Wireless Networks 8, 467–479, 2002

TCP Westwood: End-to-End Congestion Control for Wired/Wireless Networks

CLAUDIO CASETTI
Politecnico di Torino, Italy

MARIO GERLA
UCLA Computer Science Department, USA

SAVERIO MASCOLO
Politecnico di Bari, Italy

M.Y. SANADIDI and REN WANG
UCLA Computer Science Department, USA

Abstract. TCP Westwood (TCPW) is a sender-side modification of the TCP congestion window algorithm that improves upon the performance of TCP Reno in wired as well as wireless networks. The improvement is most significant in wireless networks with lossy links. In fact, TCPW performance is not very sensitive to random errors, while TCP Reno is equally sensitive to random loss and congestion loss and cannot discriminate between them. Hence, the tendency of TCP Reno to overreact to errors. An important distinguishing feature of TCP Westwood with respect to previous wireless TCP "extensions" is that it does not require inspection and/or interception of TCP packets at intermediate (proxy) nodes. Rather, TCPW fully complies with the end-to-end TCP design principle. The key innovative idea is to continuously measure at the TCP sender side the bandwidth used by the connection via monitoring the rate of returning ACKs. The estimate is then used to compute congestion window and slow start threshold after a congestion episode, that is, after three duplicate acknowledgments or after a timeout. The rationale of this strategy is simple: in contrast with TCP Reno which "blindly" halves the congestion window after three duplicate ACKs, TCP Westwood attempts to select a slow start threshold and a congestion window which are consistent with the effective bandwidth used at the time congestion is experienced. We call this mechanism *faster recovery*. The proposed mechanism is particularly effective over wireless links where sporadic losses due to radio channel problems are often misinterpreted as a symptom of congestion by current TCP schemes and thus lead to an unnecessary window reduction. Experimental studies reveal improvements in throughput performance, as well as in fairness. In addition, friendliness with TCP Reno was observed in a set of experiments showing that TCP Reno connections are not starved by TCPW connections. Most importantly, TCPW is extremely effective in mixed wired and wireless networks where throughput improvements of up to 550% are observed. Finally, TCPW performs almost as well as localized link layer approaches such as the popular Snoop scheme, without incurring the overhead of a specialized link layer protocol.

Keywords: congestion control, bandwidth estimation, wireless network

1. Introduction

Congestion control functions were introduced into the TCP in 1988 and have been of crucial importance in preventing congestion collapse [13]. The well-known challenge in providing TCP congestion control in a mixed (wired/wireless) environment is that current TCP implementations rely on packet loss as an indicator of network congestion. In the wired portion of the network a congested router is indeed the likely reason of packet loss. In the wireless portion, on the other hand, a noisy, fading radio channel is the more likely cause of loss. This creates problems in TCP Reno since it does not possess the capability to distinguish and isolate congestion loss from wireless loss. As a consequence, TCP Reno reacts to wireless loss with a drastic reduction of the congestion window, hence of the sender transmission rate, when the best strategy in fact would be not to decrease the retransmission rate. Thus, TCP congestion control requires supplementary link layer protocols such as reliable link-layer or split-connections approach to efficiently operate over wireless links [17]. Approaches to address this problem have been discussed and compared in the pioneering work by Balakrishnan et al. [3–5]. Three alternative approaches: end-to-end (E2E), Split Connection, and Localized Link Layer methods were carefully contrasted. The best performing approach was shown to be a localized link layer solution applied directly to the wireless links. A clever "snooping" protocol monitors the packets flowing over the wireless link as well as their related acknowledgments. The protocol entities cache copies of TCP data packets and monitor the ACKs in the reverse direction. If a packet loss is detected (i.e., through duplicate acknowledgments, DUPACKs), the cached copy is used for local retransmission, and any packets carrying DUPACK information back to the TCP sender are halted until local retransmission success (to avoid "premature" retransmission at the TCP sender). The protocol is

effective in reducing retransmissions, and, more importantly, in preventing the associated reduction in congestion window size.

Snoop, however, has its own limitations. First, it requires a snoop proxy in the base station. Also, if the TCP sender is the mobile, the TCP code must be modified to respond to Explicit Loss Notification (ELN) packets from the base station. In view of the limitations introduced by link layer solutions, it is of interest thus to explore E2E recovery solutions that are independent of the link layer, and thus, more versatile.

In this paper, we propose to handle wireless losses using a modified version of TCP Reno. This new version, which we named TCP Westwood (or TCPW for short), enhances the window control and backoff process. Namely, a TCPW sender monitors the acknowledgment stream it receives and from it estimates the data rate currently achieved by the connection. Whenever the sender perceives a packet loss (i.e., a timeout occurs or 3 DUPACKs are received), the sender uses the bandwidth estimate to properly set the congestion window (`cwin`) and the slow start threshold (`ssthresh`). By backing off to `cwin` and `ssthresh` values that are based on the estimated available bandwidth (rather than simply halving the current values as Reno does), TCP Westwood avoids reductions of cwin and ssthresh that can be excessive or insufficient. In this way TCP Westwood ensures both faster recovery and more effective congestion avoidance. Experimental studies reveal the benefits of the intelligent backoff strategy in TCPW: better throughput, goodput, and delay performance, as well as fairness even when competing connections differ in their end-to-end propagation times. In addition, our studies of TCPW friendliness when coexisting with TCP Reno is reassuring since we have observed that TCP Reno connections are not starved in the presence of TCPW connections. Most importantly, TCPW is very effective in handling wireless loss. This is because TCPW uses the current estimated rate as reference for resetting the congestion window and the slow start threshold. The current rate is only marginally impacted by loss (as long as loss is a relatively small fraction of data rate). The simulation results presented in section 4 confirm this claim. For example, a throughput improvement of up to 550% over TCP Reno has been observed. Other TCP variants that use bandwidth estimation to set the congestion window have been proposed before. To our knowledge, however, such schemes require the intervention of the network layer. For example, the BA-TCP (Bandwidth Aware-TCP) scheme [10] relies on intermediate routers to take measurements of available bandwidth and compute the "fair share" for the TCP connections. The fair share value is piggybacked in the TCP header and conveyed to the TCP source. The latter uses it to appropriately set its cwin and ssthresh parameters. BA-TCP and TCPW are similar in their reliance on bandwidth information to set congestion control parameters. However, while BA-TCP requires new network layer functions to measure available bandwidth and compute fair share, TCPW relies only on information readily available in the current TCP header. TCPW does not require any support from lower layers, and thus strictly adheres to layer separation and modularity principles. Also, TCPW does not require that any TCP option be used in segment headers (i.e., timestamps).

The paper is organized as follows. Sections 2 and 3 discuss the TCPW bandwidth estimation and the congestion control algorithm. TCPW performance behavior in wired and in mixed networks is studied in sections 4, 5 and 6, whereas section 7 concludes the paper.

2. End-to-end bandwidth measurement

2.1. The ACK-based measurement procedure

A fundamental design philosophy of the Internet TCP congestion control algorithm is that it must be performed end-to-end. The importance of the end-to-end principle [8] cannot be overstated. In fact, it is this principle that guarantees the delivery of data over any kind of heterogeneous network, in spite of all possible failures at intermediate devices (proxies). Moreover, the network is considered as a "black box". As a consequence, in our approach, a TCP source does not expect to receive any explicit congestion feedback from the network. Therefore, the source, to determine the rate at which it can transmit, must probe the path by progressively increasing the input load (through the slow start and congestion avoidance phases) until implicit feedback, such as timeouts or duplicate acknowledgments, signals that the network capacity has been reached [12,13].

The key idea of TCP Westwood is to use the "bandwidth estimate" to directly control the congestion window and the slow start threshold. The bandwidth is estimated by monitoring the TCP ACKs. Namely, the source performs an end-to-end estimate of the bandwidth available along a TCP connection by measuring and averaging the rate of returning ACKs. Bandwidth estimation (via ACK monitoring) has been used before to control the TCP window, but only indirectly, via the estimation of the bottleneck backlog [12,21].

After a congestion episode (i.e., the source receives three duplicate ACKs or a timeout) the source uses the measured bandwidth to properly set the congestion window and the slow start threshold, starting a procedure that we will call *faster recovery*. When an ACK is received by the source, it conveys the information that an amount of data corresponding to a specific transmitted packet was delivered to the destination. If the transmission process is not affected by losses, simply averaging the delivered data count over time yields a fair estimation of the bandwidth currently used by the source.

When duplicate ACKs (DUPACKs), indicating an out-of-sequence reception, reach the source, they should also count toward the bandwidth estimate, and a new estimate should be computed right after their reception.

However, the source cannot tell for sure which segment triggered the DUPACK transmission, and it is, thus, unable to update the data count by the size of that segment. An average of the segment sizes sent thus far in the ongoing connection should therefore be used, allowing for corrections when the next cumulative ACK is received. For the sake of simplicity, we assume all TCP segments to be of the same size. Following this assumption, we will further assume that sequence numbers are incremented by one per segment sent, although the actual TCP implementation keeps track of the number of bytes instead: the two notations are interchangeable if all segments have the same size.

It is important to notice that, immediately after a congestion episode, followed either by a timeout or, in general, n duplicate ACKs, the bottleneck is at saturation and the connection delivery rate is equal to the share of the best-effort bandwidth (i.e., saturation bandwidth) available to that connection. At steady state, under proper conditions, such as uniform propagation delays, for example, this is actually the "fair share". The saturation condition is confirmed by the fact that packets have been dropped, an indication that one or more intermediate buffers are full. Before a congestion episode, the used bandwidth is less than or equal to the available bandwidth because the TCP source is still increasing its window to probe the network capacity.

As a result, TCP Westwood adjusts cwin by taking into account the network capacity that is available to it at the time of congestion, whereas current TCPs "blindly" half the value of cwin.

2.2. Bandwidth estimation

The TCPW sender monitors ACKs to estimate the bandwidth currently used by, and thus available to the connection. More precisely, the sender uses (1) the ACK reception rate and (2) the information an ACK conveys regarding the amount of data delivered to the destination.

We discuss the use of the information in (2) in section 2.3. For now, let assume that an ACK is received at the source at time t_k, notifying that d_k bytes have been received at the TCP receiver. We can measure the following *sample* bandwidth used by that connection as $b_k = d_k/\Delta_k$, where $\Delta_k = t_k - t_{k-1}$ and t_{k-1} is the time the previous ACK was received.

Since congestion occurs whenever the low-frequency input traffic rate exceeds the link capacity [15] we employ a low-pass filter to average sampled measurements and to obtain the low-frequency components of the available bandwidth. Notice that this averaging is also useful in filtering out the noise due to delayed acknowledgments.

In our early design and experimentation, we used a filter similar to the one used for RTT estimation in TCP. We determined that such an exponential filter with constant coefficients is not capable of efficiently filtering out high-frequency components of the bandwidth measurements. We propose the following discrete-time filter which is obtained by discretizing a continuous low-pass filter using the Tustin approximation [2]

$$\widehat{b_k} = \alpha_k \widehat{b}_{k-1} + (1-\alpha_k)\left(\frac{b_k + b_{k-1}}{2}\right),$$

where $\widehat{b_k}$ is the filtered estimate of the available bandwidth at time $t = t_k$, $\alpha_k = (2\tau - \Delta_k)/(2\tau + \Delta_k)$, and $1/\tau$ is the cutoff frequency of the filter.

Notice that the coefficients α_k are made to depend on Δ_k to counteract the effect of non deterministic interarrival times. In fact, when the interarrival time Δ_k increases, the last value $\widehat{b}_{k-1}$ should have less significance. On the other hand, a recent $\widehat{b}_{k-1}$ should be given higher significance. The coefficient α_k decreases when the interarrival time increases, and thus the previous value b_{k-1} has less significance with respect to the last two recent samples which are weighted by $(1-\alpha_k)/2$. As an example, let $t_k - t_{k-1} = \Delta_k = \tau/10$. Then

$$\widehat{b_k} = \frac{19}{21}\widehat{b}_{k-1} + \frac{2}{21}\left(\frac{b_k + b_{k-1}}{2}\right).$$

The new value $\widehat{b_k}$ is thus made up of approximately 90% of the previous value $\widehat{b}_{k-1}$ plus approximately 10% of the arithmetic average of the last two samples b_k and b_{k-1}.

Since the TCPW filter has a cutoff frequency equal to $1/\tau$, all frequency components above $1/\tau$ are filtered out. According to the Nyquist sampling theorem, in order to sample a signal with bandwidth $1/\tau$ a sampling interval less than or equal to $\tau/2$ is necessary. But, since the ACK stream is asynchronous, the sampling frequency constraint cannot be guaranteed. Thus, to guarantee the Nyquist constraint, we establish that if a time τ/m ($m \geqslant 2$) has elapsed since the last received ACK without receiving any new ACK, then the filter assumes the reception of a *virtual* null sample $b_k = 0$. The situation is shown in figure 1, where t_{k-1} is the time an ACK is received, $\widehat{t}_{k+j}$ are the arrival times of the virtual samples, with $\widehat{t}_{k+j+1} - \widehat{t}_{k+j} = \tau/m$ for $j = 0, n-1$; and $b_{k+j} = 0$ for $j = 0, n-1$ are the virtual samples. Then, $b_{k+n} = d_{k+n}/\Delta_{k+n}$ is the bandwidth sample at t_{k+n}.

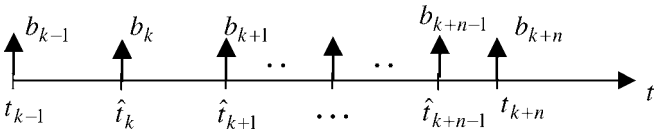

Figure 1. Bound on the maximum sampling interval obtained by inserting virtual sample.

It is desirable that after a long time without ACKs (i.e., because no new data were sent), the filter acts in a conservative fashion, progressively decreasing the bandwidth estimation as time elapses without new samples. Consider, thus, the operation of the TCPW filter when there is prolonged absence of ACKs after a time $t = t_k$. As can be inferred from

the following sequence of bandwidth estimates, the estimated bandwidth exponentially goes to zero:

$$\widehat{b}_k = \frac{2m-1}{2m+1}\widehat{b}_{k-1} + \frac{0+b_{k-1}}{2m+1},$$
$$\widehat{b}_{k+1} = \frac{2m-1}{2m+1}\widehat{b}_k,$$
$$\vdots$$
$$\widehat{b}_{k+h} = \left(\frac{2m-1}{2m+1}\right)^h \widehat{b}_k.$$

2.3. On the effects of delayed and cumulative ACKs on bandwidth measurement

As previously stated, DUPACKs should count toward the bandwidth estimation since their arrival indicates a successfully received segment, albeit in the wrong order. As a consequence, a cumulative ACK should only count as one segment's worth of data since duplicate ACKs ought to have been already taken into account. Further complications result from *delayed ACKs*. The standard TCP implementation provides for an ACK being sent back once every other in-sequence segment received, or if a 200 ms timeout expires after the reception of a single segment [18]. The combination of delayed and cumulative ACKs can potentially disrupt the bandwidth estimation process.

We, therefore, stress two important aspects of the bandwidth estimation process:

(a) the source must keep track of the number of DUPACKs it has received before new data is acknowledged;

(b) the source should be able to detect delayed ACKs and act accordingly.

The approach we have chosen to take care of these two issues can be found in the `AckedCount` procedure, detailed below, showing the set of actions to be undertaken upon the reception of an ACK, for a correct determination of the number of packets that should be accounted for by the bandwidth estimation procedure, indicated by the variable `acked` in the pseudocode. The key variable is `accounted`, which keeps track of the received DUPACKs. When an ACK is received, the number of segments it acknowledges is first determined (`cumul_ack`). If `cumul_ack` is equal to 0, then the received ACK is clearly a DUPACK and counts as 1 segment towards the BWE; the DUPACK count is also updated. If `cumul_ack` is larger than 1, the received ACK is either a delayed ACK or a cumulative ACK following a retransmission event; in that case, the number of ACKed segments is to be checked against the number of segments already accounted for (`accounted_for`). If the received ACK acknowledges fewer or the same number of segments than expected, it means that the "missing" segments were already accounted for when DUPACKs were received, and they should not be counted twice. If the received ACK acknowledges more segments than expected, it means that although part of them were already accounted for by way of DUPACKs, the rest are cumulatively acknowledged by the current ACK; therefore, the current ACK should only count as the cumulatively acknowledged segments. It should be noted that the last condition correctly estimates the delayed ACKs (`cumul_ack = 2` and `accounted_for = 0`):

```
PROCEDURE AckedCount
cumul_ack = current_ack_seqno
   - last_ack_seqno;
if (cumul_ack = 0)
   accounted_for = accounted_for + 1;
   cumul_ack = 1;
endif
if (cumul_ack > 1)
   if (accounted_for >= cumul_ack)
      accounted_for = accounted_for
         - cumul_ack;
      cumul_ack = 1;
else if (accounted_for < cumul_ack)
      cumul_ack = cumul_ack
         - accounted_for;
      accounted_for = 0;
   endif
endif
last_ack_seqno = current_ack_seqno;
acked = cumul_ack;
return(acked);
END PROCEDURE
```

3. TCP Westwood

In this section we describe how bandwidth estimation is used in the congestion control algorithm run at the sender side of a TCP connection. As will be explained, the fundamental congestion window dynamics during slow start and congestion avoidance are unchanged, that is they increase exponentially and linearly, respectively, as in current TCP Reno.

The general idea is to use the bandwidth estimate `BWE` to set the congestion window and the slow start threshold after a congestion episode. We start by describing the general algorithm behavior after n duplicate ACKs and after coarse timeout expiration.

3.1. Algorithm after n duplicate ACKs

The pseudocode of the algorithm is the following:

```
if (n DUPACKs are received)
   ssthresh = (BWE * RTTmin)/seg_size;
   if (cwin > ssthresh) /* congestion
      avoid. */
      cwin = ssthresh;
   endif
endif
```

Note that `seg_size` identifies the length of the payload of a TCP segment in bits.

During the congestion avoidance phase we are probing for extra available bandwidth. Therefore, when n DUPACKs are

received, it means that we have hit the network capacity (or that, in the case of wireless links, one or more segments were dropped due to sporadic losses). In either case, the estimated bandwidth BWE is recognized to be a "feasible", i.e., achievable bandwidth (in fact, it was just measured at packet loss time). The goal is to achieve this bandwidth (more appropriately rate) with the minimum possible window, the "ideal window" in order to avoid bottleneck congestion. By definition, the ideal window is equal to the pipe size when the bottleneck buffer is empty, i.e., `w = BWE * RTTmin`.

After the reception of n DUPACKs, TCPW sets both slow start threshold and congestion window equal to the ideal window `BWE * RTTmin`. The standard Fast Retransmit/Fast Recovery (á la Reno) then follows. It should be noted that the value `RTTmin` is set to the smallest RTT sample observed over the duration of the connection. This setting allows the queue to be drained after a congestion episode. Also, note that after `ssthresh` has been set, the congestion window is set equal to the slow start threshold *only* if `cwin > ssthresh`. The rationale behind using BWE to set the slow start threshold is that TCP exploits the slow start phase to probe for available bandwidth; it thus seems natural to set `ssthresh` to the value we *believe* represented the available bandwidth at the time of congestion.

3.2. Algorithm after coarse timeout expiration

The pseudocode of the algorithm is:

```
if (coarse timeout expires)
   ssthresh = (BWE * RTTmin)/seg_size;
   if (ssthresh < 2)
      ssthresh = 2;
   endif;
   cwin = 1;
endif
```

The rationale of the algorithm is again simple. After a timeout, `cwin` and `ssthresh` are set equal to 1 and `BWE`, respectively, so that the basic Reno behavior is still captured, while a speedy recovery is granted by the ssthresh being set to the bandwidth estimation at the time of timeout expiration.

3.3. TCP Westwood convergence to fair share

An important goal of any TCP implementation is for every connection to get its "fair share" of the bottleneck. We will use an informal argument similar to that used for Reno in [14] to show that TCPW achieves the fair share. Consider the case of two connections with the same RTT. Suppose, for the sake of example, that the RTT is X packet transmission times, and the bottleneck has X buffers. One connection, say A, starts first. Its window "cycles" between X and $2X$ (as per the TCPW algorithm described earlier in this section), each cycle terminating when buffer overflows. Later, connection B starts, first in slow start mode, and then in congestion avoidance mode. In congestion avoidance, during each cycle the windows A and B grow approximately at the same rate, i.e., one segment per RTT. Eventually, the bottleneck buffer overflows, terminating the cycle. One can show that the window at overflow is

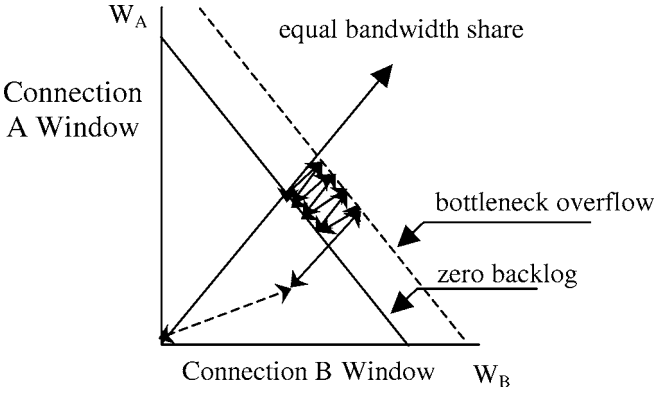

Figure 2. Convergence toward fair bandwidth sharing.

$$W_i = R_i\left(\frac{b}{C} + \text{RTT}\right) \quad \text{for } i = A, B,$$

where R is the achieved rate (i.e., `BWE`); b is the bottleneck buffer size; and C is the bottleneck trunk capacity.

This is a general property true for all TCP protocols, and in particular for TCPW. After overflow, TCPW reduces the windows to new values W_i' as follows:

$$W_i' = R_i\ (\text{RTT}) \quad \text{for } i = A, B.$$

Thus, the ratios of the windows A and B are preserved after overflow. Yet, the ratio W_B/W_A keeps increasing during congestion avoidance. Consequently, B's window and throughput ratchet up at each cycle. Equilibrium is reached when the two connections have the same windows and the same bandwidth share. Figure 2 graphically illustrates the convergence to the fix point $W_A = W_B$.

This informal proof is validated by actual simulation results. It can be generalized to many simultaneous connections (all with the same RTT). It can also be applied to the case when the bottleneck is affected by random errors equally hitting all connections.

4. TCP Westwood performance, fairness, and friendliness

In this section, we report on the basic performance behavior of TCPW, its fairness among a number of TCPW connections sharing a bottleneck link, and its friendliness to coexisting connections of other TCP variants, such as Reno.

First, the effectiveness of the bandwidth estimation algorithm is studied using a single TCP connection and a fluctuating UDP traffic rate. TCPW window dynamics (`cwin`, `ssthresh` and sequence numbers) are then considered. TCPW performance behavior is compared to the standard widely-used TCP Reno, as well as to TCP SACK [16].

All simulations presented in this paper were run using the LBL network simulator, *ns*-2 [19].

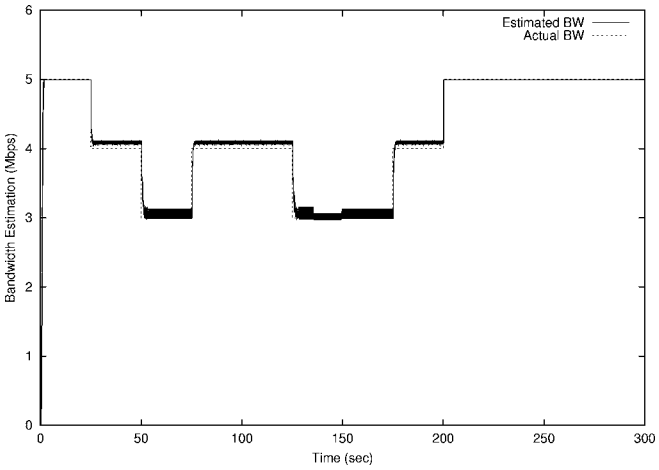

Figure 3. TCPW with concurrent UDP traffic: bandwidth estimation.

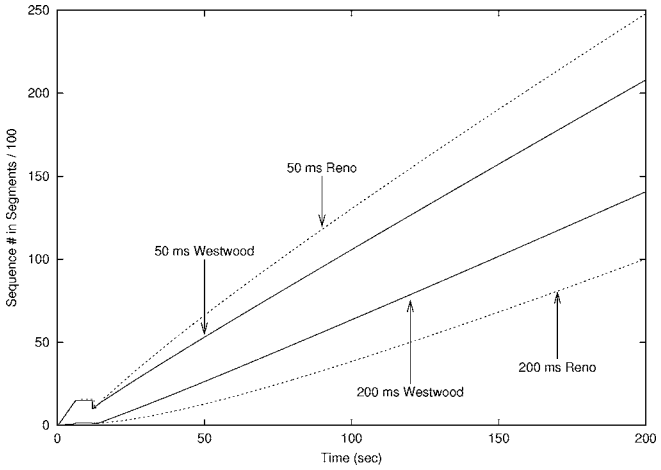

Figure 4. Sequence numbers versus time for long and short RTT connections without RED.

New simulation modules for TCP Westwood were written and they are available at [20], while existing modules for simulations involving TCP Reno and TCP SACK were used. All simulated TCP receivers implement delayed ACKs. Notice that this introduces a complication for our bandwidth estimation algorithm as delayed ACKs represent noise to be filtered.

Each scenario, involving different bottleneck link capacity, RTT or number of concurrent connections, includes a single-bottleneck link as is common in the literature. Intermediate node buffer capacity is always set equal to the bandwidth–delay product for the scenario under study. The packet size is set to 400 bytes in all experiments. The ACK arrival pattern is repetitive for each RTT in absence of packet losses (errors or buffer overflow). Thus, the interval τ should span one or more RTTs. Experimentally, we have observed that performance is not very sensitive to the choice of τ as long as $\tau >$ RTT. In our experiments, we set τ equal to 500 ms.

4.1. Bandwidth estimation effectiveness

In this section, we test the effectiveness of the proposed bandwidth estimation algorithm. For this purpose we consider a single TCPW connection sharing the bottleneck link with UDP connections. Packets are queued and transmitted on the link in FCFS order. In addition to demonstrating the accuracy of the bandwidth estimation algorithm, this scenario also illustrates the capability of a TCP Westwood connection to use the bandwidth left over by dynamic UDP flows. The configuration simulated here features a 5 Mbps bottleneck link with a one-way propagation delay of 30 ms. One TCP connection shares the bottleneck link with two ON/OFF UDP connections, and TCP and UDP packets are assigned the same priority. Each UDP connection transmits at a constant bit rate of 1 Mbps while ON. Both UDP connections start in the OFF state; after 25 s, the first UDP connection is turned ON, joined by the second one at 50 s; the second connection follows an OFF-ON-OFF pattern at times 75 s, 125 s and 175 s; at time 200 s the first UDP connection is turned off as well. The UDP connections then remain silent until the end of the simulation. The TCPW connection sends data throughout the simulation.

The scenario above is intended to demonstrate the effectiveness of the feedback control used in TCPW when subject to "step" and "impulse" stimuli. The behavior of the bandwidth estimation process is shown in figure 3.

4.2. TCPW fairness

Fair bandwidth sharing implies that all connections are provided with similar opportunity to transfer data. Our experiments show that TCPW fairness is at least as good, if not better, than that provided by the widely-used TCP Reno. In the sample results below we show that two flows with different round trip times (RTT) share the bandwidth more effectively under TCPW than under TCP Reno.

We ran simulations in which connections faced 50 ms and 200 ms RTT, respectively. Figures 4 and 5 show the sequence number progress for TCPW and Reno connections without and with RED, respectively. In all cases the short connection progresses faster as expected. We note however that TCPW provides better fairness than Reno across different propagation times. The reason for the superior fairness exhibited by TCPW is that the long connection suffers less reduction in `cwin` and `ssthresh`. In Reno, `cwin` reduction is independent of RTT. The results in figure 5 show that both protocols benefit from RED, as far as fairness is concerned. Remarkably, the improvement in TCPW due to RED was higher than the improvement in Reno.

4.3. TCPW friendliness

Friendliness is another important property of a TCP protocol. TCPW must be "friendly" to other TCP variants. That is, TCPW connections must be able to coexist with connections running TCP variants while providing opportunities for all connections to progress satisfactorily. At least, TCPW connections should not result in starvation of connections running other TCP variants. Better yet, the bandwidth share of TCPW connections should be equal to their fair share.

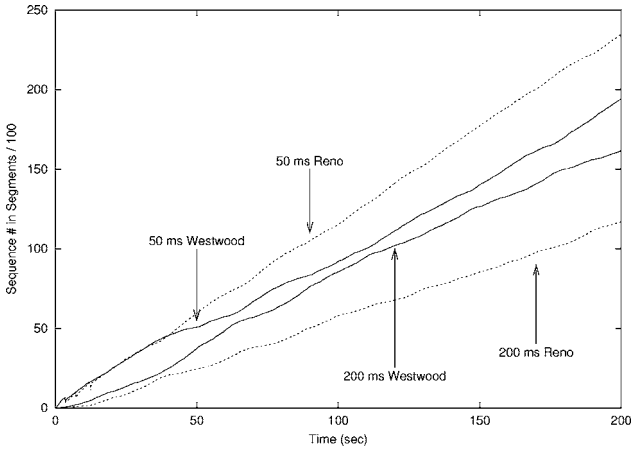

Figure 5. Sequence numbers versus time for long and short RTT connections with RED.

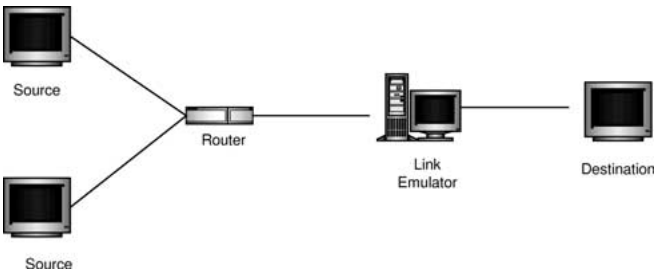

Figure 6. Experimental test bed layout.

We ran simulation experiments with the following parameters: 2 Mbps bottleneck link, 20 flows total, flows with RTT = 100 ms. With all 20 connections running TCPW, the average throughput per connection was 0.0994 Mbps. The 20 Reno connections have an average throughput of 0.0992 Mbps. As predicted, we got the same results for the two schemes. We then ran 10 Reno with 10 Westwood connections sharing the same 2 Mbps bottleneck link over a path of 100 ms RTT. The average throughput for a TCPW connection went up to 0.1078, and that of a Reno connection went down to 0.0913. This shows that TCPW behavior departs from "fair share" by 16% (TCPW gains 8% and TCP Reno loses 8%). This unfairness is rather moderate and it can be tolerated as coexistence with Reno is not compromised.

To probe the friendliness issue further, we also carried out actual measurements using our TCPW Linux implementation in our lab. Figure 6 shows the topology of our lab testbed. The link emulator is used to vary the link propagation time and error characteristics.

We measured the throughput for a total of 5 connections with a variable Reno/TCPW mix. Then, to evaluate the friendliness of TCPW under stress, we introduce a relatively high error rate on the bottleneck link, namely 1% packet loss (see figure 7). This error rate is actually appropriate for wireless links as we shall discuss later. Note that TCPW shines in presence of line errors, so friendliness in the error situation is even more difficult to establish than in error free operation.

Figure 7 shows the average throughput per connection for TCPW and for Reno with a 100 ms RTT. The lower average

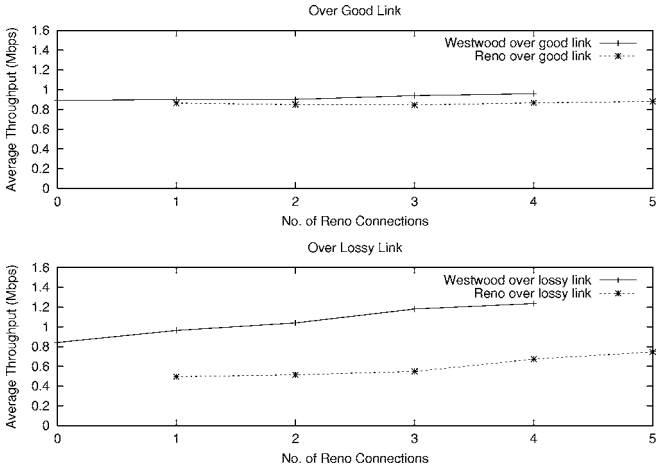

Figure 7. Average throughput versus number of Reno connections over good and lossy link (5 connections total).

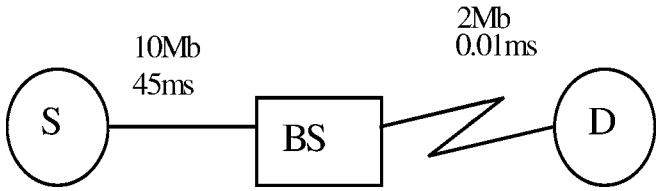

Figure 8. A simple simulation topology.

throughput line is that of Reno connections. The horizontal axis represents the number of Reno connections in the mix. For example, at the point marked 3 on the horizontal line, the measurement experiment includes 3 Reno connections and 2 TCPW connections. The results in figure 7 illustrate two important points. First, TCPW has a significant edge in a high-error-rate environment: 5 TCPW get 10% more throughput than 5 TCP Reno. We will press more on this later. Secondly, friendliness is preserved. Even though TCPW has an advantage over Reno in error-prone environments, Reno connections are not starved. Indeed, the introduction of TCPW connections into the mix reduces the average throughput of a Reno connection only by a minimal amount. Thus, for practical purposes, we can claim that TCPW is friendly.

5. TCPW performance in mixed (wired/wireless) networks

TCPW is being proposed in this paper as an end-to-end solution to error and congestion control in mixed wired and wireless networks. In view of this claim, a number of different scenarios are studied below to show the benefits of using TCPW in such wired/wireless environments. Independent and correlated loss models are used. Ground radio as well as satellite scenarios are developed and studied.

5.1. Independent loss model in ground radio environment

Figure 8 shows a topology of a mixed network with a wired portion including a 10 Mbps link between a source node and

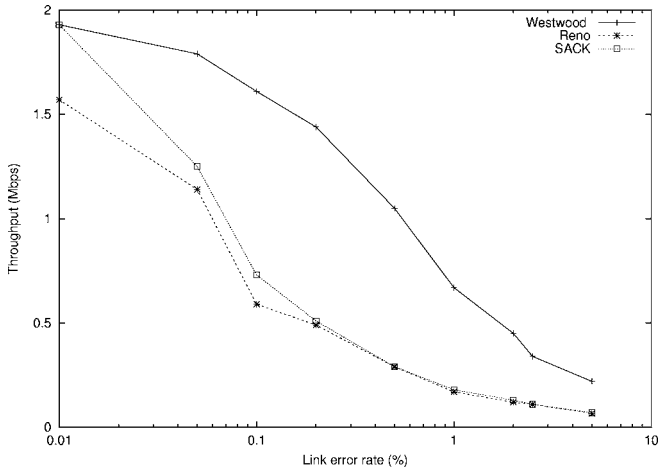

Figure 9. Throughput versus error rate of the wireless link.

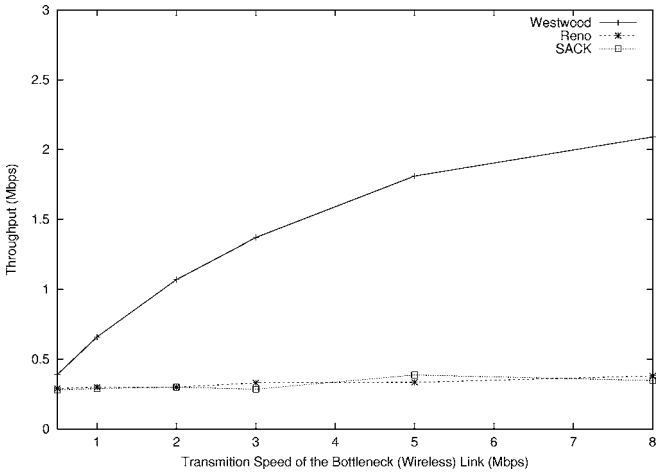

Figure 11. Throughput versus link capacity.

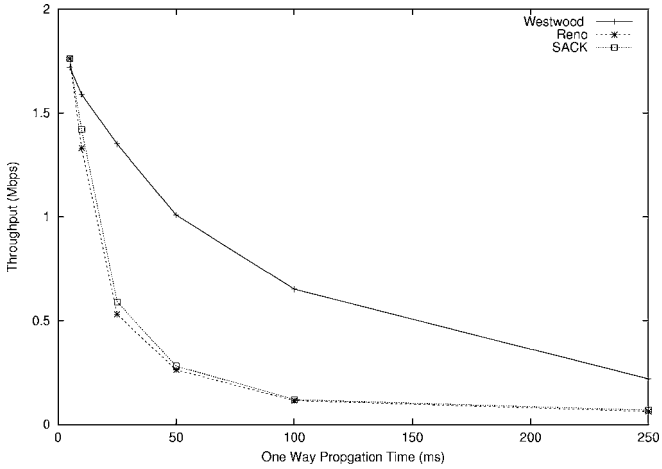

Figure 10. Throughput versus one-way propagation delay.

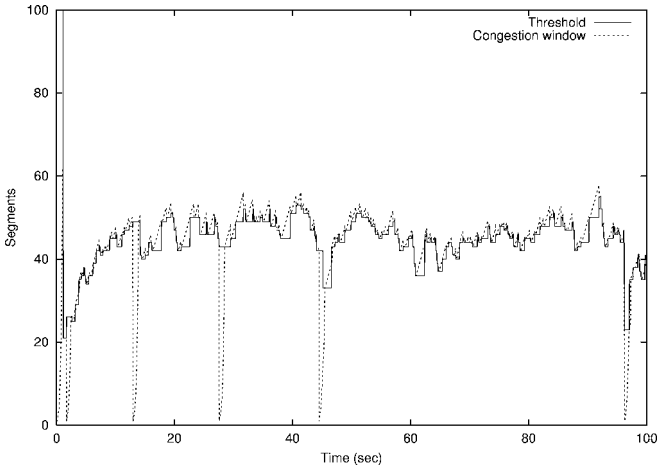

Figure 12. TCP Westwood over lossy link – `cwin` and `ssthresh`.

a base station. The propagation time over the wired link is initially assumed to be 45 ms. Later, the propagation time is varied from 0 to 250 ms to represent a variety of wired network environments ranging from campus to intercontinental connections.

The wireless portion of the network is a very short 2 Mbps wireless link with a propagation time of 0.01 ms. The wireless link is assumed to connect the base station to a destination mobile terminal. Errors are assumed to occur in both directions of the wireless link.

We compare the throughput of TCPW to that of Reno and SACK assuming independent (Bernoulli) errors ranging from 0 to 5% packet loss probability. The error model assumed here is equivalent to the "exponential error" model in which the time between back-to-back errors is exponentially distributed [3]. The range of error rates assumed here is also similar to the range used in [3]. The results in figure 9 show that TCPW gains up to 394% over Reno or SACK. This gain occurs at a realistic packet error probability of 1%.

To assess TCPW throughput gain and its relationship to the end-to-end propagation time, we ran simulations with the wired portion propagation time varying from 0 to 250 ms. The results in figure 10 show a significant gain for TCPW of up to 567%, at a propagation time of 100 ms. When the propagation time is small (say, less than 5 ms), all protocols are equally effective. This is because a small window is adequate and window optimization is not an issue. TCPW reaches the maximum improvement over Reno and SACK as the propagation time increases to about 100 ms. After that, in this experiment, the gain starts to decrease as the feedback information used to estimate the available bandwidth arrives too late to be of significant help to TCPW.

Simulation results in figure 11 show that TCPW gains also increase significantly as the bottleneck link transmission speed increases (again, because what matters is the window size determined by the bandwidth–delay product). Thus, TCPW is more effective than TCP Reno in utilizing the Gbps bandwidth provided by new-generation, high-speed networks. Figure 11 shows that the improvement obtained via TCPW increases to approximately 550% when the wireless link speed reaches 8 Mbps. The error model is still Bernoulli with parameter 0.5%, and the round trip time is 45 ms.

Window dynamics of TCPW and of TCP Reno are presented in figures 12 and 13. The graphs show the improved

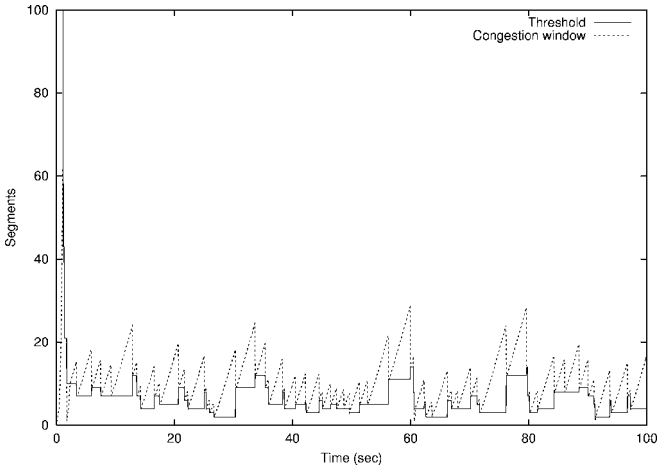

Figure 13. TCP Reno over lossy link – `cwin` and `ssthresh`.

window dynamics in TCPW. The cwin and ssthresh values are consistently higher than the corresponding values in Reno, thus yielding higher throughput.

Next, we compare TCPW to Snoop, the leading local strategy shown to provide the biggest improvement over TCP Reno [4]. Published results show that Snoop provides approximately a 400% improvement over an E2E approach based on TCP Reno when the error rate is 1 bit in 64 kBytes and the round trip propagation time is 135 ms. Our simulations with similar parameter values show that TCPW provides a 382% improvement over Reno. This shows that TCPW and Snoop gains are remarkably (and enticingly) close. We plan to probe further the issue of effectiveness of local versus E2E error recovery via simulation and measurements. From the qualitative and protocol implementation standpoint, however, we note that TCPW is completely end-to-end, and does not require any support from network or link layers. It does not have the scalability problems that Snoop may encounter as the number of mobile terminals increases. Further, the effectiveness of Snoop in wireless subnets including multiple base stations and handoffs is not clear.

Comparisons were also run with Explicit Loss Notification (ELN), which is an E2E scheme that was introduced and assessed in [3]. Basically, the method provides explicit notification from TCP receiver to TCP sender that a loss *due to a link error* has occurred. The lost packet is also identified to the Sender TCP entity. Using the same parameter values above (1 bit in 64 Kbytes error rate, and 135 ms propagation time), ELN is shown to provide a gain of approximately 200% over Reno. In comparison, TCPW provides 382%, closer to Snoop performance. Further, ELN assumes that the destinations can detect errors on a link and identify the packet and its TCP source. These assumptions are not likely to be uniformly satisfied for various error causes and various link technologies, hence the limited versatility of ELN in addition to its limited gain over Reno.

We compared, via simulation, TCPW to BA-TCP [10], an alternative strategy where the routers explicitly measure and relay the bandwidth available for each connection back to the TCP sender. At 40 ms round trip time, and 1 bit in 100 KB error rate, BA-TCP's improvement over Reno is 202%. For TCP Westwood, the throughput improvement at the same parameter values is 161%. This is quite remarkable considering that BA-TCP measures the bandwidth actually available for a connection at the bottleneck router, while TCPW works with no support from routers to estimate the available bandwidth at the bottleneck. Note that the router functionalities required by BA-TCP are not available in today's routers.

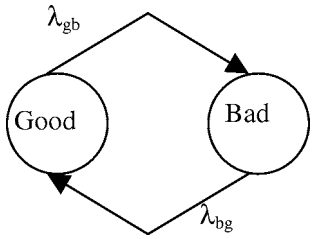

Figure 14. 2-state Markov model for burst error characterization.

5.2. Burst error models in a ground radio environment

To study TCPW performance with correlated errors, we use the 2-state Markov models following [1,5]. In such models, burst errors occur at a high rate due to a variety of conditions associated mostly with terminal mobility. Such conditions include variable fading, blackouts due to shadowing, and the like. Figure 14 depicts the 2-state Markov model. The wireless link is assumed to be in one of two states: Good or Bad. In the Good state, a bit (or packet) error Bernoulli model is assumed. The time intervals between bit errors is thus exponentially distributed (that is a memoryless model for channel errors). In addition, a link is assumed to stay in the Good state for a time interval that is exponentially distributed with parameter λ_{gb}. The time spent in the Bad state is also exponentially distributed but with parameter λ_{bg}. In the Bad state, we assume that errors are still Bernoulli-distributed, but their rates are much higher. For the simulation experiments below we vary the error rate in the Bad state depending on the specific link conditions we want to study. To represent fading conditions, the bit error rate is assumed to range from 0 to 30%. For blackouts, the error rate is 100%.

Simulation results using the 2-state Markov models show that TCPW improves the throughput for links with fading and blackouts as discussed below.

5.2.1. Fading

Let the Bad state represent fading conditions, and let the mean duration of Good and Bad states be 8 and 4 s, respectively. The error rate in the Good state is assumed to be 0.001% packet loss, and the error rate in the Bad state is varied from 0 to 30% packet loss rate. The results in figure 15 show the improvement obtained with TCPW over Reno or SACK. TCPW increases the throughput by as much as 300%. This is achieved when the error rate in the Bad state is 5%. When the error rate is higher, all protocols perform poorly. When the error rate is less than 5%, TCPW provides a 150% improve-

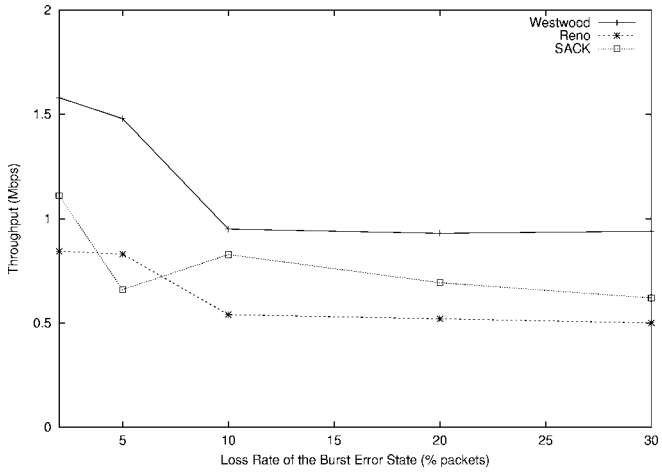

Figure 15. Throughput versus error rate of the bad state.

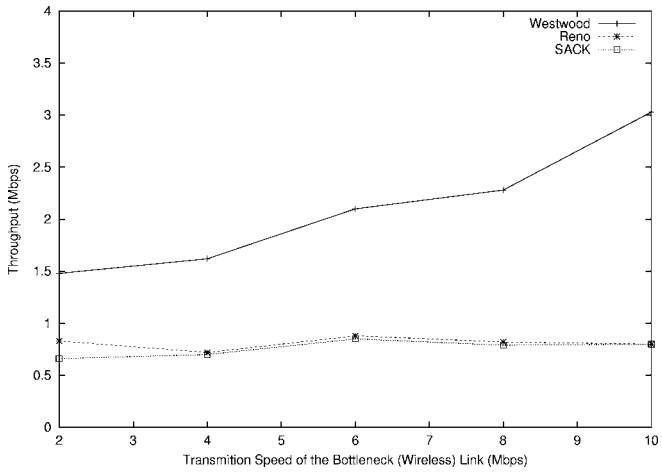

Figure 16. Throughput versus link capacity in 2-state error model.

ment. The link speed was also varied to determine its impact on protocol performance.

Figure 16 shows that TCPW improvement increases as the speed of the wireless link (bottleneck link in this case) increases (as expected, since a similar trend was observed also in wired links). Again, the error rate in the Bad state was 5%. It can be observed that, at 10 Mbps link speed, a remarkable 400% throughput improvement is achieved.

5.2.2. Blackouts

Let us now assume that the Bad state represents a blackout, where a base station becomes temporarily not visible to a terminal due to mobility. The mean duration for the Good state is 4 s; the mean duration of the Bad state varies between 0 and 0.5 s. Figure 17 shows the throughput improvement obtained by TCPW to be 167% over Reno and SACK when the mean blackout duration is 0.1 s. For longer blackouts, TCP timeouts occur and all protocols are equally affected.

5.3. LEO satellite model

Another environment where the improvement introduced by TCPW is likely to be valuable is the LEO satellite system.

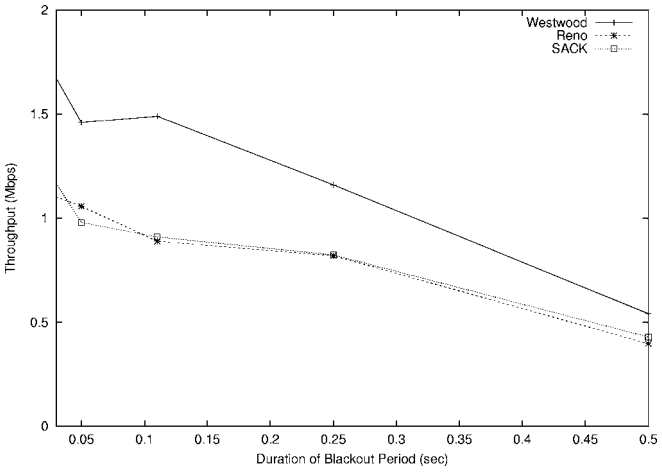

Figure 17. Throughput versus average duration of blackout.

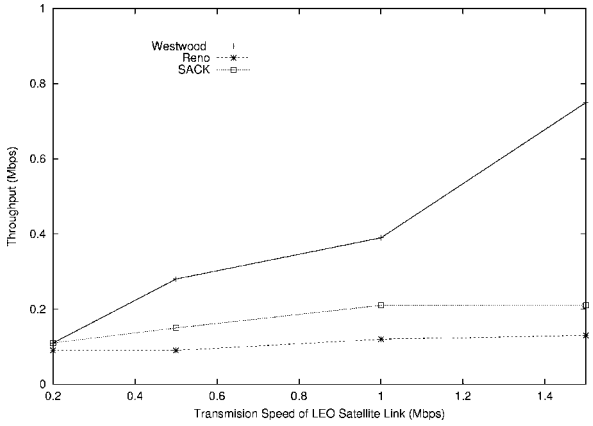

Figure 18. Throughput versus link capacity of the satellite link.

LEO Satellites present an environment with varying link quality and relatively long propagation delay [11]. Also, in the future, higher transmission speeds are expected. That is where TCPW would be most beneficial.

We considered for this study a scenario where a single hop, up to the satellite and down to an earth terminal, connects a terminal to a gateway and from there to the terrestrial network. The one-way (e.g., terminal to gateway) propagation time is assumed to be 100 ms. The error rate is assumed 0.1% in normal operating conditions. Occasionally, if the LEO system supports satellite diversity, a handoff to a different LEO satellite (different orbit) becomes necessary to overcome the blocking due to buildings, thick foliage etc. During handoff, we assume all packets are lost. In our model, the handoff from one satellite to another needs 100 ms to complete, and the period between handoffs is 4 s. Figure 18 shows the major improvements of TCPW over Reno and SACK, especially at high speeds.

6. Internet measurements

To test TCPW in an actual Internet environment, we have carried out a set of Internet experiments using the configuration depicted in figure 19. The sources are at UCLA, while

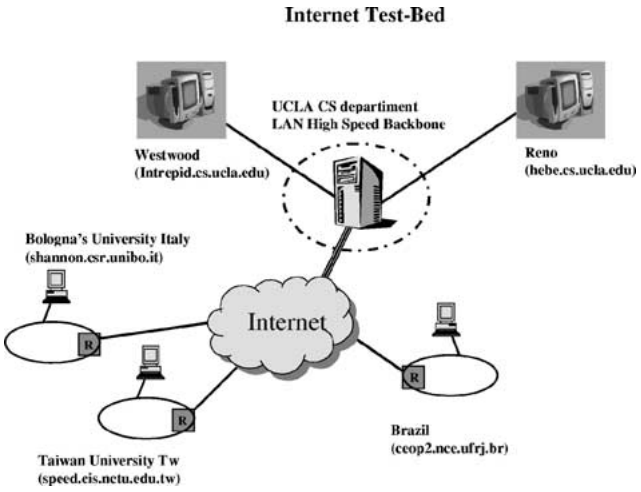

Figure 19. Internet measurement scenario.

Table 1
Internet throughput measurements.

Destination	Italy		Taiwan		Brazil	
RTT	170 ms		250 ms		450 ms	
Protocol	TCPW	Reno	TCPW	Reno	TCPW	Reno
Throughput (KB/s)	78.66	73.93	167.38	152	22.16	15.4

the destinations are chosen in three different continents (Europe, South America, and Asia). The destination hosts are, of course, unaware whether the source host runs TCPW or Reno.

Tests were scheduled during normal working hours at the destination sites. Experiments included either single or multiple file transfers. Throughput results were obtained by averaging repeated single file transfers. Multiple file transfer experiments were used to assess TCPW fairness. A rather large file size was used (10 Mbytes) to capture only steady state behavior. A standard FTP client (ncftp-3.0.2) was used as testing software with additional code for obtaining detailed logging at 1 s intervals. We measured application throughput in terms of user data/s as reported by ncftp. The average throughput achieved by Reno and TCPW on the various intercontinental connections is shown in table 1. Tests were repeated about 200 times throughout the day. The results show that TCPW performs marginally better that Reno on the Italy and Taiwan connections. It performs significantly better on the Brazil connection.

This result motivated further examination of the paths in question using the *traceroute* Linux tool. We found that Italy and Taiwan are connected using standard wired technology. In this case, link errors are expected to be minimal, thus, TCPW does not introduce much improvement over Reno. On the other hand, the Brazil path has a "lossy" satellite link provided by Teleglobe. The lossy link accounts for the TCPW improved performance.

In a second set of experiments, we compared the fairness of TCPW and Reno by injecting 20 connections simultaneously between UCLA and University of Bologna, Italy. The experiments were repeated 50 times over a 24 h period. The throughputs corresponding to the 20 flows are shown in figure 20. The results show excellent fairness across different TCP Westwood flows. The results show excellent fairness in Reno as well.

7. Conclusions and future research

In this paper we have proposed a new version of the TCP protocol, TCP Westwood (TCPW for short), aimed at improving performance under random or sporadic losses. TCPW has been tested through simulation, showing considerable throughput gains in almost all wireless scenarios.

In retrospect, the new scheme can be viewed as one more step in the TCP evolution. TCP Tahoe resets `cwin` to one after a loss. TCP Reno halves `cwin` after three duplicate ACKs. TCP Westwood introduces a "faster" recovery mechanism to avoid over-shrinking `cwin` after three duplicate ACKs. It does so by taking into account the end-to-end estimation of the bandwidth available to TCP. The use of bandwidth estimation to control the congestion window has an effect that goes beyond faster recovery. Namely, TCP window congestion control is based not solely on packet loss (which itself is an ambiguous congestion indicator in presence of wireless links), but also on the bandwidth estimate at the time of loss. The benefits of using bandwidth estimation (in addition to packet loss) have been amply demonstrated in a very broad range of wireless scenarios.

The issue of fairness and friendliness has been addressed. A proof of fair behavior (within TCPW flows) under equal propagation delay conditions has been provided. Friendliness to TCP Reno is more difficult to establish. "Unfriendly" trends due to TCPW "aggressiveness" have been detected in our experiments, but were shown to be contained and never severe enough to lead to starvation.

The code modifications required to implement TCP Westwood are comparable to the ones implemented in the transition from TCP Tahoe to TCP Reno. As in the Tahoe to Reno transition, a major advantage of the TCP Westwood modification is that it affects only the source TCP (as opposed to other variants such as TCP SACK that also require destination modifications). This allows a TCP Westwood source to "interwork" with arbitrary destinations in the Internet.

Work is in progress in many directions. The comparison of TCPW with link level techniques such as Snoop deserves further study. It is clear that link level recovery is in general much more powerful than end to end recovery since it isolates and corrects the loss "*in loco*". For instance, suppose that the bottleneck is in the wired network and one of the connections sharing the bottleneck goes over a wireless, lossy link. With end-to-end recovery (TCPW and TCP Reno alike) the wireless connection is heavily penalized with respect to the others. With link layer recovery (e.g., Snoop) fair sharing is enforced. Next, TCPW performs poorly when random packet loss rate exceeds a few percent. Snoop, on the other hand, is quite robust to high error rates. We are now investigating TCPW enhancements that will in part correct these deficiencies. We plan to further refine our bandwidth estimation and filtering

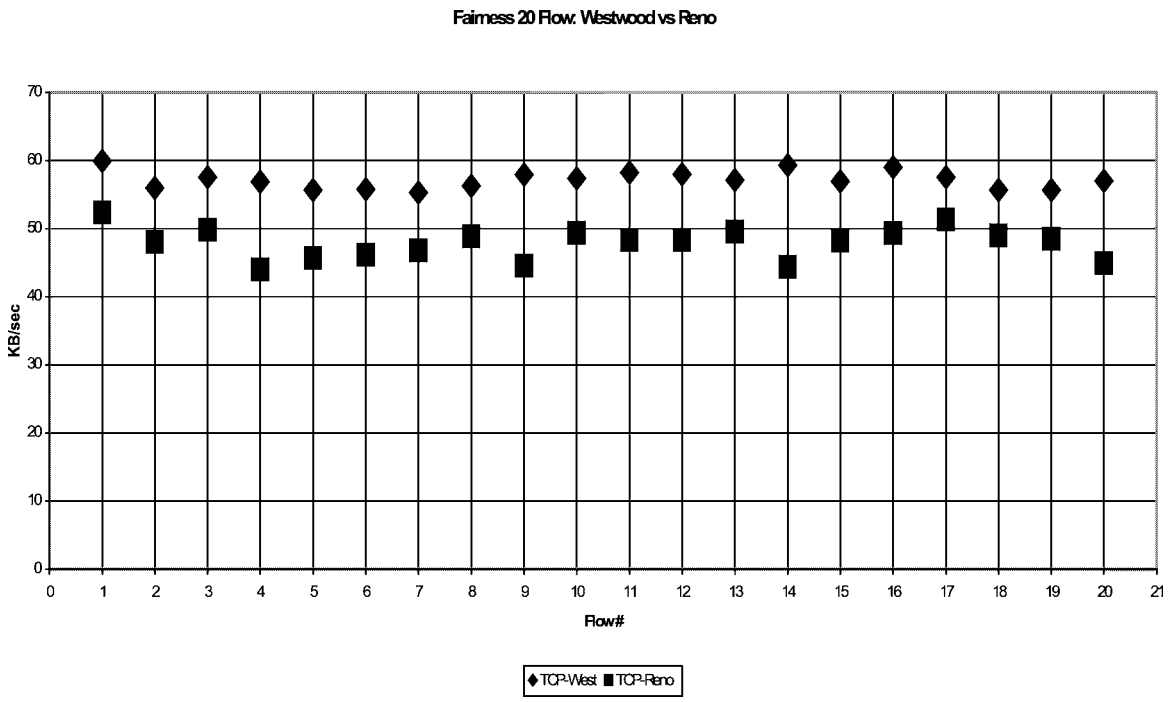

Figure 20. TCPW and Reno connections throughput (Internet measurements).

method, in order to improve TCPW "friendliness". Finally, we intend to pursue the development of control theoretical models that will enable us to study the stability of TCPW as a function of the various systems parameters.

Acknowledgements

The authors take great pleasure in acknowledging the valuable contribution to this work by Tienshiao Ma, Bryan Ng, Giovanni Pau and Cathy Yang, who valiantly implemented TCPW under Linux and conducted the lab measurements and experiments reported above. We also thank the anonymous referees for their thorough reviews and valuable comments.

This research was supported by NSF under Grant ANI-9983138 on High Speed Networks Performance Measurements and Analysis.

References

[1] A.A. Abouzeid, S. Roy and M. Azizoglu, Stochastic modeling of TCP over lossy link, in: *INFOCOM 2000*, Tel Aviv, Israel (March 2000).

[2] K.J. Åström and B. Wittenmark, Computer controlled systems (Prentice Hall, Englewood Cliffs, NJ, 1997).

[3] 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 (December 1997).

[4] H. Balakrishnan, S. Seshan, E. Amir and R.H. Katz, Improving TCP/IP performance over wireless networks, in: *MOBICOM'95*, Berkeley, CA (November 1995).

[5] H. Balakrishnan and R.H. Katz, Explicit loss notification and wireless web performance, in: *Proceedings of IEEE GLOBECOM'98 Internet Mini-Conference*, Sydney, Australia (November 1998).

[6] T. Bonald, Comparison of TCP Reno and TCP Vegas: Efficiency and fairness, in: *Proceedings of PERFORMANCE'99*, Istanbul, Turkey (October 1999).

[7] C. Casetti, M. Gerla, S. Lee, S. Mascolo and M. Sanadidi, TCP with faster recovery, in: *MILCOM 2000*, Los Angeles, CA (October 2000).

[8] D. Clark, The design philosophy of the DARPA Internet protocols, in: *Proceedings of SIGCOMM'88*, ACM Computer Communication Review 18(4) (1988) 106–114.

[9] M. Gerla, R. Lo Cigno, S. Mascolo and W. Weng, Generalized window advertising for TCP congestion control, CSD-TR 990012, UCLA, CA (February 1999).

[10] M. Gerla, W. Weng and R. Cigno, Bandwidth feedback control of TCP and real time sources in the Internet, in: *GLOBECOM'2000*, San Francisco, CA (November 2000).

[11] T. Henderson, Satellite transport protocol specification, Technical report, University of California, Berkeley (1999).

[12] B.S. Davies and L.L. Peterson, *Computer Networks: A Systems Approach*, 2nd ed. (Morgan Kaufman, 1999).

[13] V. Jacobson, Congestion avoidance and control, ACM Computer Communications Review 18(4) (August 1988) 314–329.

[14] J. Kurose and K. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet* (Addison Wesley, 2000).

[15] S.Q. Li and C. Hwang, Link capacity allocation and network control by filtered input rate in high speed networks, IEEE/ACM Transactions on Networking 3(1) (February 1995) 10–25.

[16] M. Mathis, J. Mahdavi, S. Floyd and A. Romanow, TCP selective acknowledgement options, RFC 2018 (April 1996).

[17] D. Mitzel, Overview of 2000 IAB Wireless Internetworking Workshop, RFC 3002 (2000).

[18] W.R. Stevens, *TCP/IP Illustrated*, Vol. 1 (Addison Wesley, Reading, MA, 1994).

[19] *ns*-2 network simulator, Version 2, LBL, `http://www-mash.cs.berkeley.edu/ns`

[20] TCP Westwood modules for *ns*-2, `http://www.telematics.polito.it/casetti/tcp-westwood`

[21] S. Keshav, A control-theoretic approach to flow control, in: *Proceedings of ACM SIGCOMM 1991* (September 1991).

Claudio Casetti graduated from the Politecnico di Torino with a Master's degree in electronic engineering in 1992 and a Ph.D. in electronic and telecommunication engineering in 1997. He was a visiting scholar at the University of Massachusetts at Amherst in 1996 and at UCLA in 2000. He is currently holding an Assistant Professor position at the Dipartimento di Elettronica e Telecomunicazioni of Politecnico di Torino. He has coauthored about 40 journal and conference papers in the area of communication networks. His interests are in the field of performance evaluation of TCP/IP networks and multimedia communications.
E-mail: casetti@polito.it

Mario Gerla is a Professor in the Computer Science Department at UCLA. He received his graduate degree in engineering from the Politecnico di Milano, in 1966, and his M.S. and Ph.D. degrees in engineering from UCLA in 1970 and 1973, respectively. He joined the Faculty of the UCLA Computer Science Department in 1977. His current research is in the area of analysis, design and control of communications networks. Ongoing projects include: the design and evaluation of QoS routing and multicast algorithms for IP domains; the design and evaluation of all-optical network topologies and access protocols; the design of wireless, mobile, multimedia networks for mobile computing applications; and the development of measurement methods and tools for evaluating the performance of high speed networks and applications.
E-mail: gerla@cs.ucla.edu

Saverio Mascolo is an Associate Professor of Automatic Control at the Department of Electrical and Electronics Engineer of Politecnico di Bari, Italy. He received the Laurea degree, cum laude, in electronic engineering in 1991 and the Ph.D. in 1995, both from Politecnico di Bari. He has been a visiting researcher at University of California, Los Angeles, and University of Uppsala. His current research interests are in the field of computer, control and communications. He has authored papers in the areas of data networks, manufacturing systems and nonlinear control of chaotic systems.
E-mail: mascolo@poliba.it

Yahya "Medy" Sanadidi received his Ph.D. in computer science from UCLA in 1982. He is currently a Research Professor at the UCLA Computer Science Department. As a co-principal investigator on NSF sponsored research, he is leading research in modeling and evaluation of high performance Internet protocols. He teaches undergraduate and graduate courses at UCLA on queuing systems and computer networks. Dr. Sanadidi was a manager and a senior consulting engineer at Teradata/AT&T/NCR from 1991 to 1999 and led several groups responsible for performance modeling and analysis, operating systems, and parallel query optimization. From 1984 to 1991, he held the position of a Computer Scientist at Citicorp where he pursued R&D projects in wireless metropolitan area data communications and other networking technologies. From 1981 to 1983, Dr. Sanadidi was an Assistant Professor at the Computer Science Department, University of Maryland, College Park, MD. There, he taught performance modeling, computer architecture and operating systems, and was a principal investigator for NSA sponsored research in global data communications networks. Dr. Sanadidi has consulted for industrial concerns, has co-authored conference as well as journal papers, and holds two patents in performance modeling. He participated as a reviewer and as a program committee member of professional conferences. His current research interests are focused on congestion control and adaptive application layer protocols in hybrid wired/wireless networks.
E-mail: medy@cs.ucla.edu

Ren Wang received her B.E. degree in automation from University of Science and Technology of China, her M.S. degree in computer engineering from Chinese Academy of Science, another M.S. degree in electrical engineering from University of California, Los Angeles (UCLA). Currently she is pursuing Ph.D. degree in computer science at UCLA under the guidance of Prof. Mario Gerla. Her research interests include network performance measurement and evaluation, network analysis and modeling, TCP protocol design and evaluation.
E-mail: renwang@cs.ucla.edu