

# Modeling TCP Traffic with Session Dynamics - Many Sources Asymptotics under ECN/RED Gateways

P. Tinnakornsrisuphap, R. J. La, and A. M. Makowski

Department of Electrical and Computer Engineering  
University of Maryland, College Park, MD, 20742, USA.

Short-lived TCP traffic (*e.g.*, Web mice) composes the majority of the current Internet traffic. Accurate traffic modeling of a large number of short-lived TCP flows is extremely difficult due to (i) the complex interactions between session, transport and network layers; and (ii) the state space explosion when the number of flows is large. As a result, ad-hoc assumptions are often required for the analysis to be tractable in a given regime.

We introduce a stochastic model of a bottleneck ECN/RED gateway under a large number of competing TCP flows, and show that as the number of flows becomes large, the buffer dynamics and aggregate traffic simplify and can be accurately described by simple stochastic recursions. These recursions can be evaluated independently of the number of flows, thereby leading to a scalable traffic model. Furthermore, the limiting model is consistent with other previously proposed models in their respective regime.

## 1. INTRODUCTION

Due to the growing size and popularity of the Internet, Internet traffic modeling has become an important research area. Internet traffic consists of many heterogeneous traffic sources, the majority of which utilize the TCP congestion control mechanism [4]. Some applications, such as FTP and Telnet, are relatively long-lived, while others are typically short-lived, *e.g.*, Web browsing.

Characterizing and modeling TCP traffic yield an understanding of the interactions between the transport layer (TCP) and the network layer. Such interactions are well understood in the context of a single long-lived TCP flow. However, when the number of TCP flows becomes large, straightforward modeling results in models that are typically not scalable because of the exploding size of the state space required to describe all flows. Furthermore, for short-lived TCP flows, the session layer constitutes an additional layer in the dynamics, and needs to be taken into account.

The existing literature on short-lived TCP traffic modeling usually skirts these two major obstacles by relying on ad-hoc assumptions, which causes the model to be accurate only in certain regimes. Holot *et al.* model short-lived TCP flows as exponential pulses that occur according to a Poisson process [3], and characterize their TCP windows through *shot noise* processes. This model implicitly assumes relatively low congestion levels in that short-lived flows last only a few round-trip times and do not experience packet drops or marks. Moreover, flows are always in either congestion avoidance (long-lived connections)

or slow start (short-lived connections), and do not transit from one state to the other. In other words, the session dynamics are not modeled *explicitly* with connections arriving and leaving the network after transfers are completed. A similar approach to modeling short-lived flows is also taken in [7].

At the other end of the spectrum, Kherani and Kumar [5] suggest that as the capacity at the bottleneck queue becomes very small, this queue can be accurately described as a processor sharing queue. When the capacity is large, however, this processor sharing model becomes less accurate as newly arrived TCP flows cannot fully utilize their allocated bandwidth. In fact, in the large capacity regime these short-lived flows may terminate even before they can increase their transmission rates to fully utilize their allocated bandwidth due to slow start.

The shortcomings of these models suggest a need for a *unified* model that is accurate in *all* regimes, instead of being restricted to a specific regime. Since the number of connections that share a bottleneck link is likely to be large, we follow the approach in [10], which considers “macroscale” modeling of aggregate TCP flows competing for the capacity of a bottleneck link. Macroscale TCP models can be developed by systematically applying limit theorems to derive a limiting traffic model when the number of TCP flows becomes large. The potential benefits of doing so are three-fold. First, model simplifications (with the promise of scalability) typically occur when applying limit theorems, with irrelevant details filtered out without relying on ad-hoc assumptions. Second, limit theorems are central to the modern theory of probability, and as such have been the focus of a huge literature that contains a large number of results and techniques. Hence, it is reasonable to expect the existence of suitable limit theorems (under very weak assumptions) which can be applied to the situation of interest. Finally, in the networking context, resource allocation problems are of greatest interest in networks operating at high utilization, *e.g.*, when the number of users is large. In such a scenario, the limit behavior will become increasingly accurate as the number of users increases.

In this paper we extend the model in [10] by incorporating *explicitly* an additional layer of dynamics, namely the session layer, with TCP in the transport layer and RED gateway [2] with ECN [1] capability in the network layer. As the number of sessions increases, we show that (i) the queue size per session and the workload per session brought in during a round-trip time converge to deterministic processes; and (ii) the sessions become asymptotically independent, indicating that the RED mechanism does alleviate the synchronization problem among the connections.

The paper is organized as follows: The model and the dynamics of network, transport, and session layers are described in Section 2. The asymptotic results are presented in Section 3. Section 4 gives a brief discussion of the results and a comparison with previously proposed models. Section 5 concludes the paper with suggestions for future work.

Some words on the notation in use: Equivalence in law or in distribution between random variables (rvs) is denoted by  $=_{st}$ . The indicator function of an event  $A$  is given by  $\mathbf{1}\{A\}$ , and we use  $\xrightarrow{P}_n$  (resp.  $\implies_n$ ) to denote convergence in probability (resp. weak convergence or convergence in distribution) with  $n$  going to infinity. For scalars  $a$  and  $b$  we write  $a \vee b = \max(a, b)$ .

## 2. THE MODEL

Our model captures three layers of dynamics, namely the network, transport, and session layers, which interact with each other through mechanisms to be specified shortly. At the lowest level, the network is simplified to be a single bottleneck router, more specifically, a RED gateway with ECN capability enabled. The traffic injected into the network is shaped by the TCP congestion control mechanism in the transport layer, which reacts to the marks from the network. Each TCP connection is initiated by a session, such session being either active or idle. If a session is busy, a file or an object is transferred through a TCP connection. A busy period for a session lasts until it no longer has any data to transfer, at which time it goes idle. The duration of an idle period is random and represents the idle time between consecutive file transmissions. When a new file/object to be transferred arrives, the session becomes active again and sets up a new TCP connection. We now give detailed descriptions of each of the three layers of the model and of their interactions.

Throughout this section let  $N$  be a fixed positive integer, and let  $\mathcal{N} = \{1, \dots, N\}$  be the set of sessions that share a bottleneck RED gateway. Time is slotted in contiguous time slots of duration equal to the round-trip delay of TCP connections. Hence, there is no loss of generality in taking time to be discrete with  $t = 0, 1, \dots$  indicating the start of time slot  $[t, t + 1)$ .

We fix  $t = 0, 1, \dots$ , and we write  $X^{(N)}$  to indicate the explicit dependence of the quantity  $X$  on the number  $N$  of sessions.

### 2.1. Session dynamics

Each session  $i \in \mathcal{N}$  is either active or idle. A session idle at the beginning of time slot  $[t, t + 1)$  has no packet to transmit in that time slot. An idle session in time slot  $[t, t + 1)$  becomes active at the beginning of time slot  $[t + 1, t + 2)$  with probability  $P_{ar}$ ,  $0 < P_{ar} < 1$ , independently of the past. In other words, the duration of an idle period is *geometrically* distributed with parameter  $P_{ar}$  (hence with mean  $1/P_{ar}$ ). This attempts to capture the dynamics of connection arrivals, where the interarrival times are reported to be exponentially distributed [9].<sup>1</sup> Let  $\{U_i(t), i \in \mathcal{N}; t = 0, 1, \dots\}$  be a collection of i.i.d. rvs uniformly distributed on  $[0, 1]$ , and let  $\mathbf{1}\{U_i(t + 1) \leq P_{ar}\}$  be the indicator function of the event that a new file/object arrives in the time slot  $[t + 1, t + 2)$  for an idle session  $i$ .

Let  $\{F_i(t), i \in \mathcal{N}; t = 0, 1, \dots\}$  be a collection of i.i.d. non-negative integer-valued rvs distributed according to a general probability mass function (pmf)  $F$  on  $\{1, 2, \dots\}$ . The workload of a connection for session  $i$  that becomes active at the beginning of time slot  $[t, t + 1)$  is given by  $F_i(t)$ . This workload represents the *total* number of TCP segments<sup>2</sup> the connection brings in before it is torn down. Thus, in the event a given TCP connection is used to transfer more than one object, this workload variable  $F_i(t)$  represents the total number of TCP segments brought in by all these objects. We denote by  $X_i(t)$  the remaining workload (expressed in packets) of connection  $i$  at the beginning of time slot

<sup>1</sup>Recall that an exponential rv  $X$  with parameter  $\alpha$  can be approximated by  $\lceil X \rceil$ , which is a geometric rv with parameter  $p = 1 - e^{-\alpha}$ .

<sup>2</sup>In this model each TCP segment is transmitted as a separate packet.

$[t, t + 1)$ . Clearly,  $X_i(t) = 0$  if session  $i$  is idle during  $[t, t + 1)$ . The evolution of  $X_i(t)$  is then given by the recursion

$$X_i^{(N)}(t + 1) = \mathbf{1} \{X_i^{(N)}(t) > 0\} (X_i^{(N)}(t) - A_i^{(N)}(t)) + \mathbf{1} \{X_i^{(N)}(t) = 0\} \mathbf{1} \{U_i(t + 1) \leq P_{ar}\} F_i(t + 1), \quad (1)$$

where  $A_i^{(N)}(t)$  denotes the number of packets injected into the network by connection  $i$  at the beginning of time slot  $[t, t + 1)$ . This will be explained next.

## 2.2. TCP dynamics

For each  $i \in \mathcal{N}$ , let  $W_i^{(N)}(t)$  be an integer-valued rv that encodes the congestion window size (in packets) at the beginning of time slot  $[t, t + 1)$ . We assume that the rv  $W_i^{(N)}(t)$  has range  $\{0, 1, \dots, W_{\max}\}$  where  $W_{\max}$  is a finite integer representing the receiver advertised window size of the TCP connection. The congestion window size of an idle session is taken to be zero. When an idle session becomes active at the beginning of time slot  $[t, t + 1)$ , the congestion window size of the TCP connection is set to one at the beginning of time slot  $[t + 1, t + 2)$ . This models one round-trip delay for the three-way handshake. We now describe how the congestion window sizes of active connections evolve. Each TCP source transmits as many of the remaining data packets as allowed by its congestion window in that time slot. In other words, suppose that connection  $i$  has  $X_i^{(N)}(t)$  remaining packets waiting to be transmitted at the beginning of time slot  $[t, t + 1)$ .<sup>3</sup> The number  $A_i^{(N)}(t)$  of packets which connection  $i$  transmits at the beginning of time slot  $[t, t + 1)$  is given by  $A_i^{(N)}(t) = \min(W_i^{(N)}(t), X_i^{(N)}(t))$ .

The congestion control mechanism of TCP operates in one of two different modes, namely *slow start* (SS) and *congestion avoidance* (CA). A new TCP connection starts in SS. While in SS, the congestion window size is doubled every round-trip time until one or more packets are marked. If a mark is received, then the congestion window size is halved and TCP switches to CA. The congestion window size is limited by the receiver advertised window size  $W_{\max}$ . Hence, the congestion window of connection  $i$  in SS evolves according to

$$W_{SS,i}^{(N)}(t + 1) = \min(2W_i^{(N)}(t) \vee 1, W_{\max}) M_i^{(N)}(t + 1) + \lceil \frac{W_i^{(N)}(t)}{2} \rceil (1 - M_i^{(N)}(t + 1)) \quad (2)$$

where  $M_i^{(N)}(t + 1)$  is an indicator function of the event that no packet of connection  $i$  has been marked in time slot  $[t, t + 1)$ , i.e.,  $M_i^{(N)}(t + 1) = 1$  when no packet from session  $i$  is marked in the time slot and  $M_i^{(N)}(t + 1) = 0$  when at least one packet is marked. The marking mechanism is explained in Subsection 2.3.

In CA, the congestion window size in the next time slot  $[t + 1, t + 2)$  is increased by one if no mark is received in time slot  $[t, t + 1)$ , while if one or more packets are marked in time slot  $[t, t + 1)$ , the congestion window in the next time slot is reduced by half. The congestion window size in CA is then given by

$$W_{CA,i}^{(N)}(t + 1) = \min(W_i^{(N)}(t) + 1, W_{\max}) M_i^{(N)}(t + 1) + \lceil \frac{W_i^{(N)}(t)}{2} \rceil (1 - M_i^{(N)}(t + 1)). \quad (3)$$

<sup>3</sup>We refer to a TCP connection of an active session  $i$  by connection  $i$  when there is no confusion.

The  $\{0, 1\}$ -valued rvs  $\{S_i^{(N)}(t), i \in \mathcal{N}\}$  encode the state of TCP connections, with the interpretation that  $S_i^{(N)}(t) = 0$  (resp.  $S_i^{(N)}(t) = 1$ ) if connection  $i$  is in CA (resp. in SS) at the beginning of time slot  $[t, t + 1)$ . Therefore, combining (2) and (3), we see that the complete recursion of the congestion window size can be written as

$$W_i^{(N)}(t + 1) = \mathbf{1} \left\{ X_i^{(N)}(t) > W_i^{(N)}(t) \right\} \\ \times \left( S_i^{(N)}(t) W_{SS,i}^{(N)}(t + 1) + (1 - S_i^{(N)}(t)) W_{CA,i}^{(N)}(t + 1) \right). \quad (4)$$

The first indicator function in (4) is used to reset the congestion window size to zero when session  $i$  runs out of data to transmit and returns to its idle state.

Finally, the evolution of  $S_i^{(N)}(t)$  is given by

$$S_i^{(N)}(t + 1) = \mathbf{1} \left\{ X_i^{(N)}(t) \leq W_i^{(N)}(t) \right\} \\ + \mathbf{1} \left\{ X_i^{(N)}(t) > W_i^{(N)}(t) \right\} S_i^{(N)}(t) M_i^{(N)}(t + 1). \quad (5)$$

This equation can be interpreted as follows. Connection  $i$  is in SS in time slot  $[t + 1, t + 2)$  if either (1) there is no packet left to transmit (so the connection resets) at the beginning of the time slot or (2) the connection was active and in SS in time slot  $[t, t + 1)$  and received no mark in the time slot. From (5) we assume that a new TCP connection in SS is ready to be set up after the previous connection is torn down after finishing its workload, and the new TCP connection becomes active when a new file/object arrives initiating the three-way handshake.

### 2.3. Network dynamics

We now explain how packet marking provides congestion notification to the active TCP connections. The capacity of the bottleneck link is  $NC$  packets/slot for some positive constant  $C$ . The buffer size is assumed infinite so that no packets are dropped due to buffer overflow, and there is no retransmission by TCP due to packet losses at the network layer. Thus, congestion control is achieved solely through the random marking algorithm of the RED gateway.

Let  $Q^{(N)}(t)$  denote the number of packets queued in the buffer at the beginning of time slot  $[t, t + 1)$ . Connection  $i$  injects  $A_i^{(N)}(t)$  packets into the network, and they are put in the buffer at the beginning of time slot  $[t, t + 1)$ . Let the rv  $A^{(N)}(t) := \sum_{i=1}^N A_i^{(N)}(t)$  denote the aggregate number of packets offered to the network by the  $N$  sessions at the beginning of time slot  $[t, t + 1)$ . Hence,  $Q^{(N)}(t) + A^{(N)}(t)$  packets are available for transmission during that time slot. Since the bottleneck link has a capacity of  $NC$  packets/time slot,  $\left[ Q^{(N)}(t) + A^{(N)}(t) - NC \right]^+$  packets will not be served during time slot  $[t, t + 1)$ , and will remain in the buffer. Hence, their transmission is deferred to subsequent time slots. The number of packets in the buffer at the beginning of time slot  $[t + 1, t + 2)$ ,  $Q^{(N)}(t + 1)$ , is therefore given by

$$Q^{(N)}(t + 1) = \left[ Q^{(N)}(t) - NC + A^{(N)}(t) \right]^+. \quad (6)$$

Each incoming packet into the router in time slot  $[t, t + 1)$  is marked with a probability which depends on the queue length at the beginning of time slot  $[t, t + 1)$ , namely

$f^{(N)}(Q^{(N)}(t))$  for some mapping  $f^{(N)} : \mathbb{N} \rightarrow [0, 1]$ . This model approximates the case where the memory of the queue averaging mechanism is long, which is the case for the recommended parameter settings of RED. We represent marking through the  $\{0, 1\}$ -valued rvs  $M_{i,j}^{(N)}(t+1)$  ( $j = 1, \dots, A_i^{(N)}(t)$ ) with  $M_{i,j}^{(N)}(t+1) = 0$  (resp.  $M_{i,j}^{(N)}(t+1) = 1$ ) if the  $j$ th packet from source  $i$  is marked (resp. not marked) in the RED buffer. More concretely, for each  $i \in \mathcal{N}$  and  $j = 1, 2, \dots$ , we write the rv  $M_{i,j}^{(N)}(t+1)$  as the indicator function

$$M_{i,j}^{(N)}(t+1) = \mathbf{1} \{V_{i,j}(t+1) > f^{(N)}(Q^{(N)}(t))\},$$

where the collection of i.i.d.  $[0, 1]$ -uniform rvs  $\{V_{i,j}(t+1), i, j = 1, \dots; t = 0, 1, \dots\}$  are assumed independent of all other rvs introduced so far. The indicator function of the event that no packets from connection  $i$  in time slot  $[t, t+1)$  are marked can then be written as

$$M_i^{(N)}(t+1) = \prod_{j=1}^{A_i^{(N)}(t)} M_{i,j}^{(N)}(t+1). \quad (7)$$

### 3. ASYMPTOTICS

The main result of the paper consists of the asymptotics for the normalized buffer content as the number of sessions becomes large; its proof is given in [11]. This result is discussed under the following assumptions (A1)-(A2):

(A1) There exists a continuous function  $f : \mathbb{R}_+ \rightarrow [0, 1]$  such that for each  $N = 1, 2, \dots$ ,

$$f^{(N)}(x) = f(N^{-1}x), \quad x \geq 0;$$

(A2) For each  $N = 1, 2, \dots$ , the dynamics (1), (4), (5) and (6) start with the initial conditions  $Q^{(N)}(0) = W_i^{(N)}(0) = X_i^{(N)}(0) = 0$ , and  $S_i^{(N)}(0) = 1, i = 1, \dots, N$ .

Assumption (A1) is a structural condition while (A2) is made essentially for technical convenience as it implies that for each  $N$  and all  $t = 0, 1, \dots$ , the rvs  $Z_1^{(N)}(t), \dots, Z_N^{(N)}(t)$  are *exchangeable* where we have written

$$Z_i^{(N)}(t) = (W_i^{(N)}(t), X_i^{(N)}(t), S_i^{(N)}(t)), \quad i = 1, \dots, N; \quad N = 1, 2, \dots \quad (8)$$

Assumption (A2) can be omitted but at the expense of a more cumbersome discussion. Let  $\mathcal{Z}$  denote the range of the rvs defined in (8), namely  $\mathcal{Z} := \{0, 1, \dots, W_{\max}\} \times \mathbb{N} \times \{0, 1\}$ .

**Theorem 1** *Assume that (A1)-(A2) hold. Then, for each  $t = 0, 1, \dots$ , there exist a (non-random) constant  $q(t)$  and a  $\mathcal{Z}$ -valued rv  $Z(t) := (W(t), X(t), S(t))$  such that the following holds: (i) The following convergences*

$$\frac{Q^{(N)}(t)}{N} \xrightarrow{P} q(t) \quad \text{and} \quad Z_1^{(N)}(t) \Rightarrow_N Z(t) \quad (9)$$

*are taking place; (ii) For any bounded function  $g : \mathcal{Z} \rightarrow \mathbb{R}$ , we have*

$$\frac{1}{N} \sum_{i=1}^N g(Z_i^{(N)}(t)) \xrightarrow{P} \mathbf{E}[g(Z(t))]. \quad (10)$$

(iii) For any integer  $I = 1, 2, \dots$ , the triplets  $\{Z_i^{(N)}(t), i = 1, \dots, I\}$  become asymptotically independent as  $N$  becomes large, with

$$\lim_{N \rightarrow \infty} \mathbf{P}[Z_i^{(N)}(t) = z_i \ i = 1, \dots, I] = \prod_{i=1}^I \mathbf{P}[Z(t) = z_i] \quad (11)$$

for any  $z_1, \dots, z_I$  in  $\mathcal{Z}$  with  $z_i = (w_i, x_i, s_i)$  for each  $i = 1, \dots, I$ .

In addition, with initial conditions  $q(0) = W(0) = X(0) = 0$  and  $S(0) = 1$ , it holds that

$$q(t+1) = (q(t) - C + \mathbf{E}[A(t)])^+ \quad \text{where } A(t) = \min(W(t), X(t)). \quad (12)$$

Further, the recurrence

$$\begin{aligned} X(t+1) &=_{st} \mathbf{1}\{X(t) = 0\} \mathbf{1}\{U(t+1) \leq P_{ar}\} F(t+1) \\ &\quad + \mathbf{1}\{X(t) > 0\} (X(t) - A(t)), \end{aligned} \quad (13)$$

$$A(t) =_{st} \min(W(t), X(t)), \quad (14)$$

$$W_{SS}(t+1) =_{st} \min(2W(t) \vee 1, W_{\max}) M(t+1) + \lceil \frac{W(t)}{2} \rceil (1 - M(t+1)), \quad (15)$$

$$W_{CA}(t+1) =_{st} \min(W(t) + 1, W_{\max}) M(t+1) + \lceil \frac{W(t)}{2} \rceil (1 - M(t+1)), \quad (16)$$

$$W(t+1) =_{st} \mathbf{1}\{X(t) > W(t)\} (S(t)W_{SS}(t+1) + (1 - S(t))W_{CA}(t+1)), \quad (17)$$

$$S(t+1) =_{st} \mathbf{1}\{X(t) > W(t)\} + S(t)M(t+1) \mathbf{1}\{X(t) \leq W(t)\} \quad (18)$$

holds in law, where the rv  $F(t+1)$  has distribution  $F$  and is independent of other rvs, and

$$M(t+1) = \mathbf{1}\{V(t+1) \leq (1 - f(q(t)))^{A(t)}\} \quad (19)$$

for i.i.d.  $[0, 1]$ -uniform rvs  $\{U(t+1), V(t+1); t = 0, 1, \dots\}$ .

If the workload distribution  $F$  has a finite second moment, it also holds that

$$\frac{1}{N} \sum_{i=1}^N X_i^{(N)}(t) \xrightarrow{P}_N \mathbf{E}[X(t)]. \quad (20)$$

## 4. DISCUSSION

### 4.1. Macroscale modeling

Theorem 1 shows that the queue length  $Q^{(N)}(t)$  at time  $t$  can be approximated by  $Nq(t)$  with  $q(t)$  determined via a simple deterministic recursion, which is *independent* of the number of sessions. The traffic  $A^{(N)}(t)$  injected into the network during time slot  $[t, t+1)$  can also be approximated by  $N \cdot \mathbf{E}[A(t)]$ . These approximations become more accurate as the number of sessions becomes large, and the computational complexity does not depend on  $N$ . The limiting model is therefore “scalable” as it does not suffer from state space explosion, nor does it require any ad-hoc assumption in the analysis. Theorem 1 also shows that the dependency between each session becomes negligible under a large number of sessions, i.e., “RED breaks the global synchronization when the number of sessions is large.”

Although the sequence  $\{(q(t), Z(t)), t = 0, 1, \dots\}$  is a time-homogeneous Markov chain with values in  $\mathbb{R}_+ \times \mathcal{Z}$ , we do not address here the existence of its steady state regime (when  $t \rightarrow \infty$ ) as complications arise due the deterministic (i.e., degenerate) character of the first component. Nevertheless, the *numerical* calculations for the limiting model are very simple. The number of steps required for the calculations in each time slot is independent of  $N$  [11].

#### 4.2. Limiting cases

Next, we briefly consider the resulting model from Theorem 1 in the regime when  $C$  is either very large or very small with the following assumption:

- (A3) The marking function  $f : \mathbb{R} \rightarrow [0, 1]$  is monotonically increasing with  $f(0) = 0$  and  $\lim_{x \rightarrow \infty} f(x) = 1$ ;

It is easy to see that  $\lim_{C \rightarrow \infty} q(t) = 0$  for all  $t = 0, 1, \dots$ , so that the marking probability per flow also converges to zero from (A3) for all  $t$ . Therefore, each incoming flow will always operate in the slow start (exponential growth) mode and the resulting input traffic into the network is the superposition of (discrete-time) Poisson arrival streams of random number of packets, each of which doubles its window size every round-trip. The aggregate input traffic is therefore similar to the time-reversed shot-noise processes, in agreement with [3].

On the other hand, with  $C \simeq 0$ , the queue will start building up, whence  $\lim_{t \rightarrow \infty} q(t) = \infty$ . Thus, for large  $t$ , all TCP flows (including incoming TCP flows) will experience marking probability close to one from Assumption (A3). All active connections will be able to inject only one packet per round-trip into the network because every packet transmitted will be marked with a probability going to one. As a result, each TCP congestion window size approaches one with high probability. Since the bottleneck router will transmit packets non-selectively, any active flow will receive roughly equal throughput and hence the queue behavior approaches that of processor-sharing, which is in agreement with [5].

#### 4.3. Steady state regime

Using the results from the previous sections, we now carry out a simple steady-state analysis with a few additional reasonable assumptions. Although we have not formally proved that the system converges to steady state for the reasons stated earlier, simulation results indicate that the system does converge to steady state for a large range of initial conditions and parameters. This is illustrated in [11]. In order to facilitate our further analysis we assume the following:

- (A4) The sequence  $\{(q(t), Z(t), M(t+1)), t = 0, 1, \dots\}$  admits a steady state in the sense that  $(q(t), Z(t), M(t+1)) \Rightarrow_t (q^*, Z^*, M^*)$  for some rvs  $(q^*, Z^*, M^*)$  where  $q^*$  is a constant and  $Z^* = (W^*, X^*, S^*)$  and  $M^*$  are rvs taking values in  $\mathcal{Z}$  and  $\{0, 1\}$ , respectively, with

$$M^* =_{st} \mathbf{1} \left\{ V(t+1) \leq (1 - f(q^*))^{A^*} \right\} \quad \text{where} \quad A^* = \min(W^*, X^*).$$

- (A5) Let  $F_{ar}$  be a rv with the distribution  $F$ , representing the initial workload size of a new connection. we assume  $\mathbf{E}[W^*] \ll \mathbf{E}[F_{ar}]$ .

We wish to find the steady-state queue level  $q^*$  as a fixed-point solution to

$$q^* = (q^* - C + \mathbf{E}[A^*])^+ = (q^* - C + \mathbf{E}[\min(W^*, X^*)])^+. \quad (21)$$

Because the window size and workload are both zero when a connection is idle, it holds that

$$\mathbf{E}[A^*] = \mathbf{P}[\text{active}] \mathbf{E}[\min(W^*, X^*) | \text{active}], \quad (22)$$

where the probability  $\mathbf{P}[\text{active}]$  that a session is active in steady state is given by

$$\mathbf{P}[\text{active}] = \frac{\mathbf{E}[\text{connection duration}]}{\mathbf{E}[\text{connection duration}] + \mathbf{E}[\text{idle period}]} \quad (23)$$

by elementary arguments from renewal theory. Next, we note that  $\mathbf{E}[\text{connection duration}] = \mathbf{E}[\mathbf{E}[\text{connection duration} | F_{ar}]]$ , with

$$\mathbf{E}[\text{connection duration} | F_{ar} = x] = \frac{x}{\text{mean throughput}}, \quad x = 0, 1, \dots$$

Here we approximate the average throughput of a connection by the well-known throughput formula  $\frac{K}{\sqrt{f(q^*)}}$  for some constant  $K$  in the interval  $[1, \frac{8}{3}]$  [8,6].<sup>4</sup> We assume  $K = \sqrt{\frac{3}{2}}$  which is shown in [10] to be a reasonable approximation, whence

$$\mathbf{E}[\text{connection duration}] \approx \frac{\mathbf{E}[F_{ar}] \sqrt{f(q^*)}}{K}. \quad (24)$$

Reporting this approximation into (23), we find

$$\mathbf{P}[\text{active}] \approx \frac{\frac{\mathbf{E}[F_{ar}] \sqrt{f(q^*)}}{K}}{\frac{\mathbf{E}[F_{ar}] \sqrt{f(q^*)}}{K} + \frac{1}{P_{ar}}} = \frac{P_{ar} \mathbf{E}[F_{ar}] \sqrt{f(q^*)}}{P_{ar} \mathbf{E}[F_{ar}] \sqrt{f(q^*)} + K}$$

upon using the fact  $\mathbf{E}[\text{idle period}] = P_{ar}^{-1}$ . We approximate  $\mathbf{E}[\min(W^*, X^*) | \text{active}]$  by  $\mathbf{E}[W^* | \text{active}]$ , such an approximation being reasonable under Assumption (A5). Going back to (22) and using again the approximation  $\mathbf{E}[W^* | \text{active}] = \frac{K}{\sqrt{f(q^*)}}$ , we readily

conclude  $\mathbf{E}[A^*] \approx \frac{K P_{ar} \mathbf{E}[F_{ar}]}{P_{ar} \mathbf{E}[F_{ar}] \sqrt{f(q^*)} + K}$ .

If  $f(q^*) \in (0, 1)$ , obviously the system capacity is fully utilized and  $\mathbf{E}[A^*] = C$ . Thus,

$$C \approx \frac{K P_{ar} \mathbf{E}[F_{ar}]}{P_{ar} \mathbf{E}[F_{ar}] \sqrt{f(q^*)} + K} \quad (25)$$

and it follows that  $\sqrt{f(q^*)} \approx \frac{K(P_{ar} \mathbf{E}[F_{ar}] - C)}{P_{ar} \mathbf{E}[F_{ar}] C} = K \left( \frac{1}{C} - \frac{1}{P_{ar} \mathbf{E}[F_{ar}]} \right)$ . Inverting we find

$$q^* \approx f^{-1} \left( K^2 \left( \frac{1}{C} - \frac{1}{P_{ar} \mathbf{E}[F_{ar}]} \right)^2 \right). \quad (26)$$

While (26) is very simple, numerical examples in [11] suggest that it is a reasonable approximation.

<sup>4</sup>The average throughput is expressed here in units of packets per round-trip delay.

## 5. CONCLUSIONS

We have developed a stochastic model for general TCP flows, which takes into account the interactions between the network, transport, and session layers. The resulting model is scalable and becomes more accurate as the number of sessions grows large. An approximation on the queue distribution can also be developed with a central limit theorem-type complement similar to [10]. Limited simulation results indicate a similar limiting behavior for connections with heterogeneous round-trip delays.

While the limiting result applies only to TCP flows with identical round-trip, it will be of use in a number of situations, *e.g.*, the buffer dimensioning problem in an intercontinental Internet link, where the intercontinental link is typically a bottleneck, its large propagation delay dominates the round-trip delays of the connections, and the number of sessions is extremely large.

## REFERENCES

1. S. Floyd, TCP and explicit congestion notification, *Computer Communication Review*, vol. 24, no. 5, pp. 10-23, Oct. 1994.
2. S. Floyd and V. Jacobson, Random early detection gateways for congestion avoidance, *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397-413, Aug. 1995.
3. C.V. Hollot, Y. Liu, V. Misra, and D. Towsley, Unresponsive flows and AQM performance, in *Proceedings of IEEE INFOCOM 2003*.
4. V. Jacobson, Congestion avoidance and control, in *Proceedings of SIGCOMM'88 Symposium*, Aug. 1988, pp. 314-332.
5. A. Kherani and A. Kumar, Stochastic models for throughput analysis of randomly arriving elastic flows in the Internet, in *Proceedings of IEEE INFOCOM 2002*.
6. M. Mathis, J. Semke, J. Mahdavi, and T. Ott, The macroscopic behavior of the TCP congestion avoidance algorithm, *Computer Communication Review*, vol. 27(3), 1997.
7. M. Mellia, I. Stoica, and H. Zhang, TCP model for short lived flows, *IEEE Communications Letters*, vol. 6(2), pp. 85-7, 2002.
8. J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, Modeling TCP Reno performance: a simple model and its empirical validation, *IEEE/ACM Transactions on Networking*, vol. 8(2), pp. 133-45, April 2000.
9. V. Paxson and S. Floyd, Wide area traffic: The failure of Poisson modeling, *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226-244, 1995.
10. P. Tinnakornsrisuphap and A. M. Makowski, Limit behavior of ECN/RED gateways under a large number of TCP flows, in *Proceedings of IEEE INFOCOM 2003*.
11. P. Tinnakornsrisuphap, R. J. La, and A. M. Makowski, Characterization of general TCP traffic under a large number of flows regime, *Tech. Rep.*, The Institute for Systems Research, University of Maryland, College Park, 2002.