DIMACS Series in Discrete Mathematics
and Theoretical Computer Science

Stochastic Modeling of a Single TCP/IP Session over a Random Loss Channel

Al-Hussein A. Abou-Zeid, Murat Azizoglu, and Sumit Roy

Abstract. In this paper, we present an analytical framework for modeling the performance of a single TCP session in the presence of random packet loss. This framework may be applicable to communications channels that cause random packet loss modelled by appropriate statistics of the inter-loss duration. It is shown that the analytical model predicts the throughput for LANs/WANs (low and high bandwidth-delay products) with reasonable accuracy, as measured against the throughput obtained by simulation. Random loss is found to severely affect the network throughput, higher speed channels are found to be more vulnerable to random loss than slower channels, especially for moderate to high loss rates.

1. Introduction

TCP/IP has been designed for reliable networks in which most packet losses occur primarily due to network congestion. An important aspect of TCP is its window-based congestion avoidance mechanism [**6**]. In TCP/IP, when a node successfully receives a packet, it sends an acknowledgment (ACK) back to the source. At all times, the source keeps a record of the number of unacknowledged packets that it has released into the network. This number is called the congestion window size, or simply, *the window size*. The source is allowed to increase its window size as long as packets keep being acknowledged. The source detects a packet loss by either the non-arrival of a packet ACK within a certain time (maintained by a *timer*), or by the arrival of multiple ACKs with the same next expected packet number. A packet loss is interpreted by the source as an indication of congestion, and the source responds by reducing its window size so as not to overload the network with packets, thereby indirectly controlling the data rate. Thus modelling the dynamic behavior of congestion window size is key to analyzing TCP/IP throughput performance in a variety of situations. The system (source, network and receiver) is frequently called a 'self-clocked' system since the arrival of ACKs acts as the clock that increases the window size, while the loss of a packet acts as a 'reset' for the system.

1991 *Mathematics Subject Classification.* Primary 54C40, 14E20; Secondary 46E25, 20C20.

An important system parameter that affects TCP throughput is the ratio β of the buffer size available in the bottleneck links of the network to the link bandwidth-delay product. For LANs, the round-trip delay in a connection is small, so that the bandwidth delay product could be much smaller than the buffer size (large β). WANs, on the other hand, have large round-trip delays, so that the buffer size is typically smaller than the bandwidth-delay product (small β). It is easy to see that, for the same bandwidth delay product, increasing the buffer size at a bottleneck link (increasing β) decreases the amount of congestion. It is worth noting that the bandwidth-delay product determines the maximum number of packets that can be in transit between a source-destination pair.

Network congestion is not the only source of packet loss - random packet loss may be significant in many circumstances. For example, it may arise due to intermittent faults in hardware elements in wired networks. In a wireless network, multipath fading that characterizes many terestrial links may be modelled as leading to random packet loss depending on the fade rates relative to data rate on the channel. While random packet loss on the Internet has been reported in [**8**], it was not taken into consideration in the congestion control mechanism in TCP/IP. Previous research [**2, 3, 4, 5**] has shown that random packet loss (which is not due to congestion) may severely decrease the throughput of TCP because TCP interprets random packet loss to be due to congestion and hence lowers the input data rate into the network, and consequently the throughput.

In [**2, 3**], a discrete-time model for random packet loss was used in which any given packet is assumed to be lost with probability q independent of all other packets. This model induces a geometric distribution on the number of packets successfully transmitted between consecutive packet losses, that may or may not be appropriate for specific lossy networks. A different loss model was employed in [**5**] which assumed that packet loss is characterized by an inhomogeneous Poisson process. The steady-state distribution of the window size was obtained in [**5**] under the assumption of infinite buffer size. In contrast, in this paper we assume a continuous-time packet loss model governed by a general renewal process and investigate the effect of finite buffer size on the performance.

A basic system model is shown in Figure 1. An infinite source (i.e. one that always has a packet to send) releases packets into a buffer of size B upon receiving ACKs from the destination. The packets are then sent over a single link with capacity μ packets per second and a net delay of τ (propagation delay through the channel, any other processing delays etc.) is assumed. Define $T = \tau + 1/\mu$ to be the time between the start of transmission of a packet and the reception of an ACK for this packet. Then μT is the bandwidth-delay product and the ratio $\beta = \frac{B}{\mu T}$ is the buffer size normalized by the bandwidth-delay product.

2. Ideal Channels without Random Packet Loss

We first briefly review the operation of TCP for the case of ideal channels, and summarize the key results in [**1, 3**] relevant to our work. There are different versions of TCP - the popular Tahoe version developed by Jacobson [**6**] (TCP-T) as well as the Reno version (TCP-R) that incorporates a fast retransmit option together with a method for reducing the effect of slow start [**7**]. While [**1, 3**] considered both versions, our analysis concentrates on TCP-R only, since our aim is to present

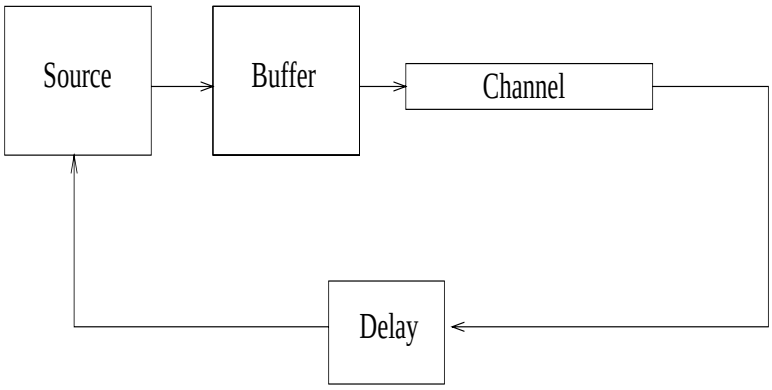

FIGURE 1. Block diagram of the system model

an analytical framework for the analysis of TCP/IP rather than to compare the different versions.

In TCP/IP, there are no explicit ACK or negative ACK signals. When the source sends a packet, it initializes a timer with an expiry time that is set depending on a current estimate of the delay τ. When a destination correctly receives a packet, it sends a signal to the source with the 'next expected' packet number. If this is received by the source prior to timer expiry, it is considered an ACK; otherwise, the packet is considered lost. In some implementations of TCP such as TCP-R, if the source receives multiple ACK signals with the same 'next expected' packet number, it interprets this as a packet loss.

Let $t' = 0$ denote the time of establishment of the TCP session under consideration, and let $W(t')$ denote the congestion window size at time t'. Then the algorithm followed by a TCP session (assuming 'ideal' operation of TCP) can be described as follows:

TCP-Tahoe:

1. Every time a packet ACK is received,
   if $W(t') < W_{th}$, set $W(t') = W(t') + 1 \ldots$ **Slow Start Phase**
   elseif$((ACKcount++) == W(t'))$ $\{$ $ACKcount = 0, W(t') = W(t') + 1$ $\}$ $\ldots$ **Congestion Avoidance Phase**

2. Every time a packet loss is detected,
   set $W_{th} = W(t')/2$ and set $W(t') = 1$. (**Go back to Slow Start Phase**)

TCP-Reno:

1. Every time a (non-repeated) packet ACK is received,
   if $W(t') < W_{th}$, set $W(t') = W(t') + 1 \ldots$ **Slow Start Phase**
   elseif$((ACKcount++) == W(t'))$ $\{$ $ACKcount = 0, W(t') = W(t') + 1$ $\}$ $\ldots$ **Congestion Avoidance Phase**

2. Every time a packet loss is detected via duplicate ACK,
   set $W_{th} = W(t')/2$, then set $W(t') = W(t')/2$ (the algorithm doesn't initiate

slow start)

3. Every time a packet loss is detected via timer expiry, the algorithm goes to slow start (as in TCP-T)
   set $W_{th} = W(t')/2$ and set $W(t') = 1$.

A TCP-T session typically evolves as follows. A packet is released from the source into the buffer just after the session is established and the transmitter enters 'slow start' phase. Each time an ACK is received, the window size is incremented according to slow start until it reaches W_{th} at which time the algorithm switches to congestion avoidance phase. The source subsequently increases its window size by one only after every window's worth of acknowledgments (congestion avoidance phase). This continues until a packet loss is detected, whence the window size is reset to one and the algorithm re-enters the slow start phase. This cycle repeats itself until all the packets at the source are transmitted and acknowledged at which time the TCP/IP session is terminated. It is worth noting that during slow start, the window size actually increases rapidly; in order to increase the window size by one when an ACK is received, the source releases two packets simultaneously into the buffer, while, to keep the window size constant, the source releases only one packet (to replace the one just acknowledged). In slow start, the source always releases two packets at each ACK reception, while in congestion avoidance, the source releases one packet for each ACK reception (to keep the window size constant) except when a window's worth of acknowledgment is received, in which case it releases two packets.

In a TCP-R session, the first cycle after session establishment is identical to that of TCP-T. However, after the first packet loss, the algorithm does not go back to slow start. Instead, the algorithm reduces the window size to half its value and continues in the congestion avoidance phase. In this description, we have assumed that the algorithm implements Selective Acknowledgments (SACKs) so that loss of multiple packets does not lead to time out phenomena, and hence the algorithm never enters Step 3 in the pseudo-code above.

Finally, we assume that fast retransmit option is used [**11**] in the event of a packet loss detection to avoid lengthy stoppage of transmission.

Denote $w_p = \mu T + B = \mu\tau + B + 1$, and note that when the window size reaches w_p, the bit pipe (the combination of the channel and the transmit buffer) is fully utilized, i.e, the buffer is fully occupied and the maximum number of packets allowable are in transit. A further increase in window size at this stage causes buffer overflow, at which point the window size is reset (for TCP-T) or halved (for TCP-R) and W_{th} is set to $w_p/2$.

Sample functions of the window size obtained from simulations for TCP-Tahoe and TCP-Reno are shown in Figure 2 when the only source of packet loss is buffer overflow.

A note about buffer overflow is in order. As can be noticed in Figure 2, the buffer build up rate in the slow start phase is very high, due to the 'fast' nature of slow start. Hence, one would expect that for small values of β, buffer overflow can take place in slow start. This is actually true, and has been noted in [**3**]. In such a case, for TCP-T, the window size is set back to one and another slow start phase follows with a lower threshold. For TCP-R, the window size is set to half its value

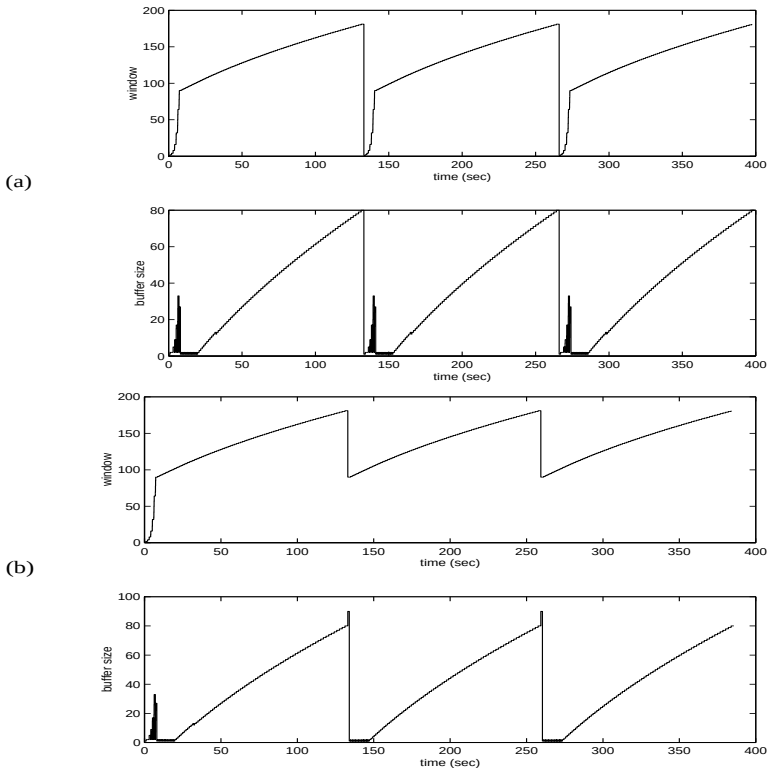

FIGURE 2. Window and buffer size evolution without random loss of packets for $\mu = 100$, $\tau = 1.0$, $\beta = 0.8$. (a)TCP-T, (b) TCP-R.

and the window evolution continues as depicted earlier. The effect of this 'double' slow start is significant for TCP-T but negligible for TCP-R.

The algorithms outlined above, defining the window size evolution during a TCP session have been studied in [**1, 3**] and mathematical expressions have been obtained for the envelope of the window size, denoted, for convenience, also by $W(t')$, for each of the phases (slow start and congestion avoidance). We summarize those expressions in a convenient way (so as to be able to use them in the following section) as follows (refer to Figure 3). Let n denote the number of packets acknowledged during a time interval t. Then the window evolution can be expressed as follows:

1. **Slow Start** $(1 < W(t') < W_{th})$. Consider two instants t'_0, $t'_0 + t$ in a slow start phase of any of the TCP cycles. Choose t'_0 such that $W(t'_0) = 1$.Then,

$$W(t'_0 + t) = 2^{t/T} \tag{2.1}$$

$$n = W(t'_0 + t) - 1 \tag{2.2}$$

2. **Congestion Avoidance - Phase I** $(W_{th} < W(t') < \mu T)$. Consider two instants t'_0, $t'_0 + t$ in a congestion avoidance phase of any of the TCP cycles.

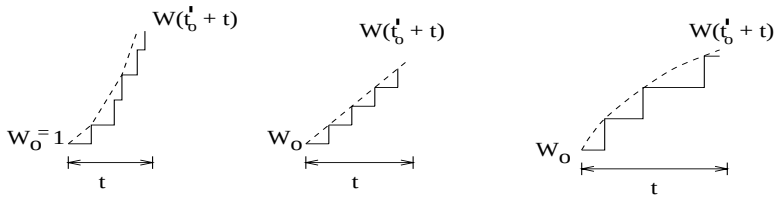

FIGURE 3. Sketch of the exponential, linear and sub-linear $O(\sqrt{t})$ phases for window evolution. Solid lines indicate the actual window size evolution while dotted lines indicate the envelope.

Choose t'_0 such that $W(t'_0) = W_0$.Then,

$$W(t'_0 + t) = W_0 + t/T \tag{2.3}$$

$$n = \frac{1}{T}(W_0 t + t^2/(2T)) \tag{2.4}$$

3. **Congestion Avoidance - Phase II** $(\mu T < W(t') < w_p)$. Consider two instants t'_0, $t'_0 + t$ in a congestion avoidance phase of any of the TCP cycles. Choose t'_0 such that $W(t'_0) = W_0$.Then,

$$W(t'_0 + t) = \sqrt{W_0^2 + 2\mu t} \tag{2.5}$$

$$n = \mu t \tag{2.6}$$

Note that in the sequel, we focus on time instants t' where $W(t')$ is discrete. Equation (1) follows from considering the time instances at which packets are sent and packet ACKs are received, and noting that two packets are sent for each ACK received. Equation 2 follows from the fact that each time an ACK is received, $W(t')$ is incremented by one; since $W(t')$ is initially equal to one, the total number of packets acknowledged at the end of the slow start phase is one less than equal to the window size at the end of the slow start phase. Equations (3)-(6) follow from a continuous time approximation of the window size. Specifically in the congestion avoidance phase, if we approximate the window size by a continuous time function, then

$$\frac{dW(t')}{dt'} = min\{\frac{1}{T}, \frac{\mu}{W(t')}\},$$

the solution to which yields (3),(5). Note that (3) indicates a linear growth of the window size with time with a slope of $1/T$, while (5) specifies a growth that is $O(\sqrt{t})$. This can be explained as follows - for $W(t') < \mu T$, the channel is able to transmit $W(t')$ packets in T seconds (linear growth with slope $1/T$), until the window size reaches μT, at which time buffer build-up set in. Thus it takes more than T seconds to send $W(t')$ packets subsequently and hence the window increase is slower than linear.

Using (2.1) - (2.6) and taking into consideration the periodic evolution of the window size outlined in the previous discussion, it is straightforward to compute the average packet transmission rate R as the ratio of the number of packets sent in one cycle of the TCP session to the time duration of the cycle. Note that for the case of TCP-R, we neglect the first cycle since it is different from the rest of the cycles (it is the only one that contains slow start). The average transmission rate is given by

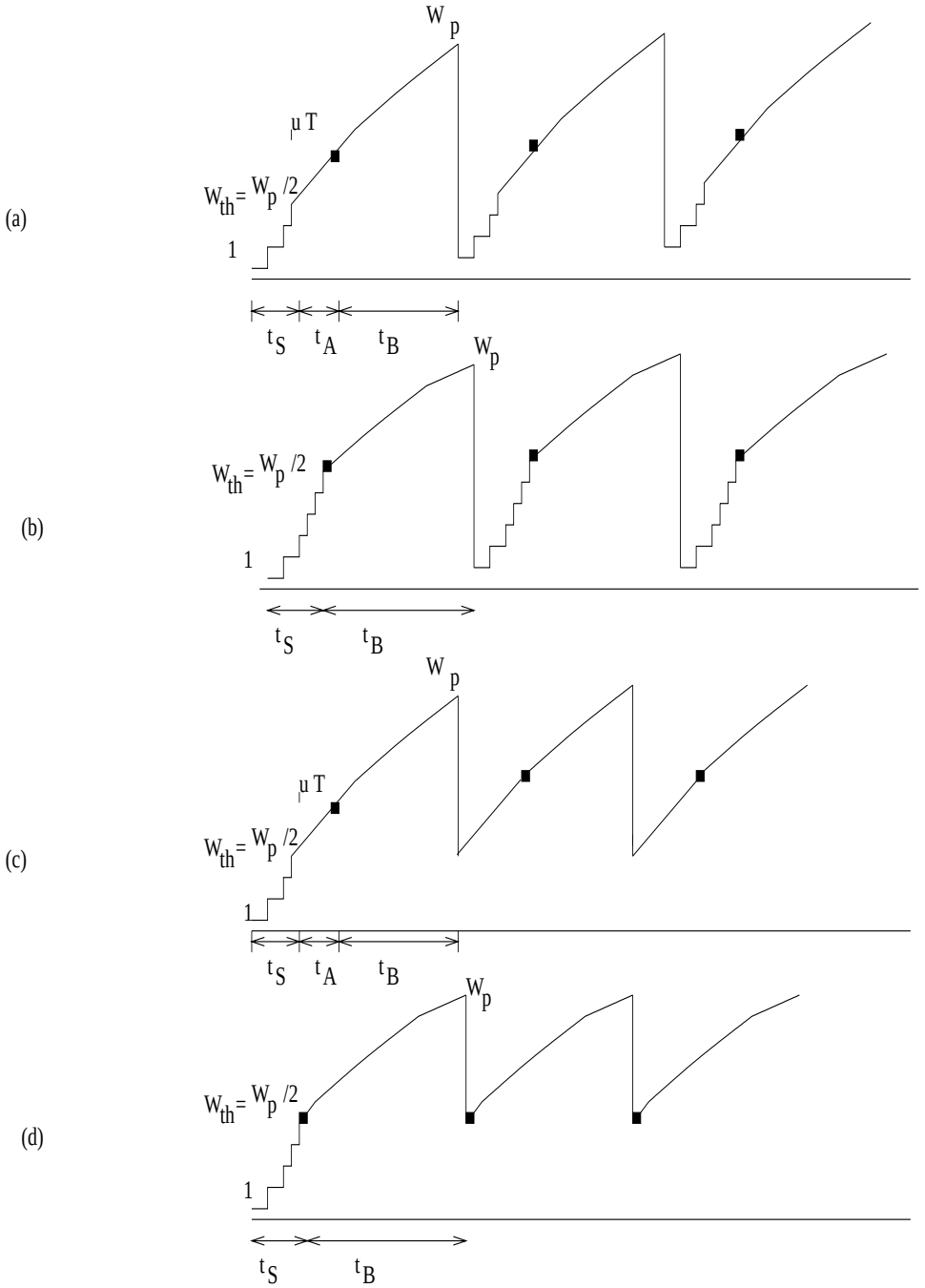

FIGURE 4. Sketch of the TCP window evolution without random loss ; (a) TCP-T with $\beta < 1$ (b) TCP-T with $\beta > 1$, (c) TCP-R with $\beta < 1$, (d) TCP-R with $\beta > 1$.Note that for $\beta < 1$, $w_p/2 < \mu T$, and hence the window evolution has a linear phase, while for $\beta > 1$, $w_p/2 > \mu T$ and hence the window evolution doesn't have a linear phase.

TCP-T:

$$\beta < 1 : R = \frac{n_S + n_A + n_B}{t_S + t_A + t_B} \tag{2.7}$$

$$\beta > 1 : R = \frac{n_S + n_B}{t_S + t_B} \tag{2.8}$$

TCP-R:

$$\beta < 1 : R = \frac{n_A + n_B}{t_A + t_B} \tag{2.9}$$

$$\beta > 1 : R \simeq \mu \tag{2.10}$$

The average throughput is then calculated as

$$\rho = \frac{R}{\mu} \tag{2.11}$$

The values above n_S, n_A, n_B, t_S, t_A and t_B are obtained by substituting for W_0 and $W(t')$ (see Figure 3) in (1)-(6) by the initial and final values of the slow start and congestion avoidance phases as indicated in Figure 4. For example, to compute n_S and t_S for TCP-T, set $W(t_S) = w_p/2$ in (1); to compute n_B and t_B for TCP-R and $\beta < 1$, let $W_0 = \mu T, W(t_B) = w_p$ in (3); to compute n_B and t_B for TCP-R and $\beta > 1$, let $W_0 = w_p/2, W(t_B) = w_p$ in (3) and so on.

Note the difference in the expressions for $\beta < 1$ and $\beta > 1$. Consider TCP-T; for $\beta < 1$, $w_p/2 < \mu T$ and hence the cyclical evolution of the window size consists of an exponential growth (the slow start phase from 1 to $w_p/2$) described by (1), then a linear growth (congestion avoidance phase from $w_p/2$ to μT) expressed by (3) and then $O(\sqrt{t})$ growth (congestion avoidance phase from μT to w_p) described by (5). On the other hand, for $\beta > 1$, $w_p/2 > \mu T$, and hence, the cyclical evolution consists of exponential growth (slow start from 1 to $w_p/2$) followed by $O(\sqrt{t})$ growth (congestion avoidance from $w_p/2$ to w_p) without having a linear phase. Similar observations apply for TCP-R; refer to Figure 4 for a schematic diagram of the window size evolution for each of the two ranges of β.

3. Channels with Random Packet Loss

3.1. Random Loss Model. Let S_i denote the time of the i^{th} packet loss, for $i = 1, 2, \ldots$. Let $X_i = S_i - S_{i-1}$ denote the time between $(i-1)^{\text{th}}$ and i^{th} packet losses with $X_1 = S_1$ by convention. As stated earlier, we will consider $\{X_1, X_2, \ldots\}$ to be a set of IID random variables with probability density function $f(x)$ and distribution function $F(x)$. Thus, the process (pdf) defined by the loss occurrence times $\{S_1, S_2, \ldots\}$ is a renewal process with interrenewal pdf $f(x)$.

Now, suppose that at a certain time instant X_1 (=S_1), the first *random* packet loss event occurs. Denote the window size at that instant by W_1. When the source detects this loss (by the arrival of duplicate ACKs for the case of TCP-R), the window size is halved. The window size now increases as depicted earlier (the window size starts from $W_1/2$ and increases till w_p, at which time a buffer overflow takes place and $W(t')$ is set to $w_p/2$, and so on) until another random packet loss takes place at a random time instant $S_2 = X_1 + X_2$. Denote the window size at this time (the time of the second loss) by W_2.

In what follows, we call one period from $w_p/2$ till w_p the free-running period or the 'typical' cycle (i.e. free from random loss effects). Note that the second

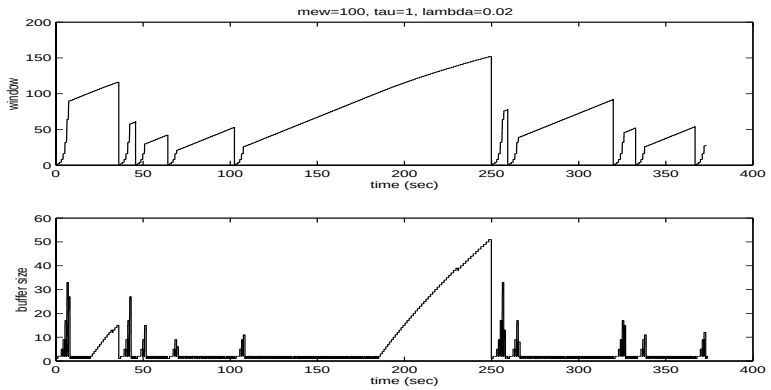

FIGURE 5. Window and buffer size evolution in with random loss of packets for a TCP-T session with $\mu = 100$, $\tau = 1.0$, $\beta = 0.8$ and $\lambda = 0.02$

random loss event can happen before the occurrence of any 'typical' cycles. The window size $W(t')$ is a semi-Markovian stochastic process, because the window size evolution after a random loss (except for its starting value which is half of that just before the random loss) is statistically independent from the window size evolution before the random loss. Further, since $\{X_1, X_2, \dots\}$ are independent and identically distributed (IID), the window sizes $\{W_1, W_2, \dots\}$ (window sizes just before the random loss) form a finite state Markov Chain (i.e. the embedded Markov Chain of the semi-Markov process $W(t')$) [**9, 10**].

3.2. Analysis Assumptions and Notations. With the above model of random loss, two quantities associated with the just defined Markov Chain are of interest
(1) $E[N|W_1 = w_1]$, the expected number of packets successfully transmitted before another random packet loss occurs, given that the most recent random loss took place at w_1; (2) The conditional probability $P[W_2 = w_2|W_1 = w_1]$ (denoted for convenience by P ; the probability that the next *random* loss takes place at $W_2 = w_2$ given that the previous *random* loss took place at $W_1 = w_1$.

Before we attempt to evaluate the above two quantities of interest, we make an approximation for TCP-R, similar to the approximation previously invoked for the channels without random loss. We ignore the first cycle of TCP-R and assume that the TCP session starts with window size $w_p/2$ instead of starting with a window size of 1. This approximation should have a negligible effect on the average throughput, due to two reasons ; (1) A source with an infinite infinite number of packets was assumed; hence the transient behavior (slow start) at the beginning of the connection is expected to be negligible, even for the case of random loss; (2) The duration as well as the number of packets sent during this slow start phase is low (recall that slow start is actually 'fast'). In other words, this transient behavior disappears very fast. We comment on the effect of this approximation in the results section following the analysis.

In the analysis that follows, two ranges of β are considered separately, $\beta < 1$ and $\beta > 1$, and expressions for $E[N|W_1]$ and $P[W_2|W_1]$ are found for each of the two ranges.

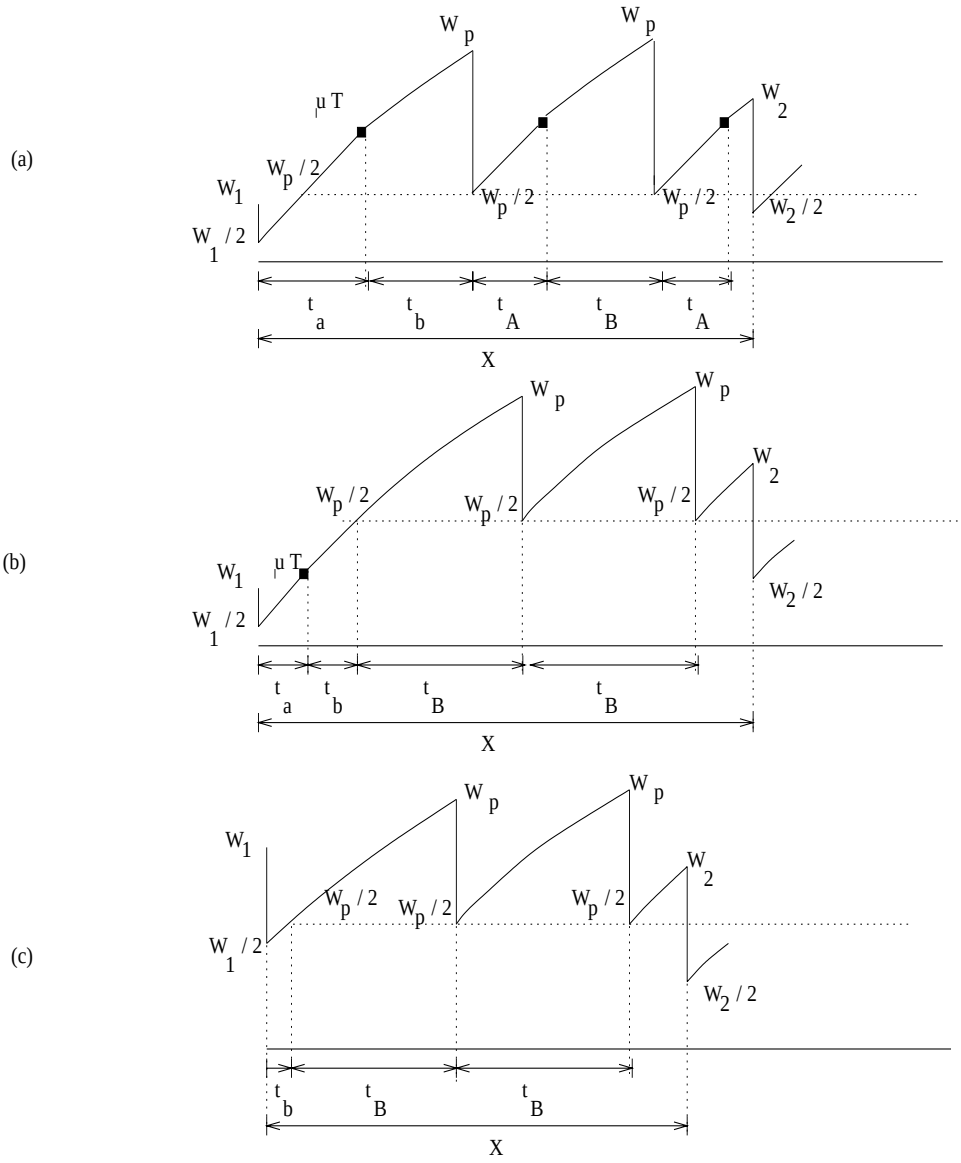

FIGURE 6. A sketch of sample functions of window size in the presence of random loss for TCP-R ; (a) $\beta < 1$ ($\mu T > w_p/2 > W_1/2$), (b) $\beta > 1$ ($w_p/2 > \mu T$) and $W_1/2 < \mu T$ and (c) $\beta > 1$ ($w_p/2 > \mu T$) and $W_1/2 > \mu T$

Define,
N_a is the number of packets sent during the linear congestion avoidance phase in the first TCP cycle just after a random packet loss at a window size $W_1 = w_1$.
N_b is the same as N_a but for the quadratic phase.
N_A is the number of packets sent during the linear congestion avoidance phase of one of the typical cycles.

N_B is the same as N_A but for the quadratic phase.
$N_p = N_A + N_B$ is the number of packets sent in a typical cycle.
The corresponding durations of time where the above number of packets is transmitted (time is counted since the beginning of the phase referenced) are t_a, t_b, t_A, and t_B respectiveley. Define,

$$t_1 = t_a + t_b,$$

$$t_p = t_A + t_B.$$

The values of N and t with capitalized subscripts (A, B) do not depend on W_1 and can be computed using (2.3)-(2.6) as in the case of channels without random packet loss. On the other hand, the values with small letter subscripts (a, b) refer to the cycle of TCP-R just after a random loss at a window size W_1. Those values depend on the value of W_1, and can be computed also by (2.3)-(2.6) by substituting by the appropriate initial and final values of the window size as shown in Figure 6.

For notaional convenience, define,

$$\begin{aligned}
T_a(n) &= T/2\sqrt{w_1^2 + 8(n+1)} - T/2w_1 \\
T_{b_1}(n) &= t_a + (n+1)/\mu \\
T_{b_2}(n) &= (n+1)/\mu \\
T_A(n) &= t_1 + jt_p + T/2\sqrt{w_p^2 + 8(n+1)} - T/2w_p \\
T_{B_1}(n) &= t_1 + t_A + jt_p + (n+1)/\mu \\
T_{B_2}(n) &= t_1 + jt_p + (n+1)/\mu \\
T_a(w_1, w_2) &= T(w_2 - w_1/2) \\
T_{b_1}(w_1, w_2) &= t_a + \frac{w_2^2 - (\mu T)^2}{2\mu} \\
T_{b_2}(w_1, w_2) &= \frac{w_2^2 - (w_1/2)^2}{2\mu} \\
T_A(w_1, w_2) &= t_1 + jt_p + T(w_2 - w_p/2) \\
T_{B_1}(w_1, w_2) &= t_1 + jt_p + t_A + \frac{w_2^2 - (\mu T)^2}{2\mu} \\
T_{B_2}(w_1, w_2) &= t_1 + jt_p + \frac{w_2^2 - (w_p/2)^2}{2\mu}
\end{aligned}$$

where, $0 \leq n < \infty$ is an incremental counter that is initialized at the beginning of each phase (linear or quadratic) and $0 \leq j < \infty$ is an invremental counter that refers to the $j + 1^{st}$ typical cycle after the first atypical cycle following a random packet loss.

The above values represent absolute times of occurances of certain events assuming that a packet loss takes place at a window size $W_1 = w_1$ and no subsequent random packet losses take place. Hence, these times, conditioned on knowing the value of w_1, are deterministic.

For example, $T_A(n)$ represents the time at which, assuming the last random packet loss took place at $W_1 = w_1$, the transmission of packet n will take place in the $(j + 1^{st})$ cycle after the first atypical cycle. $T_A(w_1, w_2)$ represents the time at which the window size will reach w_2 in the $(j + 1^{st})$ cycle after the first atypical cycle.

Finally, the subscript $_1$ for T_b and T_B is used when $\beta < 1$. When $\beta > 1$, both parameters, with $_1$ and $_2$ subscripts will be used, depending on $w_1 < 2\mu T$ or $w_1 > 2\mu T$, as shown in Figure 6.

3.3. $\beta < 1$. $E[N|W_1]$ can be written in terms of the complementary distribution function of N as

$$E[N|W_1 = w_1] = \sum_{n=0}^{\infty} Pr[N > n|W_1 = w_1] \tag{3.1}$$

Using (2.3)-(2.6) and Figure 6, the summation in (3.1) can be written in terms of the p.d.f. of X as a sum of terms of the form $F(a)$ where $F(x)$ is the complementary distribution function of X and $F(a) = Pr[X > a]$.

$$\begin{aligned} E[N|W_1 = w_1] \quad = \quad & \sum_{n=0}^{N_a-1} F[T_a(n)] + \sum_{n=0}^{N_b-1} F[T_{b_1}(n)] \\ & + \sum_{j=0}^{\infty} \sum_{n=0}^{N_A-1} F[T_A(n)] \\ & + \sum_{j=0}^{\infty} \sum_{n=0}^{N_B-1} F[T_{B_1}(n)] \end{aligned} \tag{3.2}$$

In a similar fashion, $P[W_2|W_1]$ can be written in terms of the complementary distribution function $F(x)$. Let $F(a,b) = F(a) - F(b)$. Then by using (2.3)-(2.6) (refer to Figure 6(a) for notations),

$$P = \begin{cases} 0 & 0 < w_2 < w_1/2 \\ F(T_a(w_1,w_2), T_a(w_1,w_2+1)) & w_1/2 < w_2 < w_p/2 \\ F(T_a(w_1,w_2), T_a(w_1,w_2+1)) & \\ + \sum_{j=0}^{\infty} F(T_A(w_1,w_2), T_A(w_1,w_2+1)) & w_p/2 < w_2 < \mu T \\ F(T_{b_1}(w_1,w_2), T_{b_1}(w_1,w_2+1)) & \\ + \sum_{j=0}^{\infty} F(T_{B_1}(w_1,w_2), T_{B_1}(w_1,w_2+1)) & \mu T < w_2 < w_p \end{cases} \tag{3.3}$$

Some remarks on the derivation above are appropriate. First, notice that when a random loss happens at $W_1 = w_1$, the window size immediately after the loss is set to $w_1/2$; then the TCP algorithm increases the window size, possibly passing through events of buffer overflow (depending on the time to next random loss), until another random loss occurs. This means that the window size in the interval between the two random loss events never takes a value less than $w_1/2$. This explains the value of $P[W_2|W_1]$ being zero for $0 < w_2 < w_1/2$. Second, if the next random loss takes place at $W_2 = w_2$ such that $w_1/2 < w_2 < w_p/2$, then this is equivalent to the event that the next random loss takes place somewhere between the time it reaches w_2 (as described by (2.3)) and the time it should have left w_2 to increase to $w_2 + 1$. This explains the second term. Finally, note that if the next random loss takes place at a w_2 such that $w_p/2 < w_2 < w_p$, it implies that the loss occurs somewhere between the time it reaches w_2 (as described by (2.3) and (2.5)) and the time the size is incremented from w_2 to $w_2 + 1$ in the first cycle just after a random loss at $W_1 = w_1$, or in a subsequent cycle that is $j \geq 0$ typical cycles away

from the first cycle after the random loss. This explains the last two terms in the previous equation.

3.4. $\beta > 1$. As for the case of $\beta < 1$, using (2.3)-(2.6) and Figure 6, the summation in (3.1) can be written in terms of the p.d.f. of X. However, we have to differentiate between the case when $W_1/2 < \mu T$ and $W_1/2 > \mu T$. Recall for the non-random-loss case, $w_p/2 > \mu T$, and hence the typical cycles will always consist of a quadratic evolution between $w_p/2$ and w_p. However, for the first cycle after the loss occurring at $W_1 = w_1$, the cycle starts with window size $w_1/2$. If $w_1/2 < \mu T$, the window size will have a linear growth from $w_1/2$ until μT and then a quadratic evolution (unless another random loss takes place) until w_p. On the other hand, if $w_1/2 > \mu T$, the window size will have a quadratic growth from $w_1/2$ without going into a linear phase first. Figure 6(b) and 6(c) show a schematic of a sample function of the two cases.

1. $w_1 < 2\mu T$

$$\begin{aligned} E[N|W_1 = w_1] &= \sum_{n=0}^{N_a-1} F(T_a(n)) + \sum_{n=0}^{N_b-1} F(T_{b_1}(n)) \\ &\quad + \sum_{j=0}^{\infty} \sum_{n=0}^{N_p-1} F(T_{B_2}(n)) \end{aligned} \tag{3.4}$$

$$P = \begin{cases} 0 & 0 < w_2 < w_1/2 \\ F(T_a(w_1, w_2), T_a(w_1, w_2+1)) & w_1/2 < w_2 < \mu T \\ F(T_{b_1}(w_1, w_2), T_{b_1}(w_1, w_2+1)) & \mu T < w_2 < w_p/2 \\ \sum_{j=0}^{\infty} F(T_{B_2}(w_1, w_2), T_{B_2}(w_1, w_2+1)) & w_p/2 < w_2 < w_p \end{cases} \tag{3.5}$$

2. $w_1 > 2\mu T$

$$\begin{aligned} E[N|W_1 = w_1] &= \sum_{n=0}^{N_b-1} F(T_{b_2}(n)) \\ &\quad + \sum_{j=0}^{\infty} \sum_{n=0}^{N_p-1} F(T_{B_2}(n)) \end{aligned} \tag{3.6}$$

$$P = \begin{cases} 0 & 0 < w_2 < w_1/2 \\ F(T_{b_2}(w_1, w_2), T_{b_2}(w_1, w_2+1)) & w_1/2 < w_2 < w_p/2 \\ \sum_{j=0}^{\infty} F(T_{B_2}(w_1, w_2), T_{B_2}(w_1, w_2+1)) & w_p/2 < w_2 < w_p \end{cases} \tag{3.7}$$

3.5. Steady State Distribution of the Window Size. The average packet transmission rate R of a TCP session under the current model is given by

$$R = \frac{E[N]}{E[X]} \tag{3.8}$$

where $E[N]$ is the average number of packets successfully sent in an inter-loss duration, and E[X] is the average time between two random losses (which is equal to $1/\lambda$).

$E[N]$ is given by

$$E[N] = \sum_{W=0}^{w_p} E[N|W]\pi(W) \tag{3.9}$$

where π is the steady state distribution of the MC.

It does not appear feasible that a closed-form expression for π given $P[W_2|W_1]$ can be found. Accordingly, MATLAB^{TM} routines for solving the eigenvalue problem were used to compute π for different values of λ, μ, τ and B, assuming an exponential distribution for the inter-loss times in the following section.

4. Analysis and Simulation for Exponential Random Loss

For an exponential inter-loss distribution with average rate λ, $F(a) = e^{-\lambda a}$ and $F(a,b) = e^{-\lambda a} - e^{-\lambda b}$. By substituting in (3.2)-(3.7), we obtain

4.1. $\beta < 1$**.**

$$\begin{aligned} E[N|W_1 = w_1] &= \sum_{n=0}^{n_a-1} e^{-\lambda(T/2)\sqrt{w_1^2+8(n+1)}-w_1} \\ &\quad + \frac{e^{-\lambda t_1}}{1-e^{-\lambda t_p}} \sum_{n=0}^{n_A-1} e^{-\lambda(T/2)\sqrt{w_p^2+8(n+1)}-w_p} \\ &\quad + e^{-\lambda/\mu} \frac{1-e^{-(\lambda/\mu)n_B}}{1-e^{-(\lambda/\mu)}} (e^{-\lambda t_a} + \frac{e^{-\lambda t_1}}{1-e^{-\lambda t_p}} e^{-\lambda t_A}) \end{aligned} \tag{4.1}$$

(4.2)

$$P = \begin{cases} 0 & 0 < w_2 < w_1/2 \\ e^{-\lambda T(w_2 - w_1/2)}(1-e^{-\lambda T}) & w_1/2 < w_2 < w_p/2 \\ e^{-\lambda T w_2}(1-e^{-\lambda T})(e^{\lambda T w_1/2} + \frac{e^{-\lambda(t_1 - w_p/2)}}{1-e^{-\lambda t_p}}) & w_p/2 < w_2 < \mu T \\ e^{\lambda \frac{(\mu T)^2}{2\mu}}(e^{-\lambda w_2^2} - e^{-\lambda(w_2+1)^2})(e^{-\lambda t_a} + \frac{e^{-\lambda(t_1+t_A)}}{1-e^{-\lambda t_p}}) & \mu T < w_2 < w_p \end{cases}$$

4.2. $\beta > 1$**.**

1. $w_1 < 2\mu T$

$$\begin{aligned} E[N|W_1 = w_1] &= \sum_{n=0}^{n_a-1} e^{-\lambda(T/2)(\sqrt{w_1^2+8(n+1)}-w_1)} \\ &\quad + \frac{e^{-\lambda(t_a+1/\mu)}}{1-e^{-\lambda/\mu}} (1 - e^{-(\lambda/\mu)n_b} + e^{-\lambda t_b} \frac{1-e^{-(\lambda/\mu)n_p}}{1-e^{-\lambda t_p}}) \end{aligned} \tag{4.3}$$

$$P = \begin{cases} 0 & 0 < w_2 < w_1/2 \\ e^{-\lambda T(w_2 - w_1/2)}(1-e^{-\lambda T}) & w_1/2 < w_2 < \mu T \\ e^{-\lambda(t_a - \frac{(\mu T)^2}{2\mu})}(e^{-\lambda w_2^2} - e^{-\lambda(w_2+1)^2}) & \mu T < w_2 < w_p/2 \\ \frac{e^{-\lambda(t_1 - \frac{(w_p/2)^2}{2\mu})}}{1-e^{-\lambda t_p}}(e^{-\lambda w_2^2} - e^{-\lambda(w_2+1)^2}) & \mu T < w_2 < w_p/2 \end{cases} \tag{4.4}$$

2. $w_1 > 2\mu T$

$$E[N|W_1 = w_1] = \frac{e^{-\lambda/\mu}}{1-e^{-\lambda/\mu}} (1 - e^{-(\lambda/\mu)n_b}) + e^{-\lambda t_b} \frac{1-e^{(-\lambda/\mu)n_p}}{1-e^{-\lambda t_p}}) \tag{4.5}$$

$$P = \begin{cases} 0 & 0 < w_2 < w_1/2 \\ e^{\lambda \frac{(w_1/2)^2}{2\mu}}(e^{-\lambda w_2^2} - e^{-\lambda(w_2+1)^2}) & w_1/2 < w_2 < w_p/2 \\ \frac{e^{-\lambda(t_b - \frac{(w_p/2)^2}{2\mu})}}{1-e^{-\lambda t_p}}(e^{-\lambda w_2^2} - e^{-\lambda(w_2+1)^2}) & w_p/2 < w_2 < w_p \end{cases} \tag{4.6}$$

The average throughput for different values of μ, τ, β and λ are shown in Figures 7 and 8, and discussed in the following section. Note that the above expressions for $E[N|W_1]$ simplify in the extreme cases of $\lambda \to 0$ (non-random loss) and $\lambda \to \infty$ (extremely heavy loss rate). Taking the limit of $\lambda E[N|W_1]$ as $\lambda \to 0$ yields n_p/t_p for $\beta < 1$ (where $n_p = n_A + n_B$) and μ for $\beta > 1$ which agrees with the expressions deduced in [**3**] and summarized earlier in (2.7)-(2.10). Taking the limit of $\lambda E[N|W_1]$ as $\lambda \to \infty$ yields zero as expected.

In the simulations, we considered the same set-up described in the system model in Section 2 *without approximation*. Recall that in the analysis, we approximated the TCP window evolution by neglecting the slow start phase which is expected to contribute to any (small) deviation between the analysis and the simulation results in Figures 7, 8.

5. Results and Concluding Remarks

The main results that can be deduced from the throughput behavior depicted in Figures 7, 8 are:

(1) Consider a link with a given loss rate λ and a bandwidth-delay product $\mu\tau$. The results show that increasing the buffer size (i.e. increasing β) does *not* always increase the throughput. For channels with high loss rate, increasing the buffer size has no positive effect on the throughput; however for channels with low loss rates, increasing the buffer size increases the throughput considerably.

(2) For low loss rates, faster channels (higher μ) have higher throughput. However (contrary to what may be expected) for moderate to high loss rates, slower channels have higher throughput. The explanation for this is simple though perhaps not transparent. Recall that for channels without random loss, the throughput is given by $\frac{n_p/t_p}{\mu}$. For channels with random loss, the throughput is given by $\frac{\lambda E[N]}{\mu}$. The expression in the numerator is the average transmission rate. Now, for the case of no random loss, increasing μ increases n_p significantly and hence the average transmission rate as well as the throughput increase. Similarly, for low random loss rates, increasing μ increases $E[N]$ significantly and hence both average rate and throughput increase. On the other hand, for moderate to high loss rates, increasing μ does not increase the number of packets successfully transmitted proportionately (due to the effect of random loss); hence the average transmission rate increases but the throughput actually decreases.

One practical interpretation of this result for the Internet relates to a user's dial-up modem connection to a server. Purchasing a faster modem would increase the average transmission rate, but may not be economically justifiable in the case of moderate-to-high loss rate channels since the proportion of the used bandwidth (i.e, throughput) for the new faster modem is less than that for the slower (and hence, less expensive) one.

(3) The results from the analysis of the proposed model of random loss matches closely with the simulations. However, the effect of neglecting the slow start phase at the beginning of a TCP-R session results in a deviation between the simulation and analysis results. Note the following : (a) For a given channel (i.e. bandwidth-delay product), the deviation (between the simulation and the analysis results) for low loss rates is small. This is because the slow start phase duration is sufficiently small such that the window size reaches $w_p/2$ in a very short time (compared to the average time to the first random loss) corroborating our approximation. in the

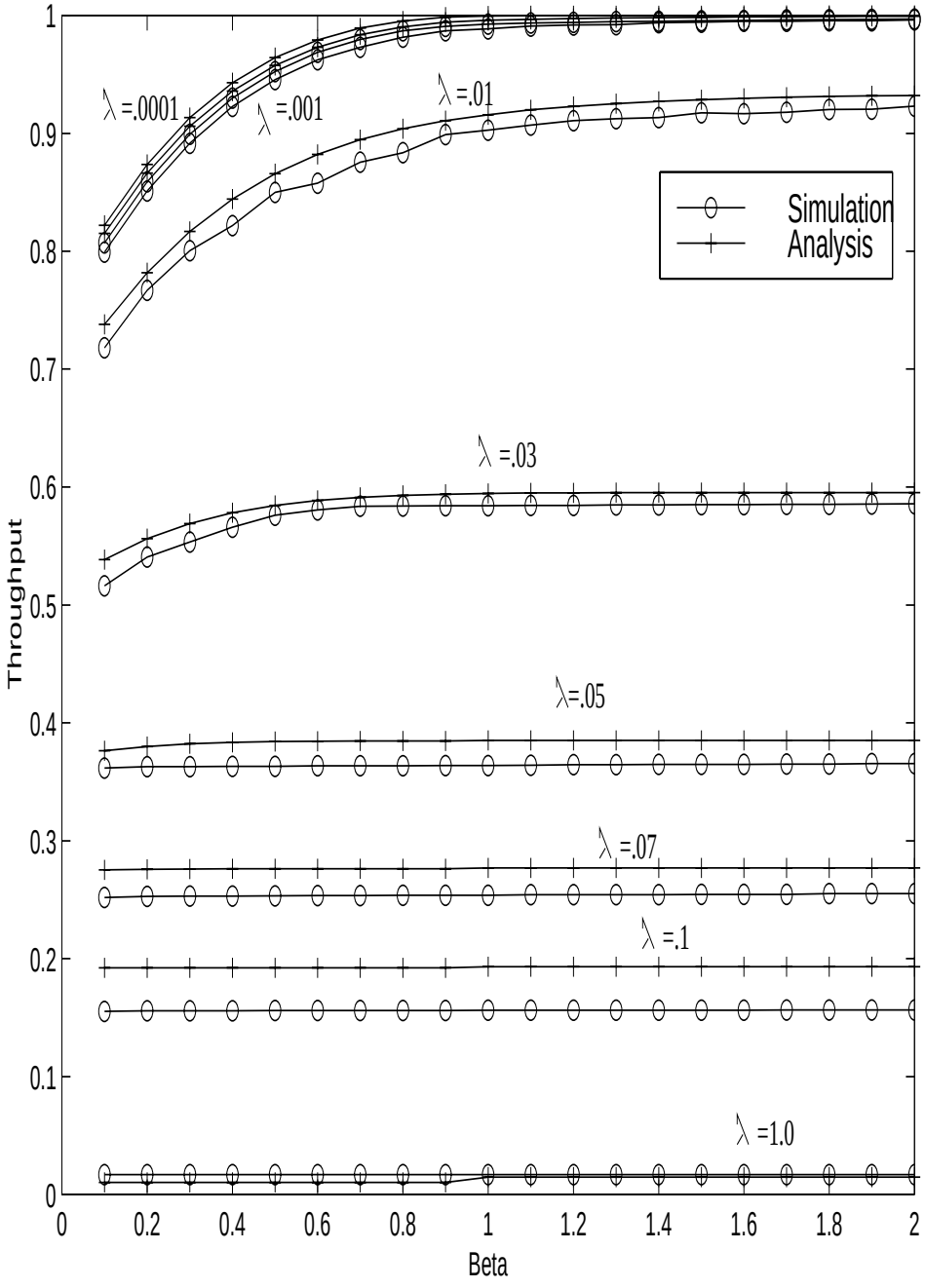

FIGURE 7. $\mu = 100, \tau = 1.0$

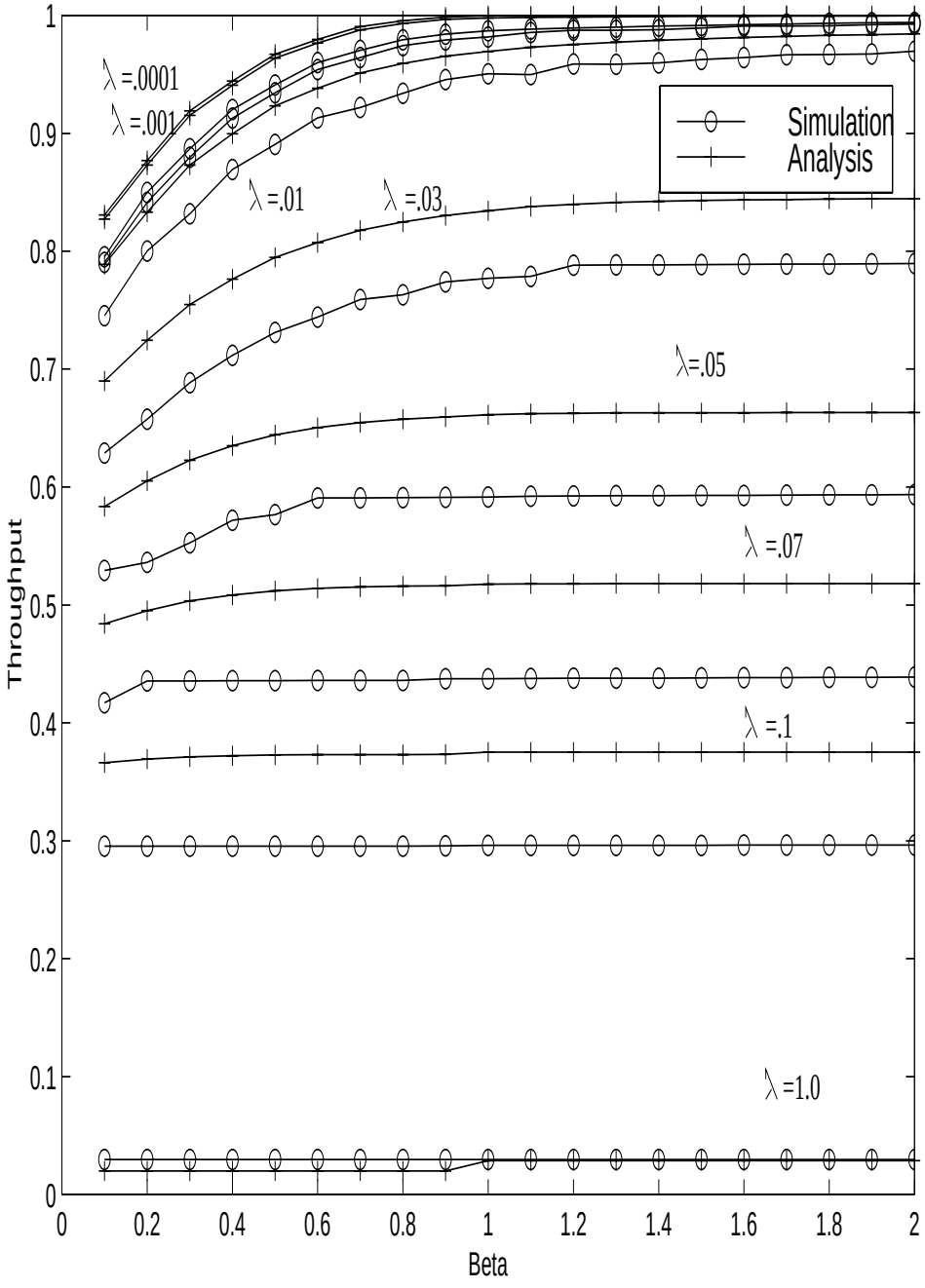

FIGURE 8. $\mu = 50, \tau = 1.0$

analysis. As λ increases, so does the deviation since it becomes increasingly probable that the first random loss takes place early in the slow start phase, thereby precipitating a congestion avoidance phase with an initial window size that is considerably smaller than $\frac{w_p}{2}$ as assumed in the approximation. Consequently, the simulated throughput (on the average) is lower than that predicted by analysis, most noticeably for moderate values of random loss. For heavy loss rates, the deviation decreases again since the approximate window size quickly decreases from its starting value of $w_p/2$ to that (i.e., the true) in the simulations.

(b) The deviation is larger for lower bandwidth-delay product. Recall that the reason for neglecting the slow start phase in the first cycle of TCP-R is that the duration as well as *the number of packets sent* during this phase is low. However, this becomes increasingly inaccurate as w_p decreases. Since $w_p = (\mu\tau + 1)(\beta + 1)$, decreasing $\mu\tau$ (the band-width delay product) decreases w_p and hence increases the deviation for the same value of β.

References

[1] S. Shenker, L. Zhang, and D. D. Clark, "Some observations on the dynamics of a congestion control algorithm," *Computer Communications Review*, pp. 30-39, Oct 1990.

[2] T. J. Ott, J. H. B. Kemperman, and M. Mathis, "The stationary behavior of ideal TCP congestion avoidance," 1996.

[3] T. V. Lakshman, and U. Madhow, "The performance of TCP/IP for networks with high bandwidth-delay products and Random loss," *IEEE/ACM Transactions on Networking*, vol. 5, no. 3, June 1997.

[4] T. V. Lakshman, and U. Madhow, "Window-based error recovery and flow control with a slow acknowledgment channel: a study of TCP/IP performance," *Proceedings of IEEE INFOCOM '97*.

[5] S. Savari, and E. Telatar, "The behavior of certain stochastic processes arising in window protocols," *preprint*, July 1998.

[6] V. Jacobson, "Congestion avoidance and control," *Proceedings of ACM SIGCOMM ' 88*, August 1988.

[7] V. Jacobson, "Modified TCP congestion avoidance algorithm," *Message to end2end-interest mailing list*, April 1990, *URL ftp://ftp.ee.lbl.gov/email/vanj.90apr30.txt*.

[8] J. Bolot, "End-to-end packet delay and loss behavior in the internet," *Proc. ACM SIGCOMM'93*.

[9] S. M. Ross, "Stochastic processes," New York: Wiley, 1983.

[10] R. G. Gallager, "Discrete stochastic processes," Boston : Kluwer, 1996.

[11] S. Floyd, "TCP and successive fast retransmits," February 1995, *URL ftp://ftp.ee.lbl.gov/email/vanj.90apr30.txt*.

Department of Electrical Engineering, University of Washington, Box 352500, Seattle WA 98195

E-mail address: hussein@ee.washington.edu

Department of Electrical Engineering, University of Washington, Box 352500, Seattle WA 98195

E-mail address: azizoglu@ee.washington.edu

Department of Electrical Engineering, University of Washington, Box 352500, Seattle WA 98195

E-mail address: roy@ee.washington.edu