Network Working Group
Request for Comments: 1110

A. McKenzie
BBN STC

August 1989

A Problem with the TCP Big Window Option

Status of this Memo

This memo comments on the TCP Big Window option described in RFC 1106. Distribution of this memo is unlimited.

Abstract

The TCP Big Window option discussed in RFC 1106 will not work properly in an Internet environment which has both a high bandwidth * delay product and the possibility of disordering and duplicating packets. In such networks, the window size must not be increased without a similar increase in the sequence number space. Therefore, a different approach to big windows should be taken in the Internet.

Discussion

TCP was designed to work in a packet store-and-forward environment characterized by the possibility of packet loss, packet disordering, and packet duplication. Packet loss can occur, for example, by a congested network element discarding a packet. Packet disordering can occur, for example, by packets of a TCP connection being arbitrarily transmitted partially over a low bandwidth terrestrial path and partially over a high bandwidth satellite path. Packet duplication can occur, for example, when two directly-connected network elements use a reliable link protocol and the link goes down after the receiver correctly receives a packet but before the transmitter receives an acknowledgement for the packet; the transmitter and receiver now each take responsibility for attempting to deliver the same packet to its ultimate destination.

TCP has the task of recreating at the destination an exact copy of the data stream generated at the source, in the same order and with no gaps or duplicates. The mechanism used to accomplish this task is to assign a "unique" sequence number to each byte of data at its source, and to sort the bytes at the destination according to the sequence number. The sorting operation corrects any disordering. An acknowledgement, timeout, and retransmission scheme corrects for data loss. The uniqueness of the sequence number corrects for data duplication.

As a practical matter, however, the sequence number is not unique; it

is contained in a 32-bit field and therefore "wraps around" after the transmission of 2**32 bytes of data. Two additional mechanisms are used to insure the effective uniqueness of sequence numbers; these are the TCP transmission window and bounds on packet lifetime within the Internet, including the IP Time-to-Live (TTL). The transmission window specifies the maximum number of bytes which may be sent by the source in one source-destination roundtrip time. Since the TCP transmission window is specified by 16 bits, which is 1/65536 of the sequence number space, a sequence number will not be reused (used to number another byte) for 65,536 roundtrip times. So long as the combination of gateway action on the IP TTL and holding times within the individual networks which interconnect the gateways do not allow a packet's lifetime to exceed 65,536 roundtrip times, each sequence number is effectively unique. It was believed by the TCP designers that the networks and gateways forming the internet would meet this constraint, and such has been the case.

The proposed TCP Big Window option, as described in RFC 1106, expands the size of the window specification to 30 bits, while leaving the sequence number space unchanged. Thus, a sequence number can be reused after 4 roundtrip times. Further, the Nak option allows a packet to be retransmitted (i.e., potentially duplicated) by the source after only one roundtrip time. Thus, if a packet becomes "lost" in the Internet for only about 5 roundtrip times it may be delivered when its sequence number again lies within the window, albeit a later cycle of the window. In this case, TCP will not necessarily recreate at the destination an exact copy of the data stream generated at the source; it may replace some data with earlier data.

Of course, the problem described above results from the storage of the "lost" packet within the net, and its subsequent out-of-order delivery. RFC 1106 seems to describe use of the proposed options in an isolated satellite network. We may hypothesize that this network is memoryless, and thus cannot deliver packets out of order; it either delivers a packet in order or loses it. If this is the case, then there is no problem with the proposed options. The Internet, however, can deliver packets out of order, and this will likely continue to be true even if gigabit links become part of the Internet. Therefore, the approach described in RFC 1106 cannot be adopted for general Internet use.

Author's Address

Alex McKenzie
Bolt Beranek and Newman Inc.
10 Moulton Street
Cambridge, MA 02238

Phone: (617) 873-2962

EMail: MCKENZIE@BBN.COM