A Survey of TCP Performance on Wireless Links

Brooke Shrader

March 11, 2002
Royal Institute of Technology (KTH), Stockholm, Sweden

Abstract

This report is a survey of both the problems that arise when TCP is used over wireless links and the proposed solutions to these problems. The problem is first characterized by the traits of wireless links and the aspects of TCP operation that affect performance. Also, the characteristics of different types of wireless networks are described as they determine the nature and severity of problems when TCP is used. Measurements taken from previous studies are given to provide quantitative examples of performance degradation. Finally, this document explores some of the proposed solutions to the problems and their effectiveness.

I. Introduction

The development of mobile computing relies on past work and standards used for wired networks, including the TCP/IP protocol suite. The definition of Mobile IP found in [1] was the first step in development of the Wireless Internet, but many problems have yet to be solved. One significant problem that has emerged and received attention in research is the performance of the Transmission Control Protocol (TCP) over a mobile, wireless link. As TCP was developed for wired links, its use in wireless links results in unanticipated problems due to the interaction between the transport and link layers. Bit errors and handoffs in the wireless link are interpreted by TCP as congestion, and the subsequent actions taken to alleviate congestion result in suboptimal performance. In Section II, this problem is further characterized while Section III deals with the proposals made to resolve the problem.

II. Problem characterization

A. Traits of wireless links

The wireless media differs tremendously from the wired media in various aspects. In general, the three key problems in wireless transmission are the path loss, ambient noise, and the sharing of the radio spectrum [2]. The path loss and ambient noise translate to higher bit error rates (BER) for a wireless link as compared to a wired link. As an example, typical values for bit error rates on a wireless channel are 10^{-6} or worse, while a fiber-optic link might have error rates of 10^{-12} or better [3]. Also, wireless links are typically employed to support mobility, and this mobility in turn causes additional degradation to the radio channel. The mobility itself causes a phenomenon referred to as fading [2], whereby the channel state can vary rapidly and sporadically, thus increasing the bit error rate. Additionally, supporting mobility while maintaining a low transmission power and sharing of the limited radio bandwidth translates to the need for a cellular approach to wireless communication. The coverage area is thus divided into smaller areas, or cells, between which resources are shared. When crossing the boundary of a cell, a handoff of the mobile user to another cell and set of resources is required. This handoff inhibits transmission, causing delays and loss of transmitted data if connectivity is lost. In general, when compared to a wired link, a wireless link is more unreliable, with rapidly changing characteristics, and intermittent connectivity. This nature of the wireless channel means that some degradation to performance is unavoidable. However, TCP's response to the wireless channel causes unnecessary additional degradation.

B. TCP's response to lost packets

TCP's response to lost packets is tuned to work well for congestion in wired networks, but in wireless networks, it causes sub-optimum performance. Some characteristics of TCP and its response are detailed here as a means

of explaining this sub-optimum performance and for reference in later exploring possible solutions. Below is a summary of TCP's response taken from the more detailed description in [4].

A sent TCP packet is determined to have been lost either if no acknowledgement (ACK) is received within the retransmission timeout (RTO) period or if multiple duplicate ACKs arrive for the packet sent prior to the one that was lost. The RTO is based on measurements of the round-trip delay time (RTT) for packets to travel over the link. These RTT measurements are collected and the RTO is set to the sum of the smoothed RTT (or approximate average) and four times its mean deviation. A reasonable RTO is crucial to effective utilization of resources. If the RTO is excessive, retransmission will be unnecessarily delayed, resulting in slow recovery of network operation. If the RTO is too short, unnecessary retransmissions will occur and effective throughput will be decreased.

When a lost packet is determined by expiration of the RTO, TCP initiates an exponential backoff of the RTO and enters the slow start and congestion avoidance mode. The exponential backoff of the RTO involves doubling its value with every expiration after the packet has been retransmitted. Then measures are taken to reduce the packet transmission rate so that congestion can be avoided. Slow start involves setting the congestion window, which indicates the number of packets that can be sent without causing congestion, to one packet. With each ACK of a received packet, the congestion window is exponentially increased. When the congestion window reaches a threshold value corresponding to half its value when the loss was determined, congestion avoidance takes over. In this phase, the congestion window is increased only linearly. When considering transmission over a wireless link, it is important to note here that multiple lost packets will cause the slow start threshold to be repeatedly reduced, and thus the congestion avoidance mode will dominate and the packet transmission rate will grow very slowly. This can lead to degradations in throughput.

On the other hand, if a packet is determined to be lost by the reception of duplicate ACKs, TCP begins fast retransmit and fast recovery. In this case, TCP is responding to what it believes is not congestion, so it does not wait for the RTO to expire before retransmitting the packet. The fast recovery algorithm then involves skipping over the slow start phase to avoid excessively decreasing the transmission rate. Instead, the congestion window is halved and congestion avoidance mode begins. Fast retransmit and fast recovery were introduced as an alternative to the traditional congestion control described above, and they are helpful in retaining good performance after losses occur on a wireless link.

C. Network scenarios

Different types of wireless networks exhibit different characteristics such as bandwidth and cell coverage area. These characteristics determine RTTs and frequency of handoffs, which effects how and to what extent TCP performance is degraded over the wireless link. A few types of wireless networks are briefly described below, followed by a description of how these traits effect TCP performance.

The focus of most studies of TCP over wireless links is for current wireless local area network (WLAN) and wide area network (WAN) environments. WLAN cell coverage radii are usually on the order of tens of meters, with bit rates in the range of 2 to 54 Mbps. Resulting RTTs are typically a few milliseconds. For example, RTTs of 1ms are measured in [3] and 3.5ms in [5]. An example frame error rate (FER) is given as 1.5% for a link over 85 feet with 1400 byte frames [6]. On the other hand, WANs cover larger areas at lower bit rates and thus higher RTTs. Cell coverage radii range in hundreds of meters, while bit rates of today's systems are tens of kilobits per second. For example, the Global System for Mobile Communications operates at a bit rate of 9.6kbps. Typical RTTs are then hundreds of milliseconds and are more widely varying than those for WLANs. In one case, the RTT on a GSM network averaged 600ms with a standard deviation of 20ms [6]. The FER is similar to that in WLAN although frames are shorter on average. Also, frame errors tend to be less bursty than bit errors due to the use of interleaving in most systems.

Wireless Internet can also be implemented over satellite links, whose characteristics differ greatly from those of WLAN and WAN systems. Two different types of satellite systems display different characteristics. Low earth orbit (LEO) satellites, usually located between 500 and 2000km above the surface of the earth, orbit the earth such that their motion requires handoff between satellites for a connection to a user on the ground. The area covered by a LEO ranges from 3000 to 4000km in radius, and handoffs may occur every few minutes [7]. Typical RTTs are

cited as 6-70ms in [7]. On the other hand, geostationary earth orbit (GEO) satellites, located at altitudes of roughly 40,000km above the earth, remain stationary relative to points on the earth. Thus, handoffs are not necessary for GEO satellite, which should alleviate problems associated with TCP use. However, GEO RTTs are much greater than those for LEO, and for this reason, it is concluded in [7] that TCP performs better over LEO constellations.

Given the traits of these various types of wireless systems, it is important to note how TCP performance will vary over the wireless links [6]. Links with higher delay times will be more severely affected by packet losses since TCP must maintain larger transmission windows in order to keep the data flowing. Also, higher speed links are more vulnerable to packet losses since it takes longer for TCP to reach its peak throughput after a packet is lost. Finally, as earlier stated, smaller coverage areas can translate to more frequent handoffs, which cause many problems with TCP.

In addition, the trend in wireless networks is towards hierarchical structures, whereby different types of networks are overlaid. For example, a WLAN microcell located within a WAN macrocell should in the future allow for seamless handoff between the two systems. These constitute a different type of handoff than those within a single system, and cause additional problems with TCP.

D. Measured performance

An early comprehensive study aimed at measuring TCP's performance in wireless links was performed in [3]. This study focuses on the effect of crossing cell boundaries and makes measurements for a WLAN scenario. Three different handoff situations are examined and the throughput is found to decrease compared to situations without handoffs. Additionally, measured adverse effects include pauses in communication, loss of packets, and slow recovery. Some details of results found in [3] are described below.

The reduction in throughput found in [3] depends on the nature of the handoff. For handoffs between overlapping cells, the mobile never loses connectivity, and the throughput decreases by only 6%. This loss in throughput is attributed to encapsulation and delays due to forwarding after handoff. However, when cell boundaries are non-overlapping, the degradation is more pronounced. In this case, packets are lost while routing tables are being updated. Also, waiting periods for excessively high RTOs are found to cause long pauses in transmission, which is determined to be the main cause of the decrease in throughput. For a handoff occurring at the instant the mobile crosses the cell border, throughput decreases by 12%. When there is a lag time at handoff and the mobile loses connectivity for some period of time, the throughput is measured to decrease by 31%.

Other measurements taken in [3] involve the long pauses occurring after handoff. For TCP implementations without fast retransmission and recovery, pauses are measured to last 800ms after a handoff from non-overlapping cells. The benefit of implementing fast retransmissions is indicated by a reduction of these pauses to 200-300ms. It is noted that as a result of the exponential backoff of the RTO, the pauses grow exponentially with increasing lag time during handoff. In extreme cases, these pauses are measured to last several seconds after the mobile user enters a new cell and handoff is completed.

Other results point out the difference in reduction to throughput due to wireless errors for WLANs and WANs. A single WLAN link is cited to suffer a throughput reduction of 47%, while a WAN path (involving a WLAN and 15 wired links) suffers only a 23% reduction [6].

Measurements of TCP performance in hierarchical systems provide interesting results in [5]. First, handoffs between a WLAN and a GSM system are observed. The smaller value of the RTO before the handoff cannot be adjusted quickly enough after the sudden and substantial increase in the RTT after handoff. This causes spurious TCP timeouts and unnecessary retransmissions. TCP's recovery time is measured to be up to 40 seconds, which is substantially worse than found in previous studies. Also, handoffs from a GSM system to a WLAN are observed and the additional bandwidth available after the handoff is found to be inefficiently used. Even after the handoff, packets in the buffer on the link layer continue to flow over the lower-bandwidth GSM system, thus wasting the high bandwidths made available by the WLAN system. The handoff itself is not found to cause timeouts or unnecessary retransmissions, but if packets are lost during or just after the handoff, the excessive RTO causes needed retransmissions to be delayed.

III. Proposed solutions

Many solutions have been put forth for improving TCP performance over wireless links. This section describes a number of these proposed solutions, though the possibilities described here are certainly not exhaustive. As the problem is caused by poor interaction between the link and transport layers, most solutions are suggested for either one of these two layers. Link layer solutions work with the aim of improving link characteristics or hiding non-congestion-related losses from the transport layer. Transport layer solutions instead attempt to adapt the TCP protocol to make it aware and respond appropriately to losses that are not related to congestion. In many cases, a cross-layer approach is taken where one layer must be cognizant of characteristics of another layer. Additionally, a split-connection approach is described for application to a network that connects to a wired infrastructure [8].

A. Link layer solutions

When the adverse effect of crossing cell boundaries was measured in [3], a reasonable initial proposal was to require that all handoffs be soft, or implement 'make then break' connections. This would eliminate loss of packets during cell crossings that cause pauses in communication. However, characteristics of wireless networks, including scarcity of bandwidth and the need for low-power solutions as well as accurate location information, make it advantageous to avoid overlap between cells. It is then noted that in the near future, it is unlikely that all cellular networks will be able to ensure soft handoffs.

Performance on the link itself could be improved by altering link-layer protocols. These protocols involve forward error correction (FEC), retransmission in response to automatic repeat request (ARQ) messages, or a hybrid of the two. Link-layer protocols traditionally work independently of the transport layer, and this can directly cause problems for TCP. Some ARQ schemes retransmit packets out of order, which can cause duplicate TCP ACKs, unnecessary invocation of fast retransmissions, and thus degraded throughput [8]. To resolve this problem, it is suggested in [8] that the link-layer protocols be given knowledge of TCP. This could allow the link-layer to block duplicate ACKs from TCP and avoid retransmissions being initiated by both layers. A TCP-aware link-layer protocol investigated in [8] is shown to give 10-30% higher throughput than one that works without knowledge of TCP.

Another popular link-layer solution that uses knowledge of TCP is the Snoop protocol [9], which works to hide losses over the wireless link from TCP. At a base station, a 'snoop agent' keeps a cache of transmitted TCP segments that have not been acknowledged by the mobile user. When a link-local timeout or duplicate acknowledgements indicate a packet loss over the link, the snoop agent accesses its cache and resends the packet without notifying TCP. Thus TCP is shielded from duplicate ACKs, which avoids unnecessary fast retransmissions and initiation of congestion avoidance.

B. Transport layer solutions

One modification to TCP that would improve performance over wireless links is the use of Selective Acknowledgement (SACK). The use of cumulative ACKs in traditional TCP result in poor performance when multiple packets are lost during one transmission window. The cumulative ACKs do not provide information quickly enough to allow for fast recovery. This is a particular problem in satellite networks with large RTTs [7]. The SACK proposal [10] would modify ACKs to contain segment sequence numbers. This could allow for lost segments to be quickly identified and resent within a single RTT. In [8] it is noted that SACK implementation is particularly useful in bursty error channels. Performance improvements due to SACK are shown to be significant, though not as great as improvements provided by other link-layer solutions [8].

Explicit loss notification (ELN) is a solution aimed at providing the sender with information to distinguish between losses due to congestion and those due to errors on the wireless link. This is achieved by marking cumulative ACKs to identify a loss on the wireless link that is not related to congestion. Thus, the sender can retransmit segments without initiating congestion-control mechanisms that would unnecessarily reduce throughput. A drawback of ELN is that it may be difficult to identify which packet losses are due to errors on the wireless link. One ELN scheme simulated in [8] is shown to improve throughput by a factor of two over a scheme without ELN.

C. Split connection solutions

As wireless networks usually connect to a fixed, wired infrastructure, a split-connection approach can be used whereby distinct connections are set up for the wired and wireless links. A TCP session could be used on each link, or a specialized transport protocol that is suited to a wireless environment could be used for the wireless link. Transport over the wireless link could then be improved, while separating the congestion control between the different types of links could provide additional improvements. However, there are a number of drawbacks to a split-connection approach. First, an acknowledgement to the sender over a wired link may be received even before a mobile user receives the packet. These acknowledgements should be delayed, thus decreasing throughput. Also, at the point connecting the wired and wireless links, TCP protocol processing must be performed twice and information on the state of both connections must be maintained. This translates to additional overhead and slower handoffs. Studies in [8] compared split connection solutions to link and transport layer solutions, and concluded that good performance does not require splitting the connection.

IV. Conclusions

This report is a survey of the issue of TCP performance over wireless links. The problem is characterized by describing the fundamentals of wireless links and TCP's response to lost packets. The effects on different types of wireless networks are noted, and measured performance results from past studies are given. Some proposed solutions and their effectiveness are then explored and compared. This survey is certainly not an exhaustive coverage of the issue, but instead provides an introduction to a hot topic of research. Research and advancements proceed, while the problems with the performance of TCP over wireless links demonstrate the need to work across layers in the development of the wireless Internet.

References

[1] C.E.Perkins, ed., "IP Mobility Support", *IETF RFC 2002*, 1996.

[2] L.Ahlin and J.Zander, *Principles of Wireless Communications*, Studentlitteratur, Lund, 1997.

[3] R.Caceres and L.Iftode, "Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments", *IEEE Journal on Selected Areas in Communications*, June 1995, vol.13 no.5, pp.850-857.

[4] W.R.Stevens, *TCP/IP Illustrated, Volume 1: The Protocols*, Addison-Wesley Longman, Reading, Massachusetts, 1994.

[5] M.Ronquist, "TCP Reaction to Rapid Changes of the Link Characteristics due to Handover in a Mobile Environment", Master thesis report, Royal Institute of Technology (KTH), Stockholm, Aug 1999.

[6] G.Xylomenos, G.C.Polyzos, P.Mahonen, and M.Saaranen, "TCP Performance Issues over Wireless Links", *IEEE Communications Magazine*, April 2001.

[7] Y.Chotikapong, H.Cruickshank, and Z.Sun, "Evaluation of TCP and Internet Traffic via Low Earth Orbit Satellites", *IEEE Personal Communications*, June 2001.

[8] H.Balakrishnan, V.N.Padmanabhan, S.Seshan, and R.H.Katz, "A Comparison of Mechanisms for Improving TCP Performance over Wireless Links", *IEEE/ACM Transactions on Networking*, Dec 1997, vol.5 no.6, pp.756-769.

[9] H.Balakrishnan, S.Seshan, and R.H.Katz, "Improving reliable transport and handoff performance in cellular wireless networks", *ACM Wireless Networks*, vol.1, Dec 1995.

[10] M.Mathis, J.Mahdavi, S.Floyd, and A.Romanow, "Selective acknowledgement options", *IETF RFC 2018*, 1996.