Extending the Space Available for TCP Options

draft-eddy-tcp-loo-02

Abstract

This document describes a method for increasing the space available for TCP options. Two new TCP options (LO and SLO) are detailed which reduce the limitations imposed by the TCP header’s Data Offset field. The LO option provides this extension after connection establishment, and the SLO option aids in transmission of lengthy connection initialization and configuration options.
1. Requirements Notation

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.
2. Introduction

Every TCP header contains a 4-bit Data Offset (DO) field implying the length of that segment’s TCP header. The DO field has been specified as: "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" [1]. For a TCP implementation, this means that the boundary separating TCP control data and application data is always exactly DO * 4 bytes from the beginning of the TCP header.

As a 4-bit unsigned integer, DO’s value is bounded between 0 and 15. This allows for a maximum TCP header length of 60 bytes (15 * 4 bytes). The required fields in a TCP header occupy a fixed 20 bytes, leaving 40 bytes as the maximum amount of space for use by TCP options.

While 40 bytes is a reasonable amount of space, sufficient for the concurrent use of several presently defined TCP options, there are cases where more space might be useful. For example, the SACK option [2] uses a fixed 2 bytes for kind and length fields, and requires an additional 8 bytes per SACK block. Thus, the maximum number of SACK blocks a TCP acknowledgement may carry is limited to 4 (with 6 bytes left over). Since SACK is commonly used with the Timestamp option [3], which uses 10 bytes, this further limits the number of SACK blocks that may be carried to 3. For specific scenarios involving large windows and combinations of data and acknowledgement loss, additional capacity for SACK blocks is useful [4].

Creation of new TCP options is also hindered by the lack of space left over after currently-used options are accounted for. For long options that must be present at connection-startup time, this is a particular problem, as all negotiable options need to share 40 bytes of space in a SYN segment. One way that has been used to get around this limitation is overloading the Timestamp bytes in the SYN segments [5]. There are other header fields that might be similarly overloaded (e.g. the urgent pointer), but this approach is of obviously limited utility, as it does not address the fundamental limitation imposed by the DO field, and there are a finite number of overloadable bits.

This document specifies two new TCP options, LO and SLO. The Long Options (LO) option allows two hosts to negotiate for the ability to use TCP headers longer than 60 bytes (and thus options space of greater than 40 bytes) on subsequent segments. This is accomplished by ignoring the DO field’s value and adding a 16-bit field at a fixed location in the header’s options to replace it. The format and usage of the LO option is detailed in Section 3.
Attempting to process initial SYN segments with greater than 60 bytes of TCP headers might cause errors if received by hosts that consider anything past the DO-specified boundary to be application data. For backwards compatibility reasons, the maximum length of options on a connection-initiating SYN segment remains 40. The SYN Long Options (SLO) option is used in the case where these 40 bytes are not enough space to carry the desired startup configuration options, and negotiates for later reliable delivery of the left-off options. Section 4 describes the format and usage of the SLO option.
3. The Long Options (LO) Option

A host might implement some set of TCP options allowing it to predict that greater than 40 bytes of TCP options space may be useful (for example SACK, Timestamps, alternate checksums, etc). In this case, a host MAY implement the LO option. When initiating connections through an active open, hosts implementing the LO option SHOULD place a LO option of the form shown in Figure 1 somewhere in the SYN segment’s options. The 16-bit field labelled "Header Length" should be filled in with the same value as the DO field in the required portion of the TCP header, left-padded with zeros.

```
            0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------------------------------+
|     Kind: #    |  Length = 4   |        Header Length          |
+---------------+---------------+-------------------------------+
```

TCP Long Options (LO) Option

Figure 1

Receipt of an acknowledgement covering the SYN and also containing a LO option means the LO option MUST be used as the first option on all subsequent segments, and the DO field on all subsequent segments SHOULD be set to 6. The value 6 represents the length of the required portions of the TCP header plus the LO option. The Header Length field of a LO option overrides the DO field in the fixed header, and has an identical meaning, but with 16 bits of unsigned precision rather than 4. Semantically, this still represents the offset from the beginning of the TCP header bounding the start of application data bytes. Since the LO option is found in a fixed place on all subsequent segments, it essentially becomes part of the required header, and looking up the Header Length field is of similar computational complexity to that required when the DO field is used.

Since a LO option’s Header Length field is of the same range as the IP header’s Total Length field [6], this allows TCP options to consume an entire maximum-sized IP datagram’s length (minus the IP header and required TCP header fields). No matter what size the options section of a TCP header is, it must still be appended with zero-padding to make the total header a multiple of 32 bits, per RFC 793 [1].

Listening hosts that implement the LO option, after reception of a SYN segment with the LO option present, SHOULD reply with a LO option in their SYN-ACK. The LO option is then used on all subsequent segments to override the DO field. It can be seen that in both the
normal case where one host passively opens and another actively opens, and the more rare case where two hosts simultaneously initiate active opens, the LO option’s use can be successfully negotiated.
4. The SYN Long Options (SLO) Option

If the LO option has been successfully negotiated, an active-opening host that has more bytes of initialization options than would fit in the SYN, can use the SYN Long Options (SLO) option. If a host supports the LO option, then it MUST support the SLO option.

Any option bytes transmitted using the SLO option will be treated as if they were carried on the SYN segment. Since there is no guarantee that the LO option will be successfully negotiated, the additional 36 bytes left over aside from the 4 byte LO option on a SYN segment should be filled with the most important remaining options that will fit. A host issuing a passive open, MUST NOT use the SLO option, as it can use the LO option on SYN-ACK segments if it needs to send long initialization options. The SLO option only serves the needs of an active-opening host that, for backwards compatibility reasons, could not send more than 40 bytes of options on the SYN segment.

After successful LO negotiation, if a host has any options that did not fit on the SYN, then additional data or acknowledgement segments MUST carry a SLO option until the first data byte has been acknowledged. The SLO option’s format is shown in figure Figure 2. The trailing 2 bytes hold a 16-bit unsigned count of the additional bytes that would have been in the SYN segment’s options, if they had been possible to include. This represents an offset from the end of the SLO option, to the last byte that should be considered a SYN option. The next "Additional Byte Count"-number of bytes trailing the SLO option MUST be the ones that did not fit in the SYN segment. The SLO option should always immediately follow the LO option, followed by the additional SYN options, and then by normal options, and finally application data.

Since TCP connection establishment is often concluded by a pure acknowledgement (carrying no data), only placing the SLO option and additional SYN options in such a single, unreliable segment would be risky. This is why a host MUST continue transmitting SLO options on all segments until its first byte of sent data is acknowledged. Acknowledgement of the first data-byte implicitly covers the SLO and
trailing options, as these must have been received end-to-end with the first data byte.

If a host does not send any data bytes, but if by some means (perhaps through the received options) it is possible to derive either an explicit or implicit acknowledgement of even a single option transmitted in a SLO-carrying segment (for example via a Timestamp echo), then a host MAY choose to stop transmitting the SLO data. This special case overrides the previously specified MUST condition.

A host SHOULD NOT continue sending SLO options after it has received acknowledgement of the first data byte, nor should a host process incoming SLO options other than on the first valid segment it receives that carries them.
5. Middlebox Interactions

The large number of middleboxes (firewalls, proxies, protocol scrubbers, etc) currently present in the Internet pose some difficulty for deploying new TCP options. Some firewalls may block segments that carry unknown options. For instance, if the LO option is not understood by a firewall, incoming SYNs advertising LO support may be dropped, preventing connection establishment. This is similar to the ECN blackhole problem, where certain faulty hosts and routers throw away packets with ECN bits set [7]. Some recent results indicate that for new TCP options, this may not be a significant threat, with only 0.2% of web requests failing when carrying an unknown option [8].

More problematic, are the implications of TCP connection-splitting middleboxes and protocol scrubbers that do not understand the LO option. Since such middleboxes may operate on a packet’s contents (aggregating application data between multiple segments, rewriting sequence numbers, etc), if the LO option is not understood, then there may be a mangling of the data passed to the application, as control data could end up inter-mingled with the application data. Such errors could be difficult to detect at the transport layer, and many applications might not perform their own integrity checks.
6. Comparison to Extended Segments

Another proposal that solves the same problem as the LO and SLO options is that of TCP "extended segments" [9]. The extended segments technique was proposed following the initial introduction and discussion of the LO and SLO options within the IETF's TCP Maintenance and Minor Extensions working group. The two methods solve the same problem in rather different ways, and have several minor comparative advantages and disadvantages.

The LO and SLO options are designed using the philosophy of using the TCP options space to compensate for insufficiency of the standard header. This is in keeping with the way that several currently-used options work. For example, the Window Scale option deals with the limited space in the advertised receive window field, and the Selective Acknowledgement option solves the lack of information in the cumulative acknowledgement field. Extended segments approach overloads the meaning of the standard Data Offset field, keeping its original meaning for values of 5 and greater, but redefining it for values less than 5. This is seen as acceptable since values less than 5 are currently impossible, illegal, and unusable. Extended segments avoid the need for new options by changing the way that the existing standard header is parsed.

A key advantage of the extended segments approach is that it does not increase the TCP header size, whereas the LO option adds 4 bytes of space to TCP headers. The severity or triviality of this bloat in header overhead depends entirely upon the network properties and application traffic for particular use cases.

It is also not altogether clear that extended segments will always save space in comparison to LO options. The granularity of option lengths that extended segments can support is limited to the number of unusable Data Offset values (5, 0 through 4). Currently, the extended segments proposal defines 4 fixed lengths, and one "infinite" length that means the entire segment is options, with no application data. The fixed option lengths are 48, 64, 128, and 256 bytes. If the required per-data-segment options space for some extension or combination of extensions does not map to exactly these values, then padding bytes are required. If 129 bytes of options are required on a data segment, then a length of 256 must be used, and 127 bytes of useless padding are added. The LO option has a single-byte granularity and avoids the need for all wasteful padding, aside from that mandated to make the header a perfect multiple of 4-bytes. It is possible that the overhead on a single extended segment could be more than that of several segments using the LO option.
Some networkers have found the SLO mechanism that is required for processing of long initialization options to be somewhat "ugly". Extended segments avoid this by sending long initialization options on the initial SYN and SYN-ACK segments. If the other side does not support extended segments, this adds needless confusion and delay in connection setup. The protocol dance to negotiate use of extended segments is arguably much worse than using SLO. If an extended SYN is not understood, a non-reliably transmitted RST segment signals the initiating host to retry without extended segments. Such a retry mechanism is not commonly found in existing TCP implementations. If the LO option is not understood, a SYN-ACK is still immediately generated and the connection goes on uninterrupted, without any additional retry mechanisms. Furthermore, extended SYN-ACKs may be sent in response to non-extended SYNs. This complicates the recovery procedure even more, if not understood, and goes against the way that all current negotiable TCP extensions operate (only used on SYN-ACK if advertised on SYN).

Over-zealous middleboxes are immensely troublesome for the deployment of most transport layer extensions. It is unclear whether LO and extended segments have any real difference in robustness in the presence of different types of middleboxes. Both types of segments may appear as invalid to some middleboxes, and both may be mangled if rewritten by a middlebox.
7. Security Considerations

The TCP options presented in this document open no additional vulnerabilities that we are aware of.
8. Acknowledgements

This document benefitted specifically from discussions with Josh Blanton and Shawn Ostermann concerning another proposed TCP extension. Some comments from Eddie Kohler motivated the discussion of middlebox interactions.

9 References


Author’s Address

Wesley M. Eddy
NASA GRC/Verizon FNS

EMail: weddy@grc.nasa.gov