1. The top block shows the case already described 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.

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. Therefore, as soon as an AccECN-enabled TCP server (B) receives the SYN shown, it MUST set both its half connections into the feedback mode shown in the rightmost column.

4. The fourth block displays combinations that are not valid or currently unused and therefore both ends MUST fall-back to Not ECN for both half connections. Especially the first case (marked ‘broken’) where all bits set in the SYN are reflected by the receiver in the SYN/ACK, which happens quite often if the TCP
The following exceptional cases need some explanation:

ECN Nonce: An AccECN implementation, whether client or server, sender or receiver, does not need to implement the ECN Nonce behaviour [RFC3540]. AccECN is compatible with an alternative ECN feedback integrity approach that does not use up the ECT(1) codepoint and can be implemented solely at the sender (see Section 4.3).

Simultaneous Open: An originating AccECN Host (A), having sent a SYN with NS=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 (see the third block above).

3.2. AccECN Feedback

Each Data Receiver maintains four counters, r.cep, r.ceb, r.e0b and r.e1b. The CE packet counter (r.cep), counts the number of packets the host receives with the CE code point in the IP ECN field, including CE marks on control packets without data. r.ceb, r.e0b and r.e1b count the number of TCP payload bytes in packets marked respectively with the CE, ECT(0) and ECT(1) codepoint in their IP-ECN field. When a host first enters AccECN mode, it initialises its counters to r.cep = 6, r.e0b = 1 and r.ceb = r.e1b. = 0 (see Appendix A.5). Non-zero initial values are used to be distinct from cases where the fields are incorrectly zeroed (e.g. by middleboxes).

A host feeds back the CE packet counter using the Accurate ECN (ACE) field, as explained in the next section. And it 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.

3.2.1. The ACE Field

After AccECN has been negotiated on the SYN and SYN/ACK, both hosts overload the three TCP flags ECE, CWR and NS in the main TCP header as one 3-bit field. Then the field is given a new name, ACE, as shown in Figure 2.
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, it merely overloads them with another name and definition once an AccECN connection has been established.

A host MUST interpret the ECE, CWR and NS flags as the 3-bit ACE counter on a segment with SYN=0 that it sends or receives if both of its half-connections are set into AccECN mode having successfully negotiated AccECN (see Section 3.1). 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).

The ACE field encodes the three least significant bits of the r.cep counter, therefore its initial value will be 0b110 (decimal 6). This non-zero initialization allows a TCP server to use a stateless handshake (see Section 4.1) but still detect from the TCP client’s first ACK that the client considers it has successfully negotiated AccECN. If the SYN/ACK was CE marked, the client MUST increase its r.cep counter before it sends its first ACK, therefore the initial value of the ACE field will be 0b111 (decimal 7). These values have deliberately been chosen such that they are distinct from [RFC5562] behaviour, where the TCP client would set ECE on the first ACK as feedback for a CE mark on the SYN/ACK.

If the value of the ACE field on the first segment with SYN=0 in either direction is anything other than 0b110 or 0b111, the Data Receiver MUST disable ECN for the remainder of the half-connection by marking all subsequent packets as Not-ECT.
3.2.2. 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, the 3-bit counter in the ACE field might have cycled between two ACKs arriving at the Data Sender.

Therefore an AccECN Data Receiver SHOULD immediately send an ACK once ‘n’ CE marks have arrived since the previous ACK, where ‘n’ SHOULD be 2 and MUST be no greater than 6.

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 under the prevailing conditions, it SHOULD conservatively assume that the counter did cycle. 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 neutralise the effect of any action it took based on a conservative assumption that it later found to be incorrect.

3.2.3. The AccECN Option

The AccECN Option is defined as shown below in Figure 3. It consists of three 24-bit fields that provide the 24 least significant bits of the r.e0b, r.ceb and r.elb counters, respectively. 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 = TBD1 | Length = 11 |          EE0B field           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
| EE0B (cont’d) |           ECEB field                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
|                  EE1B field                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: The AccECN Option

The Data Receiver MUST set the Kind field to TBD1, which is registered in Section 6 as a new TCP option Kind called AccECN. An experimental TCP option with Kind=254 MAY be used for initial experiments, with magic number 0xACCE.
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 there is no field to feedback 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.5.

Whenever a Data Receiver sends an AccECN Option, the rules in Section 3.2.5 expect it to always send a full-length option. To cope with option space limitations, it can omit unchanged fields from the tail of the option, 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=11: EE0B, ECEB, EE1B
- Length=8: EE0B, ECEB
- Length=5: EE0B
- Length=2: (empty)

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. For initial experiments, the Length field MUST be 2 greater to accommodate the 16-bit magic number.

All implementations of a Data Sender MUST be able to read in AccECN Options of any of the above lengths. They MUST ignore an AccECN Option of any other length.

### 3.2.4. Path Traversal of the AccECN Option

An AccECN host MUST NOT include the AccECN TCP Option on the SYN. Nonetheless, if the AccECN negotiation using the ECN flags in the main TCP header (Section 3.1) is successful, it implicitly declares that the endpoints also support the AccECN TCP Option.

If the TCP client indicated AccECN support, a TCP server tha confirms its support for AccECN (as described in Section 3.1) SHOULD also include an AccECN TCP Option in 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 need not include an AccECN Option in any of these three cases 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.

If the TCP client has successfully negotiated AccECN but does not receive an AccECN Option on the SYN/ACK, it 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 ACK or on the first data segment, it switches into a mode that assumes that the AccECN Option is not available for this half connection.

While a host is in the mode that assumes the AccECN Option is not available, it MUST adopt the conservative interpretation of the ACE field discussed in Section 3.2.2. However, it cannot make any assumption about support of the AccECN Option on the other half connection, so it MUST continue to send the AccECN Option itself.

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 host SHOULD fall back to NS=CWR=ECE=0 and no AccECN Option on the retransmission of the SYN/ACK. Implementers MAY use other fall-back strategies if they are found to be more effective (e.g. retransmitting a SYN/ACK with AccECN TCP flags but not the AccECN Option; attempting to retransmit a second AccECN segment before fall-back (most appropriate during high levels of congestion); or falling back to classic ECN feedback rather than non-ECN).

Similarly, if the TCP client detects that the first data segment it sent 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 the result of previous attempts.

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

Currently the Data Sender is not required to test whether the arriving byte counters in the AccECN Option have been correctly initialised. This allows different initial values to be used as an additional signalling channel in future. If any inappropriate zeroing of these fields is discovered during testing, this approach will need to be reviewed.
3.2.5. Usage of the AccECN TCP Option

The following rules determine when a Data Receiver in AccECN mode sends the AccECN TCP Option, and which fields to include:

Change-Triggered ACKs: If an arriving packet increments a different byte counter to that incremented by the previous packet, the Data Receiver SHOULD immediately send an ACK with an AccECN Option, without waiting for the next delayed ACK. Certain offload hardware might not be able to support change-triggered ACKs, but otherwise it is important to keep exceptions to this rule to a minimum so that Data Senders can generally rely on this behaviour;

Continual Repetition: Otherwise, if arriving packets continue to increment the same byte counter, the Data Receiver can include an AccECN Option on most or all (delayed) ACKs, but it does not have to. If option space is limited on a particular ACK, the Data Receiver MUST give precedence to SACK information about loss. It SHOULD include an AccECN Option if the r.ce0b counter has incremented and it MAY include an AccECN Option if r.ec0b or r.ec1b has incremented;

Full-Length Options Preferred: It SHOULD always use full-length AccECN Options. It MAY use shorter AccECN Options if space is limited, but it MUST include the counter(s) that have incremented since the previous AccECN Option and it MUST only truncate fields from the right-hand tail of the option to preserve the order of the remaining fields (see Section 3.2.3);

Beaconing Full-Length Options: Nonetheless, it MUST include a full-length AccECN TCP Option on at least three ACKs per RTT, or on all ACKs if there are less than three per RTT (see Appendix A.4 for an example algorithm that satisfies this requirement).

The following example series of arriving marks illustrates when a Data Receiver will emit an ACK if it is using a delayed ACK factor of 2 segments and change-triggered ACKs: 01 -> ACK, 01, 01 -> ACK, 10 -> ACK, 10, 01 -> ACK, 01, 11 -> ACK, 01 -> ACK.

For the avoidance of doubt, the change-triggered ACK mechanism ignores the arrival of a control packet with no payload, because it does not alter any byte counters. The change-triggered ACK approach will lead to some additional ACKs but it feeds back the timing and the order in which ECN marks are received with minimal additional complexity.

Implementation note: sending an AccECN Option each time a different counter changes and including a full-length AccECN Option on every
delayed ACK will satisfy the requirements described above and might be the easiest implementation, as long as sufficient space is available in each ACK (in total and in the option space).

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 the above rules.

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

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.

Another large class of middleboxes intervene 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 'normalise' the TCP wire protocol by checking that all values in header fields comply with a rather narrow interpretation of the TCP specifications. To comply with the present AccECN specification, such a middlebox MUST NOT change the ACE field or the AccECN Option and it MUST attempt to preserve the timing of each ACK (for example, if it coalesced ACKs it would not be AccECN-compliant). A middlebox claiming to be transparent at the transport layer MUST 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 are correct. This is because blocking apparently invalid values does not improve security (because AccECN hosts are required to ignore invalid values anyway), while it prevents the standardised set of values being extended in future (because outdated normalisers would block updated hosts from using the extended AccECN standard).

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'. Leeway has been allowed in the present AccECN specification in the expectation that offload hardware could comply and still serve its function. Nonetheless, such hardware MUST attempt to preserve the timing of each ACK (for example, if it coalesced ACKs it would not be AccECN-compliant).
4. Interaction with Other TCP Variants

This section is informative, not normative.

4.1. 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 the first ACK it receives contains an ACE field with the value 0b110 or 0b111, it can assume that:

- the TCP client must have requested AccECN support on the SYN
- 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.

4.2. Compatibility with Other TCP Options and Experiments

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 3.2.5 provides guidance on how important it is to send an AccECN Option and whether it needs to be a full-length option.

4.3. Compatibility with Feedback Integrity Mechanisms

The ECN nonce [RFC3540] is an experimental IETF specification intended to allow a sender to test whether ECN CE markings (or losses) introduced in one network are being suppressed by the receiver or anywhere else in the feedback loop, such as another network or a middlebox. The ECN nonce has not been deployed as far
as can be ascertained. The nonce would now be nearly impossible to deploy retrospectively, because to catch a misbehaving receiver it relies on the receiver volunteering feedback information to incriminate itself. A receiver that has been modified to misbehave can simply claim that it does not support nonce feedback, which will seem unremarkable given so many other hosts do not support it either.

With minor changes AccECN could be optimised for the possibility that the ECT(1) codepoint might be used as a nonce. However, given the nonce is now probably undeployable, the AccECN design has been generalised 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.

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 [I-D.moncaster-tcpm-rcv-cheat]. Unlike the ECN Nonce, this approach does not waste the ECT(1) codepoint in the IP header, it does not require standardisation 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 sparsely.

- o Networks generate congestion signals when they are becoming congested, so they 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 [I-D.ietf-conex-abstract-mech]. Whether the receiver or a downstream network is suppressing congestion feedback or the sender is unresponsive to the feedback, or both, ConEx audit can neutralise any advantage that any of these three parties would otherwise gain.

ConEx is a change to the Data Sender that is 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.
The 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.

5. 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. The receiver can control how frequently it sends the AccECN TCP Option and therefore 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.

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 option 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.

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, to ensure that any blocking of anomalous values is
always at least under reversible policy control.

6. IANA Considerations

This document defines a new TCP option for AccECN, assigned a value
of TBD1 (decimal) from the TCP option space. This value is defined
as:

<table>
<thead>
<tr>
<th>Kind</th>
<th>Length</th>
<th>Meaning</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD1</td>
<td>N</td>
<td>Accurate ECN (AccECN)</td>
<td>RFC XXXX</td>
</tr>
</tbody>
</table>

[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 implementation before the IANA allocation MUST follow [RFC6994]
and use experimental option 254 and magic number 0xACCE (16 bits)
(ToDo register this with IANA), then migrate to the new option after
the allocation.

7. 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). 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 presented, the AccECN Data Sender is then required
to switch to more conservative assumptions about wrap of congestion
indication counters (see Section 3.2.2 and Appendix A.2).

Section 4.1 describes how a TCP server can negotiate AccECN and use
the SYN cookie method for mitigating SYN flooding attacks.
There is concern that ECN markings could be altered or suppressed, particularly because a misbehaving Data Receiver could increase its own throughput at the expense of others. Given the experimental ECN nonce is now probably undeployable, AccECN has been generalised for other possible uses of the ECT(1) codepoint to avoid obsolescence of the codepoint even if the nonce mechanism is obsoleted. AccECN is compatible with the three other schemes known to assure the integrity of ECN feedback (see Section 4.3 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.

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.

8. Acknowledgements

We want to thank Koen De Schepper, Praveen Balasubramanian and Michael Welzl 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.

Bob Briscoe was part-funded by the European Community under its Seventh Framework Programme through the Reducing Internet Transport Latency (RITE) project (ICT-317700) and through the Trilogy 2 project (ICT-317756). The views expressed here are solely those of the authors.

9. 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.

10. References

10.1. Normative References


10.2. Informative References


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 modulo operator.

On the arrival of an AccECN Option, the Data Sender uses the TCP acknowledgement number and any SACK options to calculate newlyAckedB, the amount of new data that the ACK acknowledges in bytes. If newlyAckedB is negative it means that a more up to date ACK has already been processed, so this ACK has been superseded and the Data Sender has to ignore the AccECN Option. Then the Data Sender calculates the minimum difference d.ceb between the ECEB field and its local s.ceb counter, using modulo arithmetic as follows:

    if (newlyAckedB >= 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

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 and Section 3.2.4); 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 a CE marked packet arrives, 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 uses the TCP acknowledgement number and any SACK options to calculate newlyAckedB, the amount of new data that the ACK acknowledges. If newlyAckedB is negative it means that a more up to date ACK has already been processed, so this ACK has been superseded and the Data Sender has to ignore the AccECN Option. If newlyAckedB is zero, to break the tie the Data Sender could use timestamps (if present) to work out newlyAckedT, the amount of new time that the ACK acknowledges. Then the Data Sender calculates the minimum difference
d.cep between the ACE field and its local s.cep counter, using modulo arithmetic as follows:

\[
\text{if } ((\text{newlyAckedB} > 0) \lor (\text{newlyAckedB} == 0 \&\& \text{newlyAckedT} > 0)) \\
\quad \text{d.cep} = (\text{ACE} + \text{DIVACE} - (\text{s.cep} \mod \text{DIVACE})) \mod \text{DIVACE}
\]

Section 3.2.2 requires the Data Sender to assume that the ACE field did cycle if it could have cycled under prevailing conditions. The 3-bit ACE field in an arriving ACK could have cycled and become ambiguous to the Data Sender if a row 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 lost ACKs are 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 the Data Sender to take account of the prevailing size of data segments and the prevailing CE marking rate just before the sequence of ACK losses. 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

\[
\text{dSafer.cep} = \text{newlyAckedPkt} - ((\text{newlyAckedPkt} - \text{d.cep}) \mod \text{DIVACE}),
\]

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).

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 says "the Data Sender MAY attempt to neutralise 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. However, if for some reason it needs CE-marked packets, if dSafer.cep is different from d.cep, it can calculate the average marked segment size that each implies to determine whether d.cep is likely to be a safe enough estimate. Specifically, it could use the following algorithm, where d.ceb is the amount of newly CE-marked bytes (see Appendix A.1):

```plaintext
SAFETY_FACTOR = 2
if (dSafer.cep > d.cep) {
    s = d.ceb/d.cep
    if (s <= MSS) {
        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:
The following examples give the reasoning behind the algorithm, assuming MSS=1,460 [B]:

- if d.cep=0, dSafer.cep=8 and d.ceb=1,460, then s=\infty 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.

- if d.cep=2, dSafer.cep=10 and d.ceb=1,460, 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.

- if d.cep=7, dSafer.cep=15 and d.ceb=10,200, then s=1,457 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.
To calculate s_ave, it 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 it could update once per
RTT. Either way, it would estimate s_ave as:

\[ s_\text{ave} \approx \frac{\text{flightsize}}{\text{packets}_\text{in_flight}}, \]

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 \( \lg(\text{packets}_\text{in_flight}) \), where \( \lg() \) means log base 2.

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

\[ s_\text{ave} = a * s + (1-a) * s_\text{ave}, \]

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 Beacon AccECN Options

Section 3.2.5 requires a Data Receiver to beacon a full-length AccECN
Option at least 3 times per RTT. This could be implemented by
maintaining a variable to store the number of ACKs (pure and data
ACKs) since a full AccECN Option was last sent and another for the
approximate number of ACKs sent in the last round trip time:

\[
\text{if}(\text{acks}_\text{since}_\text{full}_\text{last}_\text{sent} > \text{acks}_\text{in}_\text{round} / \text{BEACON}_\text{FREQ})
\text{send}_\text{full}_\text{AccECN}_\text{Option}() \]

For optimised integer arithmetic, \( \text{BEACON}_\text{FREQ} = 4 \) could be used,
rather than 3, so that the division could be implemented as an
integer right bit-shift by \( \lg(\text{BEACON}_\text{FREQ}) \).

In certain operating systems, it might be too complex to maintain
acks_in_round. In others it might be possible by tagging each data
segment in the retransmit buffer with the number of ACKs sent at the
point that segment was sent. This would not work well if the Data
Receiver was not sending data itself, in which case it might be
necessary to beacon based on time instead, as follows:

\[
\text{if}(\text{time}_\text{now} > \text{time}_\text{last}_\text{option}_\text{sent} + \text{RTT} / \text{BEACON}_\text{FREQ})
\text{send}_\text{full}_\text{AccECN}_\text{Option}() \]
However, this time-based approach does not work well when all the ACKs are sent early in each round trip, as is the case during slow-start.

(ToDo: A simple and robust beaconing algorithm for all circumstances is still work-in-progress.)

A.5. 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.elb. Note that, because r.e0b is initialised to 1 and the other two counters are initialised to 0, the initial sum will be 1, which matches the initial offset of the TCP sequence number on completion of the 3WHS.

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 transmissions 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. Alternative Design Choices (To Be Removed Before Publication)

This appendix is informative, not normative. It records alternative designs that the authors chose not to include in the normative specification, but which the IETF might wish to consider for inclusion:

Feedback all four ECN codepoints on the SYN/ACK: The last two negotiation combinations in Table 2 could also be used to indicate AccECN support and to feedback that the arriving SYN was ECT(0) or ECT(1). This could be used to probe the client to server path for incorrect forwarding of the ECN field [I-D.kuehlewind-tcpm-ecn-fallback]. Note, however, that it would be unremarkable if ECN on the SYN was zeroed by security devices,
given RFC 3168 prohibited ECT on SYN because it enables DoS
attacks.

Feedback all four ECN codepoints on the First ACK: To probe the
server to client path for incorrect ECN forwarding, it could be
useful to have four feedback states on the first ACK from the TCP
client. This could be achieved by assigning four combinations of
the ECN flags in the main TCP header, and only initialising the
ACE field on subsequent segments.

Empty AccECN Option: It might be useful to allow an empty (Length=2)
AccECN Option on the SYN/ACK and first ACK. Then if a host had to
omit the option because there was insufficient space for a larger
option, it would not give the impression to the other end that a
middlebox had stripped the option.

Appendix C. Open Protocol Design Issues (To Be Removed Before
Publication)

1. Currently it is specified that the receiver 'SHOULD' use Change-
   Triggered ACKs. It is controversial whether this ought to be a
   'MUST' instead. A 'SHOULD' would leave the Data Sender uncertain
   whether it can rely on the timing and ordering information in
   ACKs. If the sender guesses wrongly, it will probably introduce
   at least 1RTT of delay before it can use this timing information.
   Ironically it will most likely be wanting this information to
   reduce ramp-up delay. A 'MUST' could make it hard to implement
   AccECN in offload hardware. However, it is not known whether
   AccECN would be hard to implement in such hardware even with a
   'SHOULD' here. For instance, was it hard to offload DCTCP to
   hardware because of change-triggered ACKs, or was this just one
   of many reasons? The choice between MUST and SHOULD here is
   critical. Before that choice is made, a clear use-case for
   certainty of timing and ordering information is needed, plus
   well-informed discussion about hardware offload constraints.

2. There is possibly a concern that a receiver could deliberately
   omit the AccECN Option pretending that it had been stripped by a
   middlebox. No known way can yet be contrived to take advantage
   of this downgrade attack, but it is mentioned here in case
   someone else can contrive one.

3. The s.cep counter might increase even if the s.ccb counter does
   not (e.g. due to a CE-marked control packet). The sender’s
   response to such a situation is considered out of scope, because
   this ought to be dealt with in whatever future specification
   allows ECN-capable control packets. However, it is possible that
   the situation might arise even if the sender has not sent ECN-
capable control packets, in which case, this draft might need to
give some advice on how the sender should respond.

Appendix D. Changes in This Version (To Be Removed Before Publication)

The difference between any pair of versions can be displayed at
history/>

From 04 to 05::

* Corrected ambiguity between Classic ECN and Classic ECN
  feedback throughout

* Changed MUST to SHOULD send AccECN option on SYN/ACK last ACK
  of 3WHS and first data segment from client, to allow for cached
  knowledge of option traversal problems.

* Removed duplication of normative language about sending a full-
  length option in the sections on "The AccECN Option" and "Usage
  of the AccECN Option", and mutually cross referenced.

* Acknowledged Koen De Schepper and Praveen Balasubramanian

* Noted in Appendix that algo to beacon a full-length option is
  work-in-progress

* Editorial corrections and clarifications throughout

Authors’ Addresses

Bob Briscoe
Simula Research Laboratory
EMail: ietf@bobbriscoe.net
URI: http://bobbriscoe.net/

Mirja Kuehlewind
ETH Zurich
Gloriastrasse 35
Zurich 8092
Switzerland
EMail: mirja.kuehlewind@tik.ee.ethz.ch
A-PAWS: Alternative Approach for PAWS
draft-nishida-tcpm-apaws-02

Abstract

This document describes a technique called A-PAWS which can provide protection against old duplicates segments like PAWS. While PAWS requires TCP to set timestamp options in all segments in a TCP connection, A-PAWS supports the same feature without using timestamps. A-PAWS is designed to be used complementary with PAWS. TCP needs to use PAWS when it is necessary and activates A-PAWS only when it is safe to use. Without impairing the reliability and the robustness of TCP, A-PAWS can provide more option space to other TCP extensions.

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 http://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 April 19, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect
1. Introduction

PAWS (Protect Against Wrapped Sequences) defined in [RFC1323] is a technique that can identify old duplicate segments in a TCP connection. An old duplicate segment can be generated when it has been delayed by queueing, etc. If such a segment has the sequence number which falls within the receiver’s current window, the receiver will accept it without any warning or error. However, this segment can be a segment created by an old connection that has the same port and address pair, or a segments sent $2^{32}$ bytes earlier on the same connection. Although this situation rarely happens, it impairs the reliability of TCP.

PAWS utilizes timestamp option in [RFC1323] to provide protection against this. It is assumed that every received TCP segment contains a timestamp. PAWS can identify old duplicate segments by comparing the timestamp in the received segments and the timestamps from other segments received recently. If both TCP endpoints agree to use PAWS, all segments belong to this connection should have timestamp. Since PAWS is the only standardized protection against old duplicate segments, it has been implemented and used in most TCP
implementations. However, as some TCP extensions such as [RFC2018], [RFC5925] and [RFC6824] also requires a certain amount of option space in non-SYN segments, using 10-12 bytes length in option space for timestamp in all segments tends to be considered expensive in recent discussions.

In addition, although PAWS is necessary for connections which transmit more than $2^{32}$ bytes, it is not very important for other connections since [RFC0793] already has protection against segments from old connections by using timers. Moreover, some research results indicates that most of TCP flows tend to transmit small amount of data, which means only small fraction of TCP connections really need PAWS [QIAN11]. Timestamp option is also used for RTTM (Round Trip Time Measurement) in [RFC1323]. Gathering many RTT samples from the timestamp in every TCP segment looks useful approach to improve RTO estimation. However, some research results shows the number of samples per RTT does not affect the effectiveness of the RTO [MALLMAN99]. Hence, we can think if PAWS is not used, sending a few timestamps per RTT will be sufficient.

Based on these observations, we propose a new technique called A-PAWS which can archive similar protection against old duplicates segments. The basic idea of A-PAWS is to attain the same protection against old all duplicate segments as PAWS while reducing the use of TS options in segments. A-PAWS is designed to be used complementary with PAWS. This means an implementation that supports A-PAWS is still required to supports PAWS. A-PAWS is activated only when it is safe to use. This sounds the applicability of A-PAWS is limited, however, we believe TCP will have a lot of chances to save the option space if it uses A-PAWS.

There are some discussions that PAWS can also be used to enhance security, however, we still believe that A-PAWS can maintain the same level of security as PAWS. Detailed discussions on this point are provided in Section 5. A-PAWS is an experimental idea yet, but we hope it will contribute to facilitating the use of TCP option space.

2. Conventions and Terminology

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 [RFC2119].

3. The A-PAWS Design

A-PAWS assumes PAWS as it is designed to be used complementary with PAWS. Hence, a node which supports A-PAWS MUST support PAWS. The following mechanisms are required in TCP in order to perform A-PAWS.
3.1. Signaling Methods

An endpoint that supports A-PAWS can use the following signaling methods to activate A-PAWS logic.

1) Option Exchange in SYN

This method uses a new experimental TCP option defined in [RFC6994] and exchanges it during SYN negotiation. The format of the option is depicted in Figure 1. The option does not have any content as it simply indicates the endpoint supports A-PAWS. In this signaling method, when an endpoint wants to use A-PAWS, it MUST put A-PAWS option in SYN or SYN-ACK segment. If an endpoint does not find A-PAWS option in received SYN or SYN-ACK segment, it MUST not send segments with A-PAWS logic in Section 3.3. However, it MUST activate A-PAWS receiver logic in Section 3.4 if it has sent A-PAWS option in SYN or SYN-ACK segment. This is because some middleboxes may remove A-PAWS option in SYN or SYN-ACK segment. A-PAWS receiver logic in Section 3.4 can interact with both A-PAWS and PAWS sender. This signaling requires additional option space in SYN segments, hence non-SYN segment signaling should be used when there is not enough space in SYN option space.

2) Option Exchange in non-SYN Segments

This method uses the option in Figure 1 as well as the SYN segment signaling. However, the options are not exchanged during SYN negotiation. When a endpoint sets A-PAWS option in the segments, it indicates that it can receive the segments from A-PAWS senders. Hence, it MUST activate A-PAWS receiver logic in Section 3.4 if it sends the options. However, it MUST not send segments with A-PAWS logic in Section 3.3 until it receives A-PAWS options. This approach does not require extra option space or special timestamp value in SYN segments. However, negotiating features in non-SYN segments will require to address further arguments such as when to send the options or how to retransmits the options. We discuss these points in the next section and provide some recommended rules for implementations.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+------------------------------+
|  Kind = 254   |  Length = 4   |       16-bit ExID = TBD      |
+---------------+---------------+------------------------------+
```

Figure 1: A-PAWS option format
3.2. A-PAWS Negotiation Logic for non-SYN Segment Signaling

One important characteristic for A-PAWS is its signaling mechanism does not require tight synchronization between endpoints since A-PAWS receivers can interact with both A-PAWS senders and PAWS senders. This allows us not to invent another three-way handshake like mechanisms for non-SYN segments. This approach will require drastic changes in the current TCP semantics. Instead, we propose a relatively simple and easy mechanism for feature negotiation by using the following rules on A-PAWS endpoints.

Rule 1: An endpoint MUST activate A-PAWS receiver logic in Section 3.4 before it sends A-PAWS option.

Rule 2: An endpoint MUST not send segments with A-PAWS logic in Section 3.3 until it receives A-PAWS option from the other endpoint.

These rules can avoid situations where an endpoint sends segments by A-PAWS logic to an endpoint that doesn’t use A-PAWS logic.

Another discussion point for this signaling method is when to set A-PAWS option in segments. As A-PAWS employs asynchronous signaling, both endpoints basically can set A-PAWS option in segments anytime they want. However, it is recommended to use the following rules for setting A-PAWS options.

Rule 3: An endpoint SHOULD use a data segment when it sets A-PAWS option in a segment.

Rule 4: When an endpoint receives a data segment with A-PAWS option, it SHOULD set A-PAWS option for its ACK segment.

Rule 5: An endpoint MAY use A-PAWS options in retransmitted segments.

These rules allow endpoints to have loose synchronized signaling so that they can at least solicit responses from their peers. Of course, even an endpoint solicit a response by setting A-PAWS option in a data segment, it might not receive A-PAWS option in the ACK segment. This can be caused by the lost of the ACK segment or middleboxes that remove unknown options. In order to address these cases, the following rules can be used.

Rule 6: As long as an endpoint does not violate the other rules, it MAY set A-PAWS option in multiple data segments with a certain interval in case no A-PAWS options has been sent from the peer.
This rule can address the cases where A-PAWS options has been removed by middleboxes or segments with A-PAWS options has been lost.

3.3. Sending Behavior

A-PAWS enabled TCP transmits segments, it needs to follow the rules below.

1. TCP needs to check how many bytes has been transmitted in a connection. If the transmitted bytes exceeds $2^{32} - \text{`Sender.Offset'},$ TCP migrates PAWS mode and MUST set timestamp option in all segments to be transmitted. The value for \textit{`Sender.Offset'} is discussed in Section 5.

2. If the number of bytes transmitted in a TCP connection does not exceeds $2^{32} - \text{`Sender.Offset'},$ TCP MAY omit timestamp option in segments as long as it does not affect RTTM. This draft does not define how much TCP can omit timestamps because it should be determined by RTTM.

3.4. Receiving Behavior

A-PAWS enabled TCP receives segments, it needs to follow the rules below.

1. TCP needs to check how many bytes has been received in a TCP connection. If it exceeds $2^{32}$ bytes, A-PAWS nodes SHOULD discard the received segments which does not have timestamp option. TCP MUST perform PAWS check when received bytes exceeds $2^{32}$ bytes.

2. If the number of bytes received in a TCP connection does not exceeds $2^{32}$ bytes, A-PAWS nodes SHOULD accept the segments even if it does not have timestamp option. A-PAWS nodes MAY skip PAWS check until the received bytes exceeds $2^{32}$ bytes.

4. When To Activate A-PAWS

In basic principal, A-PAWS capable nodes can always use A-PAWS logic as long as the peers agree with them. However, the following cases require special considerations to enable A-PAWS.

1. As "When To Keep Quiet" section in [RFC0793] suggests, it is recommended that TCP keeps quiet for a MSL upon starting up or recovering from a crash where memory of sequence numbers has been lost. However, if timestamps are being used and if the timestamp clock can be guaranteed to be increased monotonically, this quiet time may be unnecessary. Because TCP can identify the segments
from old connections by checking the timestamp. We think some TCP implementations may disable the quiet time because of using timestamps from this reason. However, since A-PAWS nodes does not set timestamp options in all segments, TCP cannot rely on this approach. To avoid decreasing the robustness of TCP connection, TCP MUST NOT use A-PAWS for a MSL upon starting up or recovering from a crash.

2. Various TCP implementations provide APIs such as setsockopt() that can set SO_REUSEADDR flag on TCP connections. If this flag is set, the TCP connection allows to reuse the same local port without waiting for 2 MSL period. While this option is useful when users want to relaunch applications immediately, it makes the TCP connection a little vulnerable as TCP stack might receive duplicate segments from earlier incarnations. It has been said that PAWS can contribute to mitigate this risk by checking the timestamps in segments. In order to keep the same level of protection, TCP SHOULD NOT send A-PAWS option when SO_REUSEADDR flag is set. This rule prevents the peer from sending segments to this node with A-PAWS logic. However, the node can send segments with A-PAWS logic as long as it received A-PAWS option from the peer.

5. Discussion

As A-PAWS is an experimental logic, the following points need to be considered and discussed.

5.1. Protection Against Early Incarnations

There are some discussions that timestamp can enhance the robustness against early incarnations. Since A-PAWS does not set timestamps in all segments, some may say that it degrades the robustness of TCP. We believe that the degradation caused by A-PAWS on this point is negligible. As long as TCP limits the usage of A-PAWS as described in Section 4, duplicate segments from early incarnations should not be received by TCP.

5.2. Protection Against Security Threats

A TCP connection can be identified by a 5-tuple: source address, destination address, source port number, destination port number and protocol. Crackers need to guess all these parameters when they try malicious attacks on the connection. PAWS can enhance the protection for this as it additionally requires timestamp checking. However, we think the effect of PAWS against malicious attacks is limited due to the simplicity of PAWS check. In PAWS, a segment can be considered as an old duplicate if the timestamp in the segment less than some
timestamps recently received on the connection. The "less than" in this context is determined by processing timestamp values as 32 bit unsigned integers in a modular 32-bit space. For example, if t1 and t2 are timestamp values, t1 < t2 is verified when 0 < (t2 - t1) < $2^{31}$ computed in unsigned 32-bit arithmetic. Hence, if crackers set a random value in the timestamp option, there will be 50% chance for them to trick PAWS check. Moreover, there will be more chances if they send multiple segments with different timestamps, which will not be difficult to perform.

In addition, we think there might be a case where using PAWS increases security risks. PAWS recommends to increase timestamp over a system when TCP waives the "quiet time" described in [RFC0793]. However, if timestamps are generated from a global counter, it may leak some information such as system uptime as discussed in [SILBERSACK05]. A-PAWS might be able to allows TCP to use random timestamp values per connections.

5.3. Middlebox Considerations

A-PAWS is designed to be robust against middleboxes. This means that endpoints will not be messed up even if middleboxes discard A-PAWS option. This is because A-PAWS sender logic is activated only when TCP receives a segment with A-PAWS options. A-PAWS receiver logic does not need to know whether the sender is using PAWS or A-PAWS. Activating A-PAWS receiving logic for PAWS sender might be redundant as it requires additional overheads. However, we believe the overhead will be acceptable in most cases because of the simplicity of A-PAWS logic.

Another concern on middleboxes is that they can insert or delete some bytes in TCP connections. If a middlebox inserts extra bytes into a TCP connections, there might be a situation where an A-PAWS sender can transmit segments without timestamp, while an A-PAWS receiver perform PAWS check on them as it already has received $2^{32}$ bytes. In order to avoid discarding segments unnecessarily, we recommend that A-PAWS sender should have a certain amount of offset bytes in order to migrate PAWS mode before the receiver receives $2^{32}$ bytes. We call this protocol parameter 'Sender.Offset'. The proper value for 'Sender.Offset' needs to be discussed.

5.4. Aggressive Mode in A-PAWS

The current A-PAWS requires TCP to migrate PAWS mode after sending/receiving $2^{32}$ bytes. However, if both nodes check if 2 MSL has already passed during sending/receiving $2^{32}$ bytes, it is safe to continue using A-PAWS. We call this Aggressive mode. The use of Aggressive mode will be explored in future versions.
6. Security Considerations

We believe A-PAWS can maintain the same level of security as PAWS does, but further discussions will be needed. Some security aspects of A-PAWS are discussed in Section 5.

7. IANA Considerations

This document uses the Experimental Option Experiment Identifier. An application for this codepoint in the IANA TCP Experimental Option ExID registry will be submitted.

8. References

8.1. Normative References


8.2. Informative References


[SILBERSACK05]

Author’s Address
Yoshifumi Nishida
GE Global Research
2623 Camino Ramon
San Ramon, CA 94583
USA

Email: nishida@wide.ad.jp
Increasing Maximum Window Size of TCP

draft-nishida-tcpm-maxwin-01.txt

Abstract

This document proposes to increase the current max window size allowed in TCP. It describes the current logic that limits the max window size and provides a rationale to relax the limitation as well as the negotiation mechanism to enable this feature safely.

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 http://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 April 19, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.
1. Introduction

TCP throughput is determined by two factors: Round Trip Time and Receive Window size. It can never exceed Receive Window size divided by RTT. This implies larger window size is important to achieve better performance. Original TCP’s maximum window size defined in RFC793 [RFC0793] is $2^{16} - 1$ (65,535), however, RFC7323 [RFC7323] defines TCP Window Scale option which allows TCP to use larger window size. Window Scale uses a shift count stored in 1-byte field in the option. The receiver of the option uses left-shifted window size value by the shift count as actual window size. When Window Scale is used, TCP can extend maximum window size to $2^{30} - 2^{14}$ (1,073,725,440). This is because the maximum shift count is 14 as described in the Section 2.3 of RFC7323 [RFC7323]. However, since TCP’s sequence number space is $2^{32}$, we believe it is still possible to use larger window size than this while careful design of the logic that can identify segments inside the window is required. In this document, we propose to increase the maximum shift count to 15, which extend window size to $2^{31} - 2^{15}$.

2. Conventions and Terminology

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 [RFC2119].

3. Increasing Maximum Window Size

RFC7232 requires maximum window size to be less than $2^{30}$ as described below.
TCP determines if a data segment is "old" or "new" by testing whether its sequence number is within 2^31 bytes of the left edge of the window, and if it is not, discarding the data as "old". To insure that new data is never mistakenly considered old and vice versa, the left edge of the sender’s window has to be at most 2^31 away from the right edge of the receiver’s window. The same is true of the sender’s right edge and receiver’s left edge. Since the right and left edges of either the sender’s or receiver’s window differ by the window size, and since the sender and receiver windows can be out of phase by at most the window size, the above constraints imply that two times the maximum window size must be less than 2^31, or

max window < 2^30

However, TCP does not necessarily need to determine if a segment is old or new. Because important point is to determine if a receive segment is inside of the window or not. It basically does not matter if a segment is too old (left side of the window) or too new (right side of the window) as long as it is outside of the window. Based on this viewpoint, we propose to extend maximum window to 2^31 - 2^15, which can be attained by increasing maximum shift count to 15.

To demonstrate the feasibility of the proposal, we would like to use the following worst case example where the sender and the receiver windows are completely out of phase. In this example, we define S as the sender’s left edge of the window and W as the sender’s window size. Hence, the sender’s right edge of the window is S+W. Also, the receiver’s left edge of the window is S+W+1 and the right edge of the window is S+2W+1, as they are out of phase. This situation can happen when the sender sent all segments in the window and the receiver received all segments while no ACK has been received by the sender yet. Now, we presume a segment that contains sequence number S has arrived at the receiver. This segment should be excluded by the receiver, although it can easily happen when the sender retransmits segments.

In case of W=2^31, the receiver cannot exclude this segment as S+2W = S. It is considered inside of the window. (S+W+1 < S < S+2W+1). However, our proposed window size is W=2^31-X, where X is 2^15. In this case, when segment S has arrived, the following checks will be performed. First, TCP checks it with the left edge of the window and it considers the segment is left side of the left edge. (S < S+W+1 Note: W=2^31-X) Second, TCP checks it with the right edge of the window and it considers the segment is right of the right edge. (S > S+2W+1) You might notice that the result of the second check is not expected one as the segment S is actually an old segment. This is
the problem that the referred paragraphs from RFC7232 [RFC7323] describe. However, the segment is properly excluded by the receiver as both checks indicate it is outside of the window. It should be noted that the principle of TCP requires to accept the segment S only when it has passed both checks successfully, which means S must satisfy the following condition.

\[ S \geq \text{left edge} \land S \leq \text{right edge} \]

As we have shown in the example, our proposed maximum window size: \( W=2^{31}-2^{15} \) does not affect this principle.

4. Signaling Method

In this proposal, the following two types of signaling methods are available.

4.1. Non-synchronized Signalling

Non-synchronized signaling method does not require to verify that both endnodes supports new maximum window size. In this method, an end node that supports new maximum window size simply sets 15 to the shift count in Window Scale option. A potential problem of this approach is that an endnode cannot verify if the peer supports shift count 15 when the peer does not use shift count 15 in its Window Scale option. However, we believe this is not a significant problem. An endnode that does not support shift count 15 will simply use 14 as specified in [RFC7323]. This will not cause a fatal issue while the window size used by the sender will be smaller than the actual available window size at the receiver. One way to mitigate this issue is to advertise new maximum window size (or closer value) as much as possible. When an RFC7323 endnode sees new maximum window size value, it is automatically parsed as the current maximum window size.

4.2. Synchronized Signaling

Synchronized signaling that can verify both endnodes support new maximum window size can be achived by introducing a new TCP option. In this method, when a sender sends Window Scale option with shift count 15 in the SYN segment, it also includes new TCP option to indicate it supports new maximum value. The receiver of the segment MUST include this new option in the SYN/ACK segment if it supports new maximum value. When the sender receives this option, it can use shift count 15, otherwise it should use shift count 14. This method guarantees that both endpoints can correctly parse the new shift count.
5. Use Cases, Benefits to Explore Maximum Window Size

One of the use cases of the extended maximum window size is high volume data transfer over paths with long RTT delays and high bandwidth, called long fat pipes. The proposed extension improves and doubles at most the maximum throughput when bandwidth-latency product is greater than 1 GB. As propagation delay in an optical fiber is around 20 cm/ns, RTT will be over 100 milliseconds when the distance of the transmission is more than 10000km. This distance is not extraordinary for trans-pacific communications. In this case, the maximum throughput will be limited to 80Gbps with the current maximum window size, although network technologies for more than 100 GBps are becoming common these days.

As the current TCP sequence number space is limited to 32 bits, it will not be possible to increase maximum window size any further. However, TCP may eventually have other extensions to increase sequence number space, for example, [RFC7323] and [RFC1263] mention about increasing sequence number space to 64 bits. We believe the information in this document will be useful when such extensions are proposed as they need to define new maximum window size.

6. Acknowledgments

The authors gratefully acknowledge significant inputs for this document from Richard Scheffenegger.

7. Security Considerations

It is known that an attacker can have more chances to insert forged packets into a TCP connection when large window size is used. This is not a specific problem of this proposal, but a generic problem to use larger window. Using PAWS can mitigate this problem, however, it is recommended to consult the Security Considerations section of RFC7323 [RFC7323] to check its implications.

8. IANA Considerations

This document may request new TCP option codepoint.

9. References

9.1. Normative References


9.2. Informative References


Authors’ Addresses

Yoshifumi Nishida
GE Global Research
2623 Camino Ramon
San Ramon, CA  94583
USA

Email: nishida@wide.ad.jp

Hirochika Asai
The University of Tokyo
7-3-1 Hongo
Bunkyo-ku, Tokyo  113-8656
JP

Email: panda@wide.ad.jp
The state of implementation of TCP control block interdependence
draft-welzl-tcpm-tcb-sharing-00

Abstract

This document provides an overview of the state of implementation of
RFC 2140, in preparation for a possible future RFC2140bis document.

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 http://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 April 21, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the
document authors.  All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal
Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of
publication of this document.  Please review these documents
carefully, as they describe your rights and restrictions with respect
to this document.  Code Components extracted from this document must
include Simplified BSD License text as described in Section 4.e of
1. State of Implementation

* L = Linux, F = FreeBSD

<table>
<thead>
<tr>
<th>RFC 2140</th>
<th>Description</th>
<th>Implementation</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old-MSS</td>
<td>Maximum Segment Size</td>
<td>F:rmx_mtu</td>
<td>This is being cached and shared in FreeBSD.</td>
</tr>
<tr>
<td>Old-RTT</td>
<td>Estimated Round-Trip Time</td>
<td>L:TCP_METRIC_RTT F:rmx_rtt</td>
<td>Cached in both FreeBSD and Linux, however it is being used by a new connection in FreeBSD only.</td>
</tr>
<tr>
<td>Old-RTT</td>
<td>Estimated Round-Trip Time</td>
<td>L:TCP_METRIC_RTTVAR F:rmx_rttvar</td>
<td>Cached in both FreeBSD and Linux, however it is being used by a new connection in FreeBSD only.</td>
</tr>
<tr>
<td>Old-snd_cwnd</td>
<td>Congestion Window</td>
<td>L:TCP_METRIC_CWND F:rmx_cwnd</td>
<td>Cached in both FreeBSD and Linux, however it is not being used by a new connection.</td>
</tr>
<tr>
<td>-</td>
<td>Slow Start Thresold</td>
<td>L:TCP_METRIC_SSTHRESH F:rmx_ssthresh</td>
<td>This is being cached and shared in both FreeBSD and Linux. In Linux, it is set to max(cwnd/2, ssthresh) in most cases.</td>
</tr>
</tbody>
</table>
FreeBSD, however, it is set to either the current ssthresh if not set previously, or to the arithmetic ssthresh and previously cached metric.

- | Metric related to the extent of reordering. | L:TCP_METRIC_REORDERING | This is being cached and shared in Linux.

- | Estimated Bandwidth | F:rmx_bandwidth | Not in the specification. It is not set before caching when a connection is closed.

- | Outbound Delay - Bandwidth Product | F:rmx_sendpipe | Not in the specification. This is used for socket buffer in FreeBSD. The value is set to 0 before caching when a connection is closed.

- | Inbound Delay - Bandwidth Product | F:rmx_recvpipe | Not in the specification. This is used for socket buffer in FreeBSD. The value is set to 0 before caching when a connection is closed.

2. IANA Considerations

This memo includes no request to IANA.

3. Security Considerations

To be added
Authors’ Addresses

Michael Welzl
University of Oslo
PO Box 1080 Blindern
Oslo N-0316
Norway

Phone: +47 22 85 24 20
Email: michawe@ifi.uio.no

Safiqul Islam
University of Oslo
PO Box 1080 Blindern
Oslo N-0316
Norway

Phone: +47 22 84 08 37
Email: safiquili@ifi.uio.no

Joe Touch
USC/ISI
4676 Admiralty Way, Marina del Rey
CA 90292-6695
USA

Phone: +1 (310) 448-9151
Email: touch@isi.edu

Jianjie You
Huawei
101 Software Avenue, Yuhua District
Nanjing 210012
China

Email: youjianjie@huawei.com