Techniques for Improving the Performance of TCP in CDMA Cellular Networks

Biswajit Ghose and Vikas Jain
Department of Computer Science and Engineering
University of California, San Diego
{bghose, vjain}@cs.ucsd.edu

Abstract

In this paper we survey various techniques for improving the performance of TCP in CDMA cellular networks. The two major issues that affect the performance of TCP in cellular networks are high packet loss rates due to physical layer impairments in the radio environment, and intermittent connectivity losses due to handoffs. TCP has been designed and tuned to perform well in traditional wired networks and is not well suited to handle either of the above two issues.

We classify the various techniques surveyed in our paper into three categories depending upon the protocol layer (physical, link or transport) at which they are predominantly implemented. Our survey recognizes that improvements in lower layers have an impact on the performance of the higher layers, and that performance improvement of TCP does not necessarily involve a lot of changes in the TCP layer itself.

1. Introduction

Wireless access is poised to overtake wired access by the end of year 2003, with high-speed wireless data being a dominant service. This poses significant challenges for the de facto transport layer protocol, namely TCP, to support high data throughput over the inherently lossy wireless medium.

The performance of TCP is poor over wireless links due to its assumption that packet losses are always due to congestion and the consequent reduction of its congestion window in response. Furthermore, mobile host movements in a cellular environment result in handoffs, which cause disconnections and degrade TCP throughput due to long communication pauses, slow post-handoff recovery, and successive retransmission timeouts [5], [7].

Balakrishnan et al. [1] present a host of approaches for improving TCP performance over wireless links. The paper is remarkable for presenting details and comparative analysis on these approaches. However, the paper deals with a wireless LAN environment, and the issues of fading and handoffs, typical of a cellular wireless environment and their impact on TCP performance have not been included in their study. In our survey, we explore the various techniques for improving the performance of TCP in CDMA cellular networks.

The remainder of this paper is organized as follows. Section 2 provides a brief background on aspects of CDMA cellular environment pertinent to the subject of the paper. Section 3 sets the tone for the following three sections by suggesting that performance improvements in TCP can be brought about by changes in any of the physical, link or transport layers. Section 4, 5 and 6 present physical layer, link layer and transport layer techniques respectively, which help in improving the performance of TCP in cellular wireless systems. Section 7 concludes the paper and highlights a few areas for future research.

2. Background on CDMA Cellular Environment

CDMA [11], [12] is a spread-spectrum multiple access technique in which the mobile users occupy the same time and frequency allocations and are distinguished by unique orthogonal sequences or codes. CDMA allows a frequency reuse factor of 1, which means that every cell can use the same carrier frequency for its operation. The use of the same frequency across cells allows CDMA to support soft handoff, which is a novelty of CDMA systems and is not supported by the other cellular technologies. Soft handoff employs a make-before-break strategy while handing off the calls from one cell to another. A necessary condition for soft handoff is that the cells traversed by the user must belong to the same base station (BS). However, in the cases where the user moves across cells belonging to different base stations or makes a transition from CDMA network to an analog network (e.g. AMPS), a hard handoff (break-before-make) becomes necessary. The major benefit of soft handoff is that the user never suffers a disconnection during the process of being handed over to a new cell.

Figure 1 shows a partial cellular network layout in which cells A, B, C and D are controlled by BS1 and cells E and F are controlled by BS2. Any movement within the cells of BS1 will be a soft handoff in CDMA and only if the mobile moves into cell E or F will a hard handoff be necessary.

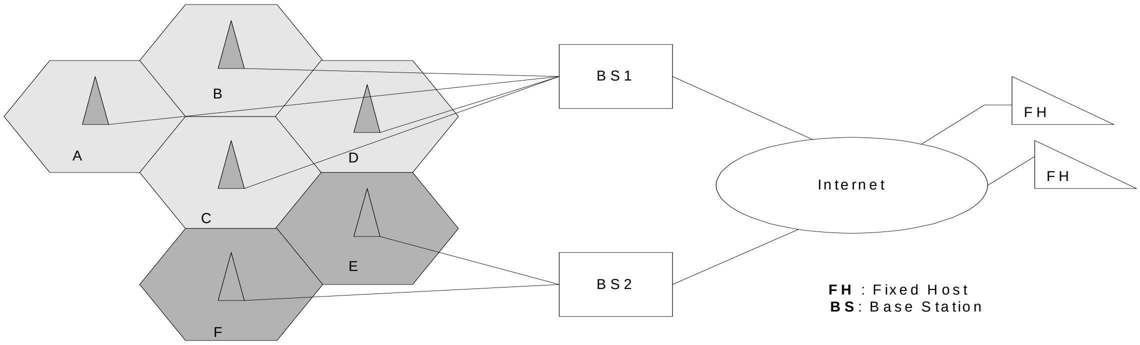

Figure 1. CDMA Cellular Network

Typical cell radius for CDMA cells is 2-5 miles, and a single base station can control hundreds of cells. Soft handoff is therefore the common case in CDMA cellular environments and accounts for 80-85% of the handoffs[1], while hard handoff is relatively less frequent.

CDMA systems use convolutional coding and also Turbo coding [13] for forward error correction (FEC) on the links between the base station and the mobile to achieve an acceptable bit error rate (BER) with lesser power. This is due to a coding gain [17], as depicted in Figure 2, which shows that Turbo coded signals require lesser power than convolutional coded signals (which require lesser power than uncoded signals) to achieve the same bit error probability. The ability of Turbo

---

[1] Here we consider the handoffs in the traffic state only, i.e., the handoffs that take place when the calls are in progress. All handoffs that take place in the idle state are hard handoffs, but that is not relevant to our discussion.

codes to achieve an acceptable BER at lesser signal power helps in supporting a larger number of users or equivalently a smaller number of users at higher data rates.

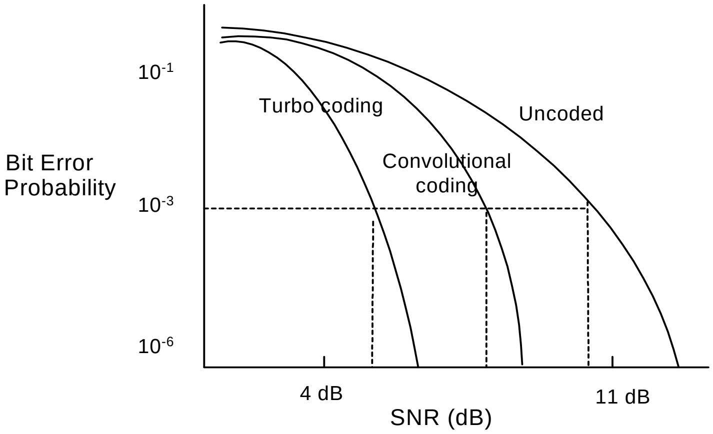

Figure 2: Coding gain due to convolutional and Turbo codes

In a mobile radio environment, the signal received by the mobile is subjected to physical layer impairments caused by path loss and multipath (Rayleigh) fading [17]. Path loss essentially characterizes the loss of signal strength as a function of distance from the transmitter. The signals are also reflected and scattered by the obstacles in the path, and arrive at the receiver through multiple paths and are thus offset in time. Since these time offsets are random, the sum total of their contribution results in rapid fluctuations over very short travel distances (a few wavelengths) or short time durations (of the order of seconds). This effect is called multipath fading and is characterized by the Rayleigh fading distribution. The multipath fading causes burst errors in the wireless link. The fade durations are comparable to the transmission times of TCP packets and thus the packet losses are correlated with one another. This is in contrast to i.i.d. (independent and identically distributed) losses, which are assumed to be independent of other losses.

3. Performance of TCP in Cellular Systems

The performance metric of TCP in cellular environment is the average throughput, which is the same as in wired networks. However, the average throughput in the cellular case is not only dependent on the congestion in the network, but also on factors like bit error rate of the wireless medium, the cell handoff time and the cell resident time [2], [7].

Figure 3 below shows the protocol layers present at the various nodes in the network including the stationary host and the mobile host. The degradation in the TCP performance as mentioned in the Section 1 is mainly due to the lossy nature of the wireless link (between the mobile host and the base station) and also due to the intermittent disconnections caused by motion of the mobile host across cells. The physical layer between the BS and the mobile host can be improved to lower the packet loss rates and thus reduce the degradation of the TCP throughput. The link layer protocol that operates on the one-hop wireless link (between the BS and the mobile host) can also be modified to hide the wireless transmission losses using local recovery mechanisms (using link layer retransmission). A direct way to improve the performance of TCP will be to modify TCP itself, since it is the inherent assumptions of TCP that are the cause of its poor performance in the wireless environment [9].

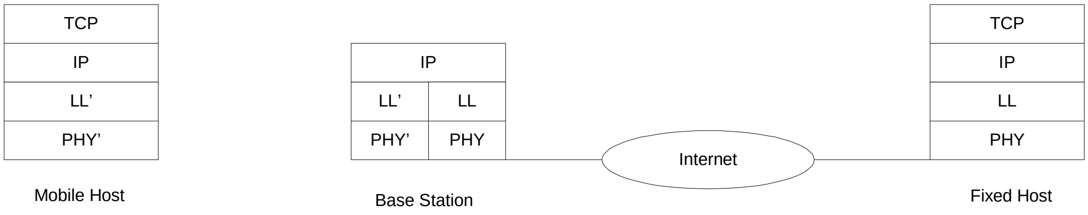

Figure 3. Protocol Stack Diagram

We classify the techniques surveyed in the paper depending upon the layer at which they are predominantly implemented, namely the physical, link or transport layer. We see that although IP layer is critical for providing inter-network mobility, it requires very little or no changes in directly effecting the performance of TCP. It will however be evident in some of the surveyed techniques that IP can aid the TCP layer in providing an awareness of host mobility, which proves useful in improving the performance. This upward propagation of information through the protocol stack is termed as vertical dependence [8] and we shall see that improvement in the performance of TCP can be effected by changes in physical, link, or transport layer of the protocol stack, in isolation or with cooperation between layers.

4. Physical Layer Techniques

In this section, we present the Forward Error Correction and Interleaving techniques, which are used at the physical layer in CDMA systems.

4.1 Forward Error Correction (FEC) and Interleaving

In their paper [2], Chockalingam et al. use a first-order two state Markov process as an analytical model for the correlated Rayleigh fading wireless radio channel. The normalized Doppler bandwidth f_dT is used to describe the correlation in the fading process. A low value of $f_dT = 0.01$ corresponds to a pedestrian user moving at 1 mi/hr, whereas a high value of $f_dT=0.64$ corresponds to a vehicular user moving at 60 mi/hr. For small values of f_dT, the fading process is very correlated (long bursts of packet errors), and for large values, successive samples of the channel are almost independent (i.i.d. errors). Table-1 in [2] clearly illustrates that as the packet error rate decreases, the length of burst errors also decreases for any given f_dT. E.g., for a f_dT value of 0.01, a packet error rate of 0.1 leads to an average burst error length of 13, and a packet error rate of 0.01 leads to an average burst error length of only 4.

The paper presents a result that for a given value of the packet error rate, TCP NewReno performs better in the case of correlated errors than i.i.d. errors. The reason for this observation is that for a given packet error rate, correlated errors correspond to fewer error events than i.i.d. errors, which cause the congestion window to be shrunk less frequently. A necessary condition for the above result to hold is that the number of errors due to channel error burstiness should be much less in comparison to the congestion window size, so as to allow sufficient number of dup acks to be generated to trigger fast retransmit. In the case of a small congestion window size, if the congestion window is exhausted before sufficient number of dup acks are generated, then the TCP sender stops and waits for the retransmission timer. A larger congestion window is therefore desirable, although the experiments show that increasing the window size above a threshold does not offer much additional improvement. By keeping the receiver buffer size in excess of the average packet error burst length, TCP NewReno is seen to offer improved throughput in the case of correlated errors as compared to i.i.d. errors. As mentioned before, the errors tend to be correlated for pedestrian users and i.i.d. for vehicular users, and thus the approach yields more benefit for pedestrian users.

The paper presents simulation results that highlight the fact that there is practically no benefit on the TCP throughput of decreasing the fast retransmit threshold from 3 to 1, and thus enhancements like TCP Vegas are unlikely to lead to much improvement in performance in the presence of correlated errors.

Discussion

The physical layer in CDMA uses convolutional and Turbo coding between the base station and mobile (in both directions) to achieve high coding gain. This helps in achieving a lower value of the packet error rate, and as per Table-1 in [2], this leads to a reduction in the burst errors, which is highly desirable.

The FEC process is followed by the interleaving process, which provides time diversity as a further safeguard against burst errors. The basic function of the interleaver is to distribute one long burst of errors into many smaller bursts of correlated errors, spread across many different physical layer frames. This prevents the concentration of errors in any particular TCP window, which allows TCP fast retransmit mechanisms to offer better throughput, as shown in the simulations in [2].

There is yet another advantage that the physical layer in CDMA provides, which we mention briefly below, but a detailed discussion of this aspect is beyond the scope of the paper. CDMA, owing to its wide-band signal (1.25 Mhz), is able to discriminate between the various multipath arrivals [11]. The use of rake receivers (essentially a set of correlators) at both the base station and the mobile allows the combination of multipaths to form a strong signal. Thus, CDMA uses multipath arrivals to its advantage in contrast to other wireless technologies, and hence suffers significantly lesser burst errors caused by multipath fading.

5. Link Layer Techniques

In this section, we present two link layer techniques: Section 5.1 presents the Delayed Duplicate Acknowledgments scheme and Section 5.2 presents the Radio Link Protocol scheme.

5.1 Delayed Duplicate Acknowledgments Scheme

Delayed duplicate acknowledgment scheme [3] presents a direct contrast to the LL-TCP-AWARE [1] scheme as it de-couples the link layer from the need to understand the semantics of TCP protocol. In this scheme, the link layer entities at the ends of wireless link perform limited number of re-transmissions to provide a localized recovery of packet losses. The link layer does not aim to provide completely reliable delivery, and it is left to the TCP layer in case the link-level retransmissions are unable to recover the packet loss. This scheme takes care of any possible interference between the TCP level retransmissions and link layer retransmissions by delaying the third and subsequent dupacks for an experimentally determined interval “d”. The value of this delay (d) is adjusted so as to allow sufficient number of re-transmissions at the link layer before the TCP layer at the receiver signals the sender to perform retransmission of the lost packets by sending triplicate dupacks.

The experimental results for this scheme are obtained from ns-2 simulations (using TCP Reno) for a 2 Mbps wireless LAN environment and for two values for delay (1ms and 20ms) on the wireless link. Exponential error model is chosen for modeling the errors on wireless link and error rates in the range of 10^{-4} to 10^{-5} are considered.

The results for wireless delay of 1ms with congestion loss rates of the order of 5% highlight an extremely interesting point that this delayed dupack scheme for d = 0 (which is equivalent to the unchanged TCP with link layer retransmission) provides a comparable performance to the LL-TCP-AWARE scheme. This is mainly due to the fact that low bandwidth-delay product links may not allow enough in-flight segments for the receiver to generate three duplicate acknowledgments and that link level retransmission may be able to recover the lost packet earlier. We feel that this experimental result provides a strong case for arguing minimal changes in the TCP to counter the errors during transmission for wireless links with low bandwidth-delay product. However, it should not be forgotten that packets on slow links also incurs more transmission delays, which may cause the sender to timeout initiating a retransmission.

It is shown through simulations that with 20ms of delay in wireless link and the value of 80ms for the parameter “d”, the delayed dupack scheme gives a TCP throughput comparable to the LL-TCP-AWARE scheme. The value of 80ms (being more than one round trip time) allows at least one retransmission, which results in appreciable performance improvement, as the chances of the retransmission being lost again are very low probabilistically.

The idea of using localized link level recovery in conjunction with delayed third dupack goes well with the end-to-end philosophy and also does not violate any layering principle. LL-TCP-AWARE [1] scheme, besides violating layering principle, also may not work well in direction towards the fixed host as it may take up to one round trip delay before the TCP acknowledgments arrive at BS in which time the retransmission timer at the wireless host may have already expired resulting in wasteful re-transmissions of data (which would also be done by the link layer at BS). It may therefore prove prudent to handle the local losses (wireless, in this case) using local link level recovery mechanisms and avoid any inter layer interactions which also violate the spirit of layering.

Discussion

This scheme does not perform well for the case when the errors are mainly due to congestion, as the third duplicate acknowledgment is delayed unnecessarily, preventing the sender from reacting to the congestion.

The third generation CDMA systems [15], [16] will be able to provide data rates of up to 2 Mbps per user, which is an appreciable improvement over the current day CDMA systems that can only provide up to 9.6kbps per user. An outcome of the increased data rates is that the transmission delays at BS are lowered which reduces the probability of interference between the TCP level and link level retransmissions significantly and thus makes the link layer mechanisms more viable. We feel that there is a strong need to further explore the link layer retransmission techniques in light of high data rates that will become available in the next generation (3G) CDMA systems.

5.2 Radio Link Protocol (RLP) And LTU Subdivision scheme

The cdma2000 protocol stack is illustrated in Figure 4. The MAC layer in CDMA [14] implements a negative acknowledgment based Radio Link Protocol (RLP), which provides limited retransmission of erroneous or lost frames [12]. LTU stands for Logical Transmission Unit, and the concept of LTU is explained later in the section.

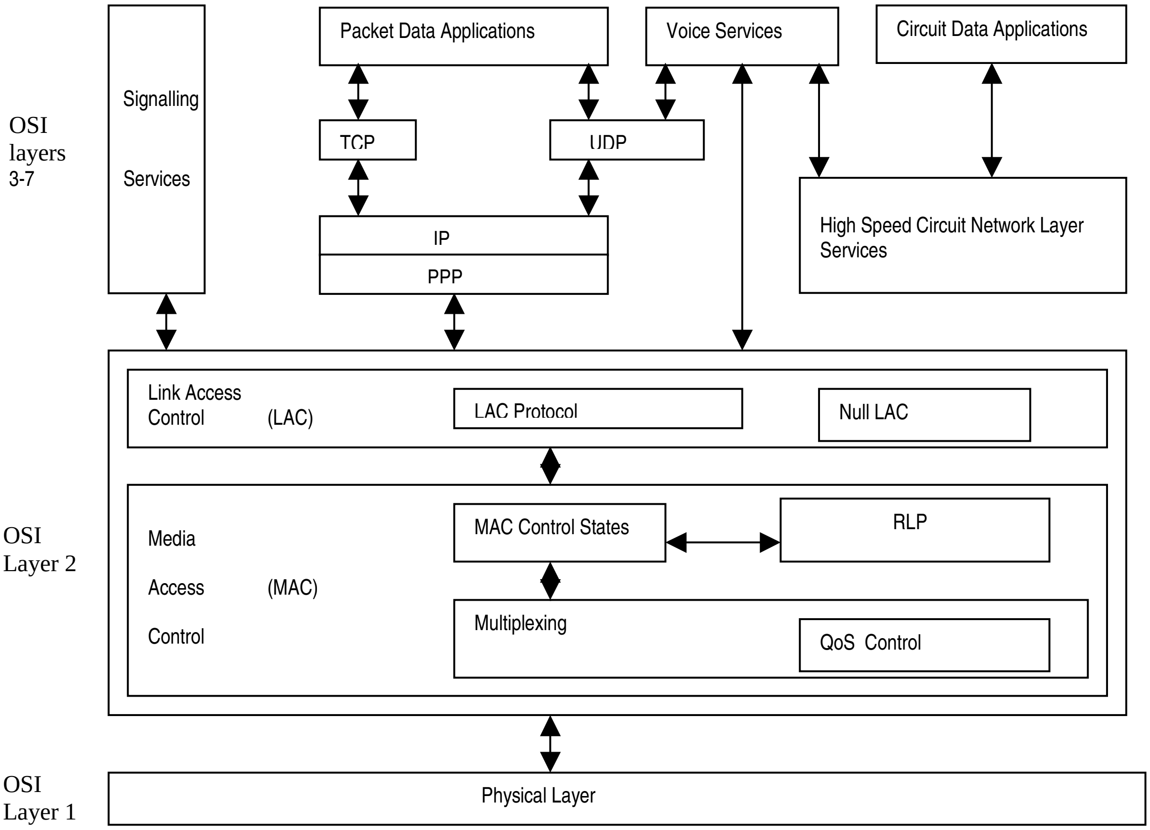

Figure 4: cdma2000 system protocol stack [4]

CDMA link layer (MAC) techniques are examined in a paper by Bai et al [4], where they focus on the following aspects of TCP performance in a CDMA network: 1) evaluation of system performance with transparent and non-transparent RLP mode and 2) performance difference with and without LTU subdivision. Both RLP and LTU subdivision are supported at the MAC layer in CDMA.

Transparent RLP does not provide any retransmission of missing data frames, whereas non-transparent RLP uses a NAK based selective ARQ protocol to retransmit lost data frames. The result of the simulation presented in the paper indicates that non-transparent RLP can withstand high FER (in the region of 10^{-1} or even more) without substantial degradation in the normalized throughput, whereas the performance of transparent RLP decays very fast at relatively small FER of 10^{-3}. The simulation experiments take into consideration a range of mobile speeds, and also highlight the effectiveness of non-transparent RLP in improving TCP performance at different mobile speeds, in addition to being able to withstand high FER rates.

The RLP has to maintain the integrity of large amount of data bits within 20 ms physical layer frames. The number of bits in the frame increases with increasing data rate. For example, corresponding to a data rate of 9.6 Kbps, the number of bits in the frame is 172, whereas for a data rate of 1 Mbps, the number of bits is 20712. The MAC layer standard in CDMA permits smaller capsules of information called Logical Transmission Units (LTUs) to be defined. A number of LTUs may be supported per physical layer frame, with the standard specifying the number of permissible LTUs at various physical layer data rates [14]. The advantage of LTU subdivision is that there is a CRC field for each LTU, rather than for each (large) physical layer frame. This facilitates faster error detection in the LTU subdivision case, and hence allows faster retransmissions of erroneous frames. Moreover, the retransmission need only be for erroneous LTUs rather than for the entire physical layer frame. The simulation experiments in the paper show that LTU subdivision results in a better normalized TCP throughput, especially as higher data rates are approached.

Discussion

LTU subdivision incurs the overhead of segmentation and re-assembly of the LTUs. On the other hand, the advantage of LTU subdivision that outweighs the overhead is that there is a CRC for each LTU, which gives the link layer a finer granularity for error checking. The sender may multiplex the LTUs to be retransmitted along with the LTUs belonging to the next link layer frame, thus leading to more efficiency in the functioning of the link layer.

6. Transport Layer Techniques

In this section, we present three transport layer techniques: Section 6.1 presents the Fast Retransmission (Explicit Handoff Notification) scheme, Section 6.2 presents the Indirect TCP scheme and Section 6.3 presents the PROBE and BUFFER+FREEZE scheme.

6.1 Fast Retransmission (Explicit Handoff Notification) Scheme

Caceres et al. [5] present a seminal analysis of the impact of handoffs resulting from the host movement on the throughput and delay of the reliable transport layer protocols. In their experiment, they analyze the performance of Tahoe TCP in a 2 Mbps wireless LAN environment with the host motion simulated in software. This setup allows for the simulation of both the possible cellular scenarios i.e. overlapping cells and non-overlapping cells with different beacon periods. The instants at which handoffs can be initiated are also precisely controlled using this host simulation software.

The handoff process when the mobile host (MH) moves across non-overlapping cells is illustrated in the Figure 5. We will use base station (BS) in our description, which is the equivalent of Mobility Support Stations (MSS) mentioned in [5]. The BS defines each cell and also acts the gateway for routing the packets to and from the mobile hosts (MH) in the cell. The BS makes the MH aware of their presence by broadcasting periodic beacons over the wireless medium. The BS for each cell runs Mobile IP protocol to provide support for inter-network mobility.

The MH senses the new beacon as it crosses over to the new cell and sends a Greeting packet to the new base station (BS2). BS2 updates its system structures as it accepts the visiting MH and sends a Greeting Acknowledgment to the MH. BS2 also notifies the old base station (BS1) that the MH has moved and that it can be reached through BS2. BS1 adjusts its routing tables in order to forward to BS2 any packets that arrive for MH and acknowledges the handoff to BS2. Finally, BS2 acknowledges the completion of handoff (via Handoff Completion signal) to the MH. The MH, earlier on the receipt of Greeting Acknowledgment, had already made BS2 as its new gateway for routing the packets.

Figure 5. Handoff Process

Owing to the non-overlapping nature of the cells, discrete beacon period and at least two packet exchanges that are required before the MH can see BS2, there is a finite time window during which the data and acknowledgements cannot be reached to either of the two BS. This results in loss of almost the entire TCP transmission window for the transmitter at MH and the fixed host (FH). The loss is more significant for the data destined for MH due the longer routing inconsistencies, as BS1 does not know that MH has left the cell until it receives an explicit notification from BS2.

Since the entire window is lost, the MH TCP sender can send no more data until the retransmission timeout occurs. This timeout throttles the congestion window and the retransmission timeout interval is also backed off exponentially. If the handoff is still in progress (which is likely in case the period between beacons is long), there may be multiple exponential back-offs which will have an effect of long pause in the active TCP connection where no data is being exchanged and the sender can do nothing but wait for the expiration of the retransmission timeout timer. The experiments indicate that this leads to a severe degradation in the average throughput (a drop of approx 30% as compared to no handoff scenario) for both the MH and SH when the beacon period is one second.

The proposed solution to alleviate this problem is to allow the IP layer to signal the end of handoff to the TCP, so that the TCP can cancel its retransmission timer (at the MH) and indicate the same to the TCP at FH by sending three duplicate acknowledgments. Another approach proposed to reduce the packet loss is to always buffer the last N bytes of data (where the value of N would be bounded by handoff latency) in the old router until the explicit notification is received from the new router. Increasingly cheaper memories and continual decrease in the handoff latency times make this approach worthwhile for further study.

We feel that it is pertinent to point the vertical dependence [8] apparent in the proposed solution in [5] in which the handoff completion signal from the IP layer to the TCP resulted in substantial gains (more than 20%) in improving the TCP throughput.

Discussion

An important issue worth considering is how the TCP at the hosts should react to this explicit information about the end of handoff. Is performing slow-start sufficient or should there be provisions for sophisticated bandwidth estimation techniques in the new cell?

The non-overlapping cells represent the break-before-make approach for the cellular environment which result in loss of connectivity in the data path between the MH and the SH. The beacon period adds additional delay and further more loss of packets, as the MH needs to sense the beacon in the new cell to resume communication.

CDMA cellular networks [11] allow the handoffs to be soft (refer Section 2) which is akin to overlapping cells mentioned in the paper [5] and allows the MH to retain the link with the old BS until the new link is completely established with the new BS. In addition, CDMA systems provide for continuous pilot signal, which solves the delay problem introduced by discrete beacons separated by time period. As majority of the handoffs in CDMA environment (80-85 %) tend to be soft, appreciable gains in the performance of TCP are realized in such cellular networks.

6.2 Indirect TCP Scheme

Bakre et al. [6] present an indirect transport layer approach, which is referred to as the split connection approach in [1]. There are two separate TCP connections: one from the FH (fixed host) to the MSR (mobility support router), and the other from the MSR to the MH (mobile host). The MSR is equivalent in functionality to the BS shown in Figure 5. This achieves the separation of the flow control and congestion control functionality on the wireless link from that on the fixed network. This allows a different (and potentially simpler) protocol to be used between the MSR and the MH. The basic idea is to modify only the wireless side of the connection (comprising of the MH and the MSR), while keeping the fixed side unchanged.

In addition to advocating an indirect transport layer approach, the paper also addresses the issue of mobile handoffs. The experiments use a similar framework as mentioned in section 6.1, and experiments are carried out for different cases of no moves, movement between overlapped cells and movement between non-overlapped cells. They are performed in both the LAN and WAN environments. The FH only sees an image of its peer MH that resides on the MSR. It is this image (i.e., state), which is handed over to the new MSR in case the MH moves to another cell. In the case of overlapped cells, the MH can continue to receive packets during handoff from the old

MSR, but during the state transfer stage, buffering may be required at the new MSR to minimize the packet losses during handoff. In the non-overlapped case, buffering may be required at both the old and the new MSR. The results of the experiments indicate that I-TCP performs better than regular TCP in all cases. The difference in performance between I-TCP and regular TCP is more pronounced in the WAN environment.

Discussion

The scheme allows the freedom to use a simpler variant of TCP on the wireless link between the MH and MSR that is commensurate with the processing and power limitations of hand-held devices.

The indirect transport layer approach is open to the criticism, as indeed done in [1], that it violates the E2E principle, since acknowledgments to packets can reach the FH even before the packets actually reach the MH. Perhaps anticipating this, the authors have presented a defense in the paper that most applications using TCP anyway rely on application layer acknowledgment and error recovery. TCP by itself cannot guarantee that the data has been received and acted upon by the receiving application, and hence they argue that I-TCP does not yield weaker E2E semantics in comparison to regular TCP. One drawback of the handoff process described in the scheme is that it requires buffering at the MSRs on a per TCP connection basis, which leads to a large increase in memory requirements at the MSRs.

6.3 PROBE and BUFFER+FREEZE Scheme

The paper by Chan et al. [7] presents two schemes at the transport layer to alleviate the impact of handoffs on TCP performance. The paper tries to address all the causes of throughput degradation associated with disconnection during handoff, which are 1) long communication pauses, 2) slow post-handoff recovery, and 3) successive retransmission timeouts due to disconnection. One of the approaches, PROBE, tries to make the TCP sender aware of the losses due to motion, and adapts the transport layer to react better to losses. The other approach, BUFFER+FREEZE, tries to hide the motion from the transport layer. Both approaches are fairly radical in that they propose a lot of changes in the TCP layers at both FH and MH, as well as the base station. The FH and MH are assumed to be the sender and the receiver respectively in the simulations presented in the paper.

In BUFFER+FREEZE, the TCP sender is kept unaware of the occurrence of handoff and its retransmission timer and states are “frozen”. To accomplish this, the base station changes the information in the TCP acknowledgments to make the sender interpret the event as the depletion of the receiver’s (MH’s) advertised window to 0 (the base station does this as part of the process of “snooping” on the TCP acks). The TCP sender stops sending further data and thus the base station need not buffer the packets. In-flight packets would be dropped and need to be retransmitted from the TCP sender once the handoff process is complete. The TCP sender keeps polling the base station periodically for the new value of the advertised window, and the old base station always replies with a 0-sized advertised window. When the handoff completes, the new base station receives the polling message and replies with an appropriate advertised window, which “unfreezes” the TCP sender and resumes normal TCP traffic flow.

In PROBE, the network layer at the MH signals the completion of handoff to the TCP layer at the MH. The MH TCP sends three copies of a special negative acknowledgment to the sender (FH) to denote the completion of handoff. Upon the receipt of any of the three messages, the FH enters a handoff recovery phase, and sends three special probe packets to the MH. On receiving the probe packets, the MH uses a technique similar to the receiver packet-pair mechanism to find out an estimate of the available wireless bandwidth in the new cell, and feeds this information back to the FH. The advantage of doing so is that it avoids injecting too many packets into the new wireless cell. The *ssthresh* variable is set to the value of the newly estimated bandwidth, and FH avoids the problems of small window size and slow recovery, and the FH exits handoff recovery phase. Normal TCP operation resumes upon the completion of retransmission of the data inferred to be missing due to the handoff. The paper also mentions that the MH to FH communication can also use a similar procedure except that some of the initial messaging is not necessary, as the MH has the exact knowledge of the handoff process.

Discussion

The scheme highlights several important considerations: 1) the prevention of the shrinking of the TCP window in the case of handoff, 2) the estimation of the available bandwidth in a new cell to avoid congestion, and 3) the non-requirement of buffering at the base station during handoffs. Even though the schemes may not be viable in terms of deployment due to the extensive nature of the proposed modifications, yet the above points are certainly worth considering in future studies.

7. Conclusions and Future Directions

We have surveyed various techniques that provide mechanisms at different layers, which help in improving the performance of TCP in CDMA cellular wireless networks. We feel that the physical and link layer techniques are absolutely necessary to deal with the harsh mobile radio environment. In addition, one or more transport layer techniques (may not be limited to the ones mentioned in the paper) may be combined judiciously with them to further improve the performance of TCP.

It was surprising that none of the surveyed techniques considered the use of the SACK TCP option in their experiments. We feel that SACK TCP may prove to be very useful in improving TCP performance in mobile radio environments, as it is known to be resilient to multiple packet losses in a single transmission window [10] and thus deserves attention in future studies. Also, the future studies of TCP performance in cellular wireless environments should use the extended version of the “ns” simulator (which allows accurate modeling of the physical layer, and includes support for the MAC layer) rather than using the standard version (used erroneously in the simulations in [7]).

The value of the delay parameter proposed in [3], for delaying the third dupack (as elaborated in section 5.1) is dependent on the network characteristics. In a cellular environment, the network characteristics change frequently due to the movement of the mobile, and thus the delay parameter value needs to be updated dynamically based on the change in network characteristics. One possible solution for the problem that came to our mind is that the link layer protocol might calculate an estimate of the time required for retransmission on the one-hop wireless link, and provides this estimate to the TCP layer at the mobile host. The TCP layer can then set the delay parameter to twice the provided estimate to allow sufficient time for at least one link layer retransmission. However, such a scheme would impose extra functionality on the link layer and would also violate the layering principle. Further study is required to be able to adapt the value of the delay parameter dynamically in accordance with changed network conditions.

[2], [3] and [4] focus on dealing effectively with wireless transmission errors but do not consider the impact of handoffs in their analysis. On the other hand, [5], [6] and [7] suggest techniques to deal with losses due to handoffs but do not include the effect of wireless transmission errors in their study. We feel that none of the surveyed techniques present a holistic approach in dealing with both the aspects of wireless transmission errors and handoffs. Except for [4], none of the papers deal specifically with the CDMA cellular environment and many of the results ([2], [3] [4] and [7]) have been drawn based on simulations. We feel that there is a strong need for carrying out real experimental studies in CDMA cellular networks and to substantiate the claims of effectiveness of these techniques with quantitative performance results. The study of TCP performance in CDMA environment has vast research potential, especially because the next generation cellular networks (3G) [15], [16] have chosen CDMA as the air interface protocol of choice.

References

1) 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, Vol.5, No.6, pp 756-769, December 1997

2) A.Chockalingam, M.Zorzi and R.R.Rao, “Performance of TCP on wireless fading links with memory,” Proc. IEEE, ICC’98, 595-600, 1998

3) N. H. Vaidya, M. Mehta, C. Perkins and G. Montenegro, "Delayed Duplicate Acknowledgements: A TCP-Unaware Approach to Improve Performance of TCP over Wireless,” Technical Report 99-003, Computer Science Dept., Texas A&M University, February 1999.

4) Y. Bai, P. Zhu, A. Rudrapatna and A. T. Ogielski. "Performance of TCP/IP over IS-2000 based CDMA radio links,”. Proc. IEEE, 52th VTC’2000, Fall 2000

5) R.Caceres and L.Iftode, “Improving the performance of reliable transport protocols in mobile computing environments,” IEEE Journal on Selected Areas in Communications, Vol. 13, No. 5, pp 850-857, June 1995

6) A.Bakre and B.R.Badrinath, “I-TCP: Indirect TCP for mobile hosts,” Proc. 15th IEEE International Conference on Distributed Computing Systems (ICDCS), pp 136-143 May 1995

7) A. Chan, Danny H.K. Tsang, and S. Gupta, "Impacts of Handoff on TCP Performance in Mobile Wireless Computing", Proceedings of IEEE International Conference on Personal Wireless Communications (ICPWC'97), India, December 1997

8) Cynthia D.Rais and Satish K. Tripathi, “Studying vertical dependence to improve NFS performance in wireless networks,” Cluster Computing, 1-2, pp 225-235, December 1998

9) G. Xylomenos and G. C. Polyzos. "Internet protocol performance over networks with wireless links". IEEE Network, 13(4):55-63, 1999

10) K. Fall, and S. Floyd. "Simulation-Based Comparisons of Tahoe, Reno and SACK TCP", ACM Computer Communications Review, Vol. 26, No. 3, July 1996.

11) William C.Y.Lee, “Overview of cellular CDMA,” IEEE Transactions on Vehicular Tehnology, Vol.40, No.2, May 1991

12) Phil Karn. "The Qualcomm CDMA digital cellular system". In Proceedings of the USENIX Mobile and Location Independent Computing Symposium, pages 35--39. USENIX Association, 1993.

13) TIA/EIA/IS-2000.2, “Physical Layer Standard for cdma2000 Spread Systems,” March 1999

14) TIA/EIA/IS-2000.3, “Medium Access Control (MAC) Standard for cdma2000 Spread Systems,” March 1999

15) 3GPP home page: http://www.3GPP.org

16) 3GPP2 home page: http://www.3GPP2.org

17) Theodore S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice Hall, 1996