Small-Buffer Networks

Mark Shifrin, Isaac Keslassy *

Department of Electrical Engineering, Technion, Haifa 32000, Israel

Abstract

Today, because of TCP dynamics, Internet backbone routers hold large packet buffers, which significantly increase their power consumption and design time. Recent models of large-buffer networks have suggested that these large buffers could be replaced with much smaller ones. Unfortunately, it turns out that these large-buffer network models are not valid anymore in small-buffer networks, and therefore cannot predict how these small-buffer networks will behave.

In this paper, we introduce a new model that provides a complete statistical description of small-buffer Internet networks. We present novel models of the distributions of several network components, such as the line occupancies of each flow, the instantaneous arrival rates to the bottleneck queues, and the bottleneck queue sizes. Later, we combine all these models in a single fixed-point algorithm that forms the key to a global statistical small-buffer network model. In particular, given some QoS requirements, we show how this new model can be used to precisely size small buffers in backbone router designs.

Key words: Small-Buffer Network, Network Model, Backbone Routers.

1 Introduction

Current backbone routers use extremely large buffers. These buffers take about half of their board space and a third of their power consumption [1]. They rely on massive amounts of SRAM and DRAM with fast access times, require complex scheduling algorithms to manage these SRAM and DRAM modules, and can take a significant amount of time to design [2–4].

* Corresponding author (phone: 972-48295738, fax: 972-48295757).
This paper was presented in part at *Networking '08*, Singapore, May 2008.
Email addresses: shifrin@comnet.technion.ac.il (Mark Shifrin), isaac@ee.technion.ac.il (Isaac Keslassy).

These large buffer sizes typically result from a widely-followed rule of thumb, stating that router buffer sizes should be equal to the product of the typical (or worst-case) round-trip-time by the router capacity [5]. This rule of thumb is derived when considering synchronized TCP flows. For instance, given a standard linecard with 40 Gbps and a 250-ms round-trip time, the rule of thumb dictates a large linecard buffer of 10 Gb, which needs several DRAM modules and cannot be practically implemented in SRAM alone.

Recent studies suggest that this rule of thumb overprovisions buffers by several orders of magnitude [6–12]. In fact, given a large number of TCP flows, the synchronization between any two flows is typically weak, and therefore the sum of all the flows is less bursty than for synchronized flows, hence incurring smaller buffer needs. More precisely, these papers argue that the many TCP flows can be modeled as independent, and therefore, by the law of large numbers, the total number of TCP packets in the network converges to a Gaussian distribution. As a consequence, in this model, also known as the *Stanford model*, it is hypothesized that the needed buffer size is smaller by a factor of about $\sqrt{n}$ than the rule of thumb for synchronized flows, where n is the number of TCP flows going through the buffer. For instance, given a million flows, the example above yields a buffer size of about 10 Mb in the Stanford model, which can be implemented in SRAM instead of DRAM. If true, such a result would obviously incur significant architectural changes in backbone routers: for the same power budget, it would be possible to pack more lines, thus increasing the router capacity; the memory architecture would be much simplified; and the input and output queues might be packed together with the switch fabric in a single chip, hence increasing its modularity as well.

Unfortunately, because it analyzes networks with large buffers, the Stanford model assumes that most of the traffic variability is *in the buffers*, and not *on the lines*. However, this assumption does not hold anymore in small-buffer networks, where the variability in the line occupancy cannot be neglected. In fact, it would seem only natural that as buffers get smaller and smaller, the variability in the buffer occupancy progressively reduces as well, until becoming negligible. Thus, in small-buffer networks, most of the network variability intuitively shifts to the line occupancy — and models of large-buffer networks with fixed line occupancies do not hold anymore. A new model is needed for small-buffer networks: this is the objective of this paper.

In this paper, we introduce a new method that provides a complete statistical description of large Internet small-buffer networks with TCP traffic. To do so, we consider each bottleneck queue, and successively build models for the distributions of several network components around this queue. We then connect all these models together in a closed loop, and derive the final network model as a result of a fixed-point equation.

In particular, contributions of this paper include: (a) to our knowledge, the first-ever model for the traffic distribution on the links entering the bottleneck queues; (b) a new model for the instantaneous arrival rate to the bottleneck queues, including a decomposition along the input lines and a Gaussian-based model for the total rate; (c) a model for the occupancy distribution and packet loss rate of bottleneck queues that does not make assumptions on the incoming traffic load and does not assume that it is Poisson; and, most importantly, (d) a closed-loop model of small-buffer networks that enables us to determine their loss rate, as well as the distributions of the major network components. We also conclude this paper with a detailed discussion of assumptions and consequences of these results. In particular, these models can be used by router designers to determine the necessary router buffer size given any loss probability target in any large network topology.

The rest of the work is organized as follows. The next section provides a summary of the related work on this topic. Section 3 presents the notations and the closed-loop model used in this paper. Then, Section 4 contains the models of the different network components, and these models are applied to link occupancy distributions in Section 5. Next, Section 6 presents simulation results that evaluate these models. Finally, Section 7 and Section 8 discuss the assumptions used and the generality of the presented results.

2 Related Work

Our work is related to several recent publications. First, the Stanford model is developed in [6–10]. A core idea in this model is that the sum of all congestion windows is distributed as a Gaussian. The Gaussian distribution is shown by demonstrating the loss of synchronization between the congestion window sizes of the different flows, as their number grows. This leads to the statistical independence between the flows. Therefore, given a large number of homogeneous flows, the buffer size can be divided by the square root of this number of flows, and still remain nearly fully utilized with an acceptable packet loss rate. However, as noted above, none of these papers provide any complete network models for small-buffer networks.

Reference [9] relies on a "paced" pattern of TCP transmissions to consider buffers that are even smaller than those in the Stanford model. The pacing phenomenon is said to correctly model networks with slow access lines or altered TCP. On the contrary, we consider a more general network, without assumption on the access lines, and without modification in the TCP dynamics. We also attempt to provide a more complete characterization of the network properties, including the occupancy distributions of the lines and the queues.

In contrast, [11–14] do provide more complete network models, but they all assume Poisson arrivals to the bottleneck queues. We will later see that these assumptions do not match simulations in small-buffer networks. Moreover, [14] also assumes a level of synchronization between the flows that we did not notice in our simulations of small-buffer networks.

An alternative model, developed in [15], allows to find the buffer size as a function of two alternative targets: the desired load and the desired loss rate. This model shows that for a limited number of flows, the number of lost packets in a congestion event is a near-linear function of the number of flows. However, the linear dependency does not scale beyond a few hundred flows, and therefore does not apply to larger networks.

Recent studies [16–18] also revisit the Stanford model by focusing on different parameters, such as the input/output capacity ratio, the percentage of persistent flows, and the percentage of flows in congestion avoidance mode. Their results complement the Stanford model, by fine-tuning it to these diverse topology parameters, and could complement our model as well.

Finally, other models also consider non-droptail queueing policies, such as RED [19], while we only study the droptail policy, which seems to be the most commonly applied policy in Internet routers.

3 Notations and Closed-Loop Model

3.1 Notations

Our objective is to model a large network with small buffers. We will first formally reduce the problem to a simpler dumbbell topology problem, and then introduce the different notations used.

Assumption 1 (Dumbbell Topology) *The large modeled network can be decomposed into subnetworks with a single bottleneck buffer in each subnetwork, each subnetwork being modeled using a dumbbell topology around its bottleneck buffer.*

This is a classical assumption (see for instance [6, 9, 13, 20, 21]), which we further discuss in Section 7. The intuition behind it is that the behavior of the TCP flows mainly derives from the congestion of the buffers on their path, and that in practice, a single buffer on their path typically causes most of the congestion.

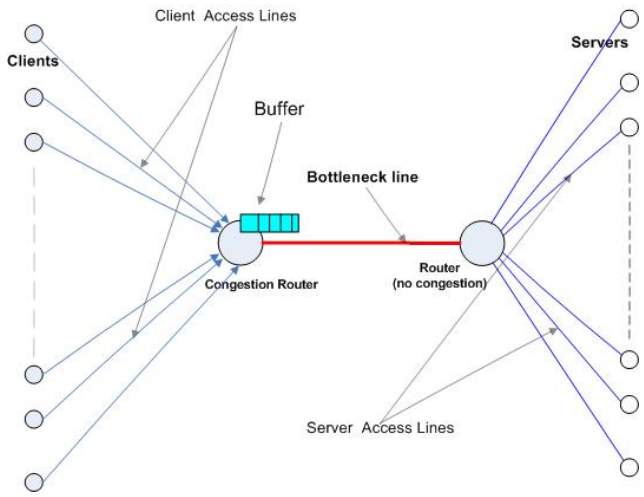

Fig. 1. *Dumbbell topology.*

As shown in Fig. 1, the dumbbell topology includes N persistent TCP flows sharing the same bottleneck buffer of capacity C and buffer size B. For each flow i, with $1 \leq i \leq N$, we will denote the congestion window size as W_i, the forward access link occupancy as L_i, the access link propagation time as T_i, and the total round-trip propagation time, not including the queueing time, as RTT_i. We will also denote the bottleneck queue size as Q, and its loss rate as p.

We assume here that most of the flows are persistent TCP flows, as formerly described in the Stanford model [6]. Other network scenarios are further discussed in Section 8.

The latencies of the forward and backward access links are assumed to follow some given bounded positive distribution, and their capacities are assumed to be so high that the bottleneck link is the only one experiencing congestion. Further, the TCP window sizes are assumed to be integer and have positive lower and upper bounds. The queueing policy is assumed to be drop-tail.

Note that the simulations in Section 6 relax some of these assumptions, e.g. by allowing for short TCP flows. Also, Section 7 further discusses the influence of these assumptions on the general results.

3.2 Closed-Loop Model

Our goal is to provide a closed-loop model for the dumbbell network. First, we will establish a general set of inter-related models. Then, by solving a fixed-point problem involving all these inter-related models, we will converge towards a final network model.

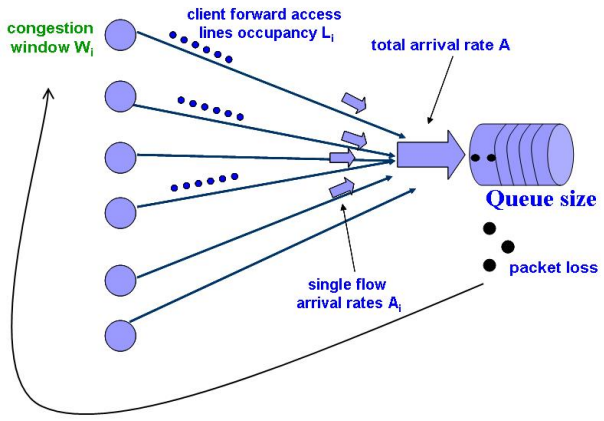

Fig. 2. *Closed-loop schematic model.*

As illustrated in Fig. 2, the inter-related models will successively provide equations for the distributions of: (1) the access link occupancies $\{L_i\}$, (2) the instantaneous arrival rates $\{\Delta A_i\}$, (3) the total instantaneous arrival rate ΔA, (4) the bottleneck queue size Q, (5) the value of the loss rate p, and (6) the congestion window sizes $\{W_i\}$. The first five models are presented in this paper. The last model for the congestion window distribution can be taken from the many literature references (see for instance [22, 23]).

Once the inter-related models are obtained, we can arrange them together in a loop using the following schematic chain:

$$p \Rightarrow \{W_i\} \Rightarrow \{L_i\} \Rightarrow \{\Delta A_i\} \Rightarrow \Delta A \Rightarrow Q \Rightarrow p, \tag{1}$$

which can be rewritten as:

$$p = f(p). \tag{2}$$

We can then simply find p by solving this fixed-point equation, and the solution provides us as well with the distributions of all the network components mentioned above. To solve for p, it is possible to use the gradient descent algorithm for a few iterations, until $|p - f(p)| < \epsilon$ for a desired ϵ.

In other words, we now have the ability to provide a model for all the main characteristics of this small-buffer network when given only the link latencies, bottleneck link capacity, and buffer size.

4 Closed Loop for the Packet Loss Rate Derivation

In this section we will successively go through the inter-related models presented above. We start with a model of the forward access link occupancies L_i.

4.1 Model of the Access Link Occupancies

The distribution of the TCP congestion windows W_i is dictated by the packet loss rate p, and for a fixed p this distribution is fixed. We will now assume that we are given the distributions of the W_i, and want to provide a model for the distribution of L_i. To do so, we will first make two simplifying assumptions.

Assumption 2 (Independence) *The $\{W_i\}_{i=1,\ldots,N}$ are independent and identically distributed.*

In the remainder, we will denote their common positive distribution function as f_W. Intuitively, this simplifying assumption relies on the fact that as the number of flows increases, their mutual synchronization decreases, and therefore the congestion windows can increasingly be modeled as independent. Further, we will consider a network growth in which the common loss rate p stays constant, and therefore the distribution f_W stays constant as well. In other words, to have an apple-to-apple comparison between networks of different sizes, we will assume that the network maintains a similar QoS level while increasing the number of flows and the link capacities. This assumption, as well as the next one, are further discussed in Section 7 and 8.

TCP flows typically send packets and ACKs (acknowledgements) in a highly bursty manner. Moreover, we can use their congestion window size to approximate the size of this burst. The following assumption models this high burstiness.

Assumption 3 (Burstiness) *Flow i has a total of W_i packets (or ACKs on the reverse path), all present as a single burst on a given link.*

Note that this simplifying assumption directly contradicts the common fluid models of TCP, which assume that the window is spread out, and assumes instead that the window is concentrated at a single point. This assumption also uses the fact that we consider small-buffer networks: since the probability of having packets in the buffer is small enough, it can be neglected in this model. We can now derive the distribution of the access link occupancies L_i:

Theorem 1 (Access Link Distribution) *The number of packets on for-*

ward access link i is distributed as:

$$Pr(L_i = k) = \begin{cases} 1 - \frac{T_i}{RTT_i} & \text{if } k = 0, \\ \frac{T_i}{RTT_i} \cdot f_W(k) & \text{otherwise.} \end{cases} \quad (3)$$

Proof. Using Assumption 3, the probability that the burst is not present on the access line is $1 - T_i/RTT_i$. The burst presence probability is independent of its size, since the propagation times are the same no matter what the burst size is. Therefore, the probability that $k > 0$ packets are present on the access line is the product of the probability that the burst size is k ($f_W(k)$) by the probability that the burst is present on the access line (T_i/RTT_i). ■

The simulation results regarding this model are presented in Section 6.

4.2 Arrival Rates of Single Flows and Total Arrival Rate

Denote the number of packets of flow i arrived to the bottleneck queue in Δt seconds as ΔA_i. Intuitively, $\frac{\Delta A_i}{\Delta t}$ represents the (instantaneous) arrival rate on line i to the bottleneck queue. We are interested in studying the distribution of ΔA_i for some small $\Delta t < T_i$, and obtain the following model:

Theorem 2 (Flow Arrival Rate Distribution) *The number of packets ΔA_i of flow i arrived during time Δt is distributed as:*

$$Pr(\Delta A_i = k) = \begin{cases} 1 - \frac{\Delta t}{T_i} + Pr(L_i = 0) \cdot \frac{\Delta t}{T_i} & \text{if } k = 0, \\ Pr(L_i = k) \cdot \frac{\Delta t}{T_i} & \text{otherwise.} \end{cases} \quad (4)$$

Proof. By Assumption 3, packets of flow i move on each line in a single burst of size W_i and at a constant speed. Therefore, the probability that on some link i, $k > 0$ packets arrive within Δt, is $Pr(L_i = k) \cdot \frac{\Delta t}{T_i}$. This gives us the probability for the packet arrival of any size larger then zero. The complementary probability, therefore, stands for the no-arrival event.

Incidentally, note that Equation (4) can be rewritten as in Equation (3):

$$Pr(\Delta A_i = k) = \begin{cases} 1 - \frac{\Delta t}{RTT_i} & \text{if } k = 0, \\ \frac{\Delta t}{RTT_i} \cdot f_W(k) & \text{otherwise.} \end{cases} \quad (5)$$

This is Equation (3), replacing the access link propagation time T_i with the propagation time Δt. ■

We now want to find the total instantaneous arrival rate $\Delta A = \sum_i (\Delta A_i)$, and show that it will converge to a Gaussian distribution given a large number of flows. Since the $\{W_i\}$ and the $\{L_i\}$ are statistically independent by Assumption 3, the $\{\Delta A_i\}$ are statistically independent as well. It now seems natural to use the classic Central Limit Theorem, which applies to the sum of independent and identically-distributed random variables. However, the $\{\Delta A_i\}$ are not identically distributed, because the round-trip times of all flows are not equal. Therefore, we will use Lindeberg's Central Limit Theorem [24] instead. We will show that the share of each flow is bounded, and therefore that the conclusions of the classic Central Limit Theorem still apply, i.e. the total arrival rate ΔA will be Gaussian for a large number of flows.

In order to characterize how the network behaves as the number of flows increases, we need to define how we scale the other parameters. We will keep the same round-trip time distribution, and assume that it is possible to scale the bottleneck link capacity C (and the buffer size B) so that the loss rate is kept constant. This will rely on the fact that smoothly increasing the capacity C will smoothly decrease the packet loss rate, as formulated in the following assumption.

Assumption 4 (Loss Rate Continuity) *The packet loss rate p is a continuous function of the capacity C.*

We will now rely on two lemmas to prove the normality of the limiting arrival rate distribution. In the first lemma, we demonstrate that it is indeed possible to scale the network by keeping a constant loss rate, as long as we can scale C.

Lemma 1 *The packet loss rate p is adjustable by changing the bottleneck link capacity C, i.e. when scaling the number of flows in the network, there exists $C' \geq C$, such that providing a link capacity of C' will maintain the previous packet loss rate p.*

Proof. For $C = 0$, the packet loss rate is 100%, because once the buffer is full, all packets are dropped. On the contrary, for $C \to \infty$, p goes to zero, because the service rate is always higher than the arrival rate (the maximum congestion window size is limited in TCP), and no packets are lost. Since p is a continuous function of C (Assumption 4), the result follows from the Intermediate Value Theorem. ■

In the next lemma, we prove that there is a common lower bound to the standard deviations of the arrival rates, a technical result that will be used for the Lindeberg condition in the Central Limit Theorem.

Lemma 2 *The standard deviations of the flow arrival rates have a common positive lower bound.*

Proof. We use the law of total variance

$$Var(\Delta A_i) = E[Var(\Delta A_i|W_i)] + Var(E[\Delta A_i|W_i]), \tag{6}$$

where the last term is always positive and we omit it to get the following inequality:

$$Var(\Delta A_i) \geq E[Var(\Delta A_i|W_i)]. \tag{7}$$

As expressed in the proof of Equation (5), when we fix $W_i = k$ (with $k > 0$), we get $\Delta A_i = k$ with probability $\frac{\Delta t}{RTT_i}$, and $\Delta A_i = 0$ otherwise. Therefore, using the variance of a scaled Bernoulli random variable, we get

$$Var(\Delta A_i|W_i = k) = \frac{\Delta t}{RTT_i}\left(1 - \frac{\Delta t}{RTT_i}\right)k^2. \tag{8}$$

Taking the expectation over W_i,

$$E[Var(\Delta A_i|W_i)] = \frac{\Delta t}{RTT_i}\left(1 - \frac{\Delta t}{RTT_i}\right)E[{W_i}^2]. \tag{9}$$

Finally, using Equation (7),

$$Var(\Delta A_i) \geq \frac{\Delta t}{RTT_i}\left(1 - \frac{\Delta t}{RTT_i}\right)E[{W_i}^2]. \tag{10}$$

In the above expression, $E[{W_i}^2] \geq 1$, because by Assumption 2, the distribution of W_i follows the positive integral distribution function f_W, which is predetermined for a given loss rate p. Further, we assumed that each RTT_i is given within a predetermined positive range, say $[RTT_{min}, RTT_{max}]$, therefore the above expression is lower-bounded by $\frac{\Delta t}{RTT_{max}}\left(1 - \frac{\Delta t}{RTT_{min}}\right)$, which is positive and independent of the flow number i. ∎

We are now ready to prove the normality of the limiting arrival rate distribution.

Theorem 3 (Total Arrival Rate Distribution) *When the number of flows $N \to \infty$, while keeping the packet loss rate constant, the normalized total arrival rate $\frac{\Delta A - \sum_i E(\Delta A_i)}{\sqrt{\sum_i Var(\Delta A_i)}}$ converges in distribution to the normalized Gaussian distribution $\mathcal{N}(0,1)$.*

Proof. The proof relies on Lindeberg's theorem, which we remind below. Using the statistical independence of the rates,

$$E[\Delta A] = \sum_{i=1}^{N} E[\Delta A_i] \tag{11}$$

and

$$Var(\Delta A) = \sum_{i=1}^{N} Var(\Delta A_i) \tag{12}$$

The distributions of the ΔA_i are independent but not identical. In order to prove the limiting Gaussian nature of the total arrival rate, we will use the Lindeberg Central Limit Theorem, showing that the Lindeberg condition is true. We quote, next, the Lindeberg Central Limit statement:
Lindeberg Theorem: Let $\{X_i\} \in \mathbb{R}^{\mathbb{N}}$ be independent random variables. Assume the expected values $E[X_i] = \mu_i$ and variances $Var(X_i) = \sigma_i^2$ exist and are finite. Also let $s_N^2 = \sum_{i=1}^{N} \sigma_i^2$. If the sequence of independent random variables X_i satisfies the Lindeberg condition (formulation according to the terminology of Zabell [24]):

$$\lim_{N\to\infty} \sum_{i=1}^{N} E\left[\left(\frac{X_i}{s_N}\right)^2 : \frac{|X_i|}{s_N} \geq \epsilon\right] = 0 \tag{13}$$

Then, the following term:

$$Z_N = \frac{\sum_{i=1}^{N}(X_i - \mu_i)}{s_N} \tag{14}$$

converges in distribution to a standard normal random variable as $N \to \infty$. ■

We substitute X_i by ΔA_i, other notations staying the same. Then, for every $\epsilon > 0$, we want to prove the Lindeberg condition as follows:

$$\lim_{N\to\infty} \sum_{i=1}^{N} E\left[\frac{(\Delta A_i - \mu_i)^2}{s_N^2} : |\Delta A_i - \mu_i| > \epsilon \cdot s_N\right] = 0 \tag{15}$$

The meaning of this condition is that there is no capture phenomenon, in which a unique flow or a small number of flows seize a major part of the link capacity. We will use the lemmas above to prove that as we scale the number of flows, s_N^2 grows as well, and no flow will be able to satisfy the condition $|\Delta A_i - \mu_i| > \epsilon \cdot s_N$.

As shown in Theorems 1 and 2, the distribution of ΔA_i depends solely on the window distribution, which is derived from the constant packet loss [23]. Therefore this distribution of ΔA_i does not depend on N. According to Lemma 1, the number of flows N can be scaled while preserving the same packet loss rate. Further, the arrival rates of the new flows will obey the same lower bounds on the variance as the arrival rates of the existing flows, as found in Lemma 2, because their propagation times have the same bounds. Consequently, s_N^2 will be lower-bounded by a linear function of N, which goes to infinity with N. Let W_{max} be the maximum window size. Then for every $\epsilon > 0$, we can find some $N_0 \geq 1$, such that for any number of flows $N \geq N_0$, s_N will satisfy the condition $s_N \geq W_{max}/\epsilon$. Thus, for $N \geq N_0$,

$$Pr(|\Delta A_i - \mu_i| > \epsilon \cdot s_N) \leq Pr\left(|\Delta A_i - \mu_i| > \epsilon \cdot \frac{W_{max}}{\epsilon}\right) = 0, \tag{16}$$

where the equality in the second term of this inequality relies on Equation (5), that is, the distribution of the arrival rate ΔA_i of each flow i, is bounded by the maximum value of the congestion window, namely W_{max}. Therefore, as we scale N, the Lindeberg condition is satisfied, and we can apply the Central Limit Theorem to the total arrival rate. ■

Note that the distribution of ΔA_i is derived from the distribution of W_i, which defines a distribution of the TCP congestion window, and therefore holds for the closed-loop model. Thus, the Gaussian distribution in the theorem proved above holds in the closed-loop model. The simulation results comparing this Gaussian model with a typical Poisson model are presented in Section 6.

4.3 Queue Size Distribution and Packet Loss Rate

Our next objective in the fixed-point model is to find the distribution of the queue size Q and the packet loss rate p given the above model for ΔA. We will decompose time into frames of size Δt, and assume that packets arrive as bursts of size ΔA every Δt seconds. Thus, we clearly get the following queueing model:

Theorem 4 (Queue Size and Loss Rate) *The queue size distribution and the packet loss rate are obtained by using a* $G_{\Delta t}^{[\Delta A]}/D/1/B$ *queueing model, in*

which every Δt seconds, packets arrive in batches of size ΔA and immediately obtain service for up to $C\Delta t$ packets.

We implemented this queueing model using both the algorithm developed in [25] and a simple Markov chain. The first method uses an iterative algorithm, which makes the steady state probabilities of the queue converge. The second method involves running a long simulation of the total arriving rate as the arrival process to the queue. Both methods yielded similar results.

We are now done with our set of network models, which can all be combined to form a fixed-point solution. In the next sections, we first apply these models to demonstrate the asymptotic normality of several network components, and then analyze their correctness using simulation results.

5 Application: Gaussian Models of Line Occupancies

After building the small-buffer network model, we would like to demonstrate its use, by proving the asymptotic normality of the distributions of several line occupancies. In fact, in the Stanford model of large-buffer networks [6–10], the *total window size* is modeled as asymptotically Gaussian, but the *line occupancies* are modeled as constant. In small-buffer networks, we will similarly demonstrate that the total window size can be modeled as asymptotically Gaussian. Then, we will use our closed-loop model to prove that the total access line occupancy is asymptotically Gaussian as well, and not constant, contrarily to the Stanford model. Intuitively, this is because with small queues, the variations in the queue occupancy cannot justify alone the variations in the total network occupancy.

5.1 Gaussian Model of the Total Congestion Window W

Our objective is to prove that the total window size $W = \sum_{i=1}^{N} W_i$ is asymptotically Gaussian (as already assumed in [6]). To do so, we will rely on the statistical independence of the congestion window sizes $\{W_i\}$.

Theorem 5 (Gaussian Model of W) *When the number of flows $N \to \infty$, while keeping the packet loss rate constant, the normalized total congestion window size $\frac{W - \sum_i E(W_i)}{\sqrt{\sum_i Var(W_i)}}$ converges in distribution to the normalized Gaussian distribution $\mathcal{N}(0,1)$.*

Proof. Using the statistical independence of the window sizes (Assumption 2),

$$E[W] = \sum_{i=1}^{N} E[W_i], \qquad \text{and} \qquad Var(W) = \sum_{i=1}^{N} Var(W_i). \tag{17}$$

The distributions of the $\{W_i\}$ are independent and identically distributed according to f_W (Assumption 2 and Lemma 1), therefore we can apply the Central Limit Theorem and obtain the resulting convergence of the normalized sum to a normalized Gaussian distribution. ■

5.2 Gaussian Model of the Total Access Line Occupancy L

We will now demonstrate that the total occupancy on the forward access links $L = \sum_{i=1}^{N} L_i$ is asymptotically Gaussian. This is contrary to the Stanford large-buffer model, in which they are assumed to be constant.

In the proof, we will rely on the independence of the line occupancies $\{L_i\}$. However, since they are independent but not identically distributed, the proof will use Lindeberg's Central Limit Theorem [24].

Theorem 6 (Gaussian Model of L) *When the number of flows $N \to \infty$, while keeping the packet loss rate constant, the normalized total forward access link occupancy $\frac{L-\sum_i E(L_i)}{\sqrt{\sum_i Var(L_i)}}$ converges in distribution to the normalized Gaussian distribution $\mathcal{N}(0,1)$.*

Proof. Using the statistical independence of the link occupancies,

$$E[L] = \sum_{i=1}^{N} E[L_i], \qquad \text{and} \qquad Var(L) = \sum_{i=1}^{N} Var(L_i). \tag{18}$$

The random variables $\{L_i\}_{1 \le i \le N}$ are independent but not identical. In order to prove the limiting Gaussian nature of the total link occupancy, we will use the Lindeberg Central Limit Theorem, as in the proof of Theorem 3. Then, the proof can follow the proof of Theorem 3, with the individual link occupancy distributions now satisfying Theorem 1, thus leading directly to the convergence result. ■

In the theorem above, the total access link occupancy is modeled as asymptotically Gaussian. In the simulations below, we will evaluate whether simulations

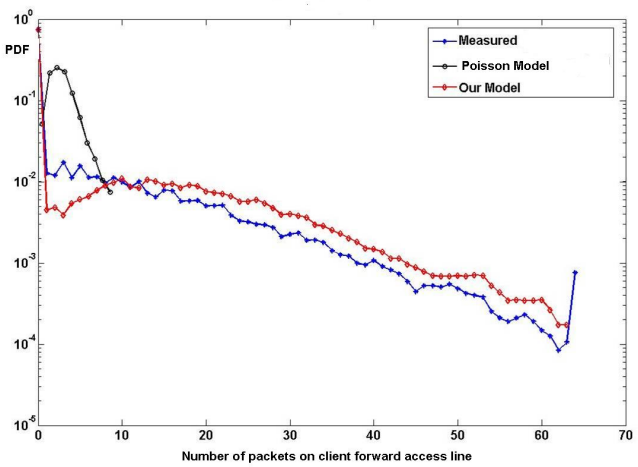

Fig. 3. *Distribution of the access link occupancy L_i (logarithmic scale).*

actually confirm such a model, and analyze other link distributions as well. In fact, the forward access links are the only links in the dumbbell network before the bottleneck, i.e. in which no packets have been lost yet. Therefore, it is expected that the independence-based Gaussian model will fare best on these links, since packet drops would not influence these as much and therefore would cause less inter-flow correlation. This will indeed be confirmed in the simulations section.

Incidentally, note that the parameters of the Gaussian distribution of L are related to those of the Gaussian distribution of W. For instance, using Theorem 1, we get

$$E[L] = \sum_{i=1}^{N} E[L_i] = \sum_{i=1}^{N} \frac{T_i}{RTT_i} \cdot \frac{E[W]}{N} = E[W] E_i \left[\frac{T_i}{RTT_i} \right] \tag{19}$$

6 Simulation Results

We will now present the simulation results for the different parts of the closed-loop model. The simulations were done in NS2 [26]. In all the simulations below, we ran 2,000 simultaneous persistent TCP NewReno flows, unless noted otherwise. The packet size was set to 1,000 bytes. The maximum allowed window size was set to $W_{max} = 64$ packets. The propagation time of the bottleneck link was fixed to 20 milliseconds, and the propagation times for all the other links were chosen randomly according to a uniform distribution.

6.1 Access Link Occupancy Model vs. Fluid Model (Theorem 1)

We want to evaluate our bursty model for $\{L_i\}$, as presented in Theorem 1. We compare it against a *fluid model*, in which the packets belonging to a flow are distributed uniformly on all the links (the queueing time is neglected).

According to this fluid model, the number of packets present on access link i at time t is thus equal to $L_i(t) = \frac{T_i}{RTT_i} \cdot W_i(t)$, because flow i has a total of $W_i(t)$ packets, and the share in the propagation time of access link i is $\frac{T_i}{RTT_i}$. Therefore, the maximum number of packets on the access link is bounded by $\frac{T_i}{RTT_i} \cdot W_{max}$.

Fig. 3 represents the pdf (probability distribution function) of the access link occupancy L_i of some random flow i using a logarithmic scale. It was obtained using a simulation involving 500 simultaneous TCP flows. It can be seen that our bursty model is fairly close to the measured results throughout the whole scale. It behaves especially better than the fluid model, for which the pdf is equal to 0 for any occupancy above $\frac{T_i}{RTT_i} \cdot W_{max} \approx 10$, and thus cannot even be represented on the logarithmic scale. This observation strengthens the intuition that Assumption 3 was reliable. Note that we obtained similar results on many simulations with different parameters.

6.2 Arrival Rate Model vs. Poisson Model (Theorems 2 and 3)

We now want to evaluate our Gaussian-based model for the distribution of the total instantaneous arrival rate ΔA, as presented Theorems 2 and 3. We will compare this model with a typical *Poisson arrival model* [11–14].

Fig. 4 plots the pdf of ΔA, using $\Delta t = 10$ ms, and compares it with simulated data obtained artificially using the two models. Our Gaussian-based model relies on the expectation and variance computed above. Of course, over this amount of time, the Poisson model yields an approximately Gaussian distribution as well, but not with the same parameters: we force its expectation to equal the measured value, and obviously the variance equals the expectation for a Poisson random variable.

As shown in the figure, our model approximately yields the correct variance and is close to the simulated results. On the contrary, the Poisson model yields too small a variance. This can be intuitively explained by the burstiness properties of the TCP flows which are not reflected in the Poisson model.

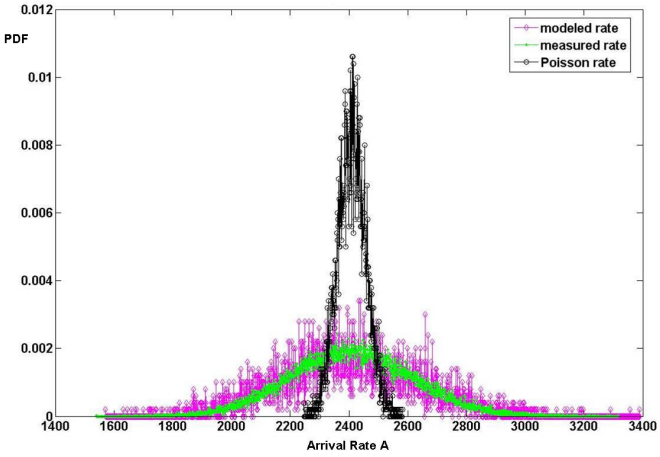

Fig. 4. *Distribution of the total arrival rate ΔA in time Δt.*

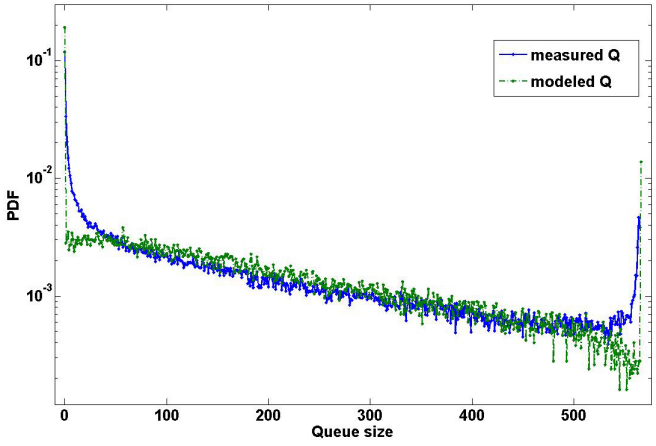

Fig. 5. *Queue size distribution.*

6.3 Queue Size Distribution (Theorem 4)

We now want to compare our model of the queue size distribution, with its measured value from a simulation. Fig. 5 illustrates such a comparison. In the simulation, we use a buffer size of 580 packets. Our Markov-chain-based queue model is obtained after the convergence of the entire closed-loop model, and therefore uses our modeled arrivals as well.

It can be seen that our queue model is fairly close on most of the simulated range of Q, but cannot exactly reproduce the smooth continuous behavior of the queue at the edges ($Q = 0$ and $Q = B$) because of its discrete bursty nature. In fact, in this example, the simulated loss rate is 0.55%, while our model gives 0.76%.

Case	Flow throughput	Queueing delay		Loss rate p	
		Measured	Modeled	Measured	Modeled
Case 1	52 pkts/sec	1.79 msec	1.9 msec	0.79 %	0.90 %
Case 2	29 pkts/sec	18.27 msec	19.87 msec	1.81 %	2.10 %
Case 3	11 pkts/sec	7.84 msec	8.54 msec	1.30 %	1.30 %
Case 4	18 pkts/sec	22.16 msec	26.15 msec	3.30 %	3.50 %

Fig. 6. *Measured versus modeled results.*

6.4 Fixed-Point Model

Let's now evaluate the solutions of our fixed-point model, which combines all the other models. In our simulations, the fixed-point solution of our network model was found using the gradient descent algorithm, in typically less than 50 iterations. The exact number of needed iterations depended of course on the desired precision and on the network parameters.

The table presented in Fig. 6 illustrates the average flow throughput, the average queueing delay (measured and modeled) and the average loss rate (measured and modeled) using the simulation results in four settings with quite different parameters:
- *Case 1:* 500 long-provisioned TCP flows with RTT_i distributed between 80 and 440 msec, $B = 232$ packets and $C = 232.5$ Mbps.
- *Case 2:* 500 long-provisioned TCP flows with RTT_i distributed between 80 and 440 msec, $B = 450$ packets and $C = 116.25$ Mbps.
- *Case 3:* 750 long-provisioned TCP flows and a constant number of 25 short TCP flows, with RTT_i distributed between 70 and 2,040 msec, $B = 244$ packets and $C = 69.75$ Mbps. (The short flows were omitted in the model.)
- *Case 4:* 1500 long-provisioned TCP flows and a constant number of 5 short TCP flows, with RTT_i distributed between 220 and 240 msec, $B = 828$ packets and $C = 209.25$ Mbps. (The short flows were omitted in the model.)

In all these cases, the modeled queueing delay and packet loss rate were close to, but slightly above, the measured results. Incidentally, using the same simulations without short flows, we also verified that the influence of the short flows on the simulation results was negligible. Note also that all these cases use small buffers, with the last case following the Stanford model, and the first three cases having even smaller buffers.

6.5 Gaussian Model of Congestion Window (Theorem 5)

We now want to check whether our Gaussian model of the total TCP congestion window is indeed confirmed in simulations. Fig. 7 illustrates this model as measured from an NS2 simulation (using $2,000$ flows, and with a measured

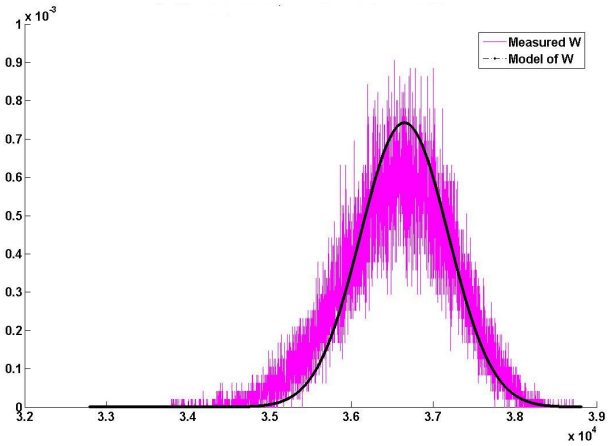

Fig. 7. *Distribution of the total congestion window size W*

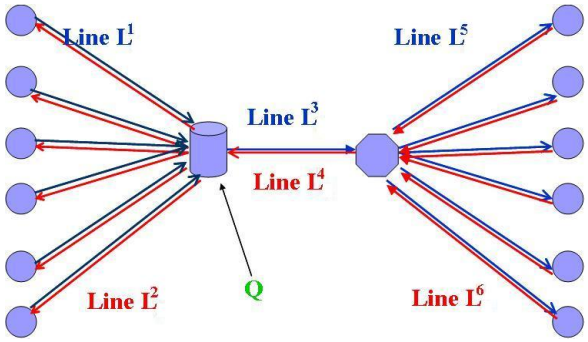

Fig. 8. *Link numbers in the dumbbell topology*

packet loss of 0.96%). The average and variance of the simulation were taken from the closed-loop network model, and not from the measured results.

As the figure shows, both the modeled mean and variance seem close to the measured values, and therefore the closed-loop model was accurate enough in this case. Indeed, the total congestion window seems approximately Gaussian, although with a slightly heavier left tail (probably mostly due to the non-Gaussian distribution of the bottleneck component).

6.6 Gaussian Model of Forward Access Links (Theorem 6)

Our next objective is to check whether the Gaussian model of the total forward access link occupancy is confirmed in simulations. In the simulations below, we will actually want to examine *all the links*, and not only the forward access links. Therefore, for simplicity, we will number the different link types in the dumbbell topology. As illustrated in Fig. 8, we will respectively denote L^1 (L^2) the forward (backward) source access links, L^3 (L^4) the forward (backward) bottleneck links, and L^5 (L^6) the forward (backward) destination links. For

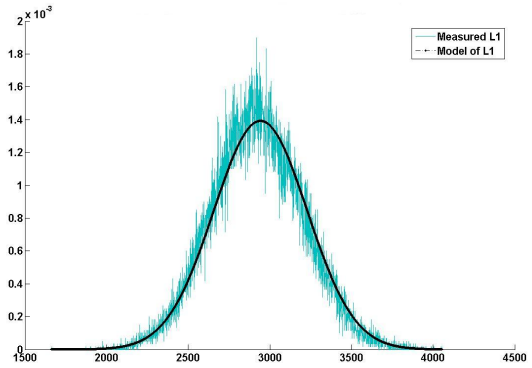

Fig. 9. *Distribution of the total forward access link occupancy* L^1

simplicity, we will use the same symbols $\{L^j\}_{j=1,\ldots,6}$ for the links and for their occupancies. Also, note that in particular $L^1 \triangleq L$ was analyzed in Section 5, and that the backward links actually carry acknowledgments, not data packets.

In Theorem 6, we found that the total forward access link occupancy L^1 could be asymptotically modeled using a Gaussian distribution. Fig. 9 illustrates this model, by comparing it to the measured results from an NS2 simulation (using 500 flows, with a measured packet loss of 0.62%).

As the figure illustrates, the model is a good fit. The modeled Gaussian parameters were taken from the closed-loop model, and not from the measured results. Still, the Gaussian model seems very close to the measured distribution. (As before, this was observed across a wide range of flow numbers, loss rates, and topology settings.)

6.7 Distribution of Link Occupancies

We found above that the inter-flow correlation on L^1 is small enough that the total link occupancy can be accurately modeled by a Gaussian distribution. We will now see that this is not necessarily the case for the other links.

In particular, we do not expect the link occupancy on the bottleneck link L^3 to have a Gaussian distribution. Link L^3 contains the packets leaving the bottleneck buffer, so its link capacity will limit its total occupancy, thus incurring a high inter-flow correlation.

Further, because of the buffer impact, the flows are not completely independent on L^2, L^4, L^5 and L^6. However, their correlation is still relatively small, and due to their different round-trip times, the correlating influence of the buffer declines, bringing their overall occupancy to be quite close to a Gaussian distribution.

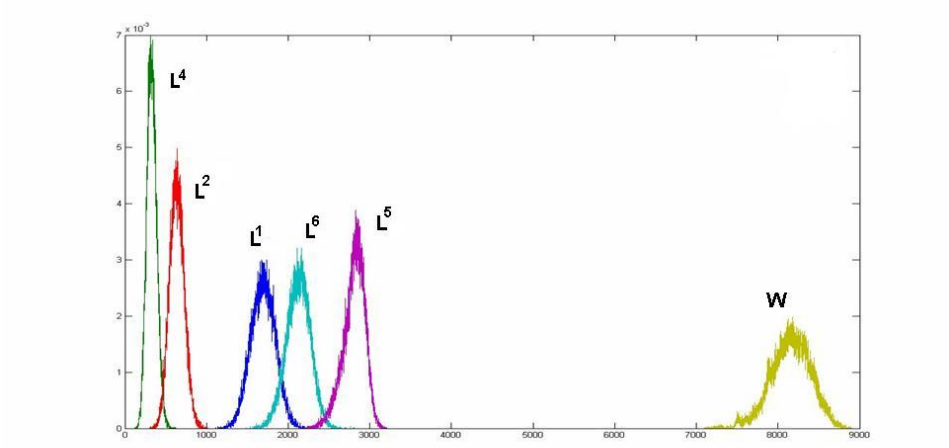

Fig. 10. *Distributions of total link occupancies and window size*

Fig. 10 illustrates the distributions of these links and of the total window. As can be seen, the five links and the total window size all seem to follow a Gaussian distribution. (Note that bottleneck link L^3 is not represented, as it consists of a sharp peak that extends well above the plot.)

However, figures can be deceptive. That's why we also represented the link occupancies using the Q-Q (Quantile-Quantile) plot method [27]. Q-Q plots illustrate the differences between the probability distribution of a measured sample and a Gaussian distribution. A linear Q-Q plot, especially near the center, is a sign that the sample can be modeled as Gaussian.

Fig. 11 presents the results for all link occupancies $\{L^j\}_{j=1,\ldots,6}$. As can be seen, these plots confirm our analysis. Link L^1 is highly Gaussian, as suggested in Theorem 6. Links L^2, L^4, L^5 and L^6 are nearly Gaussian, to varying degrees (arguably less so for L^5, which follows the bottleneck link in the network). Finally, the bottleneck link L^3 is clearly not Gaussian.

7 Discussion of Assumptions

Let's now discuss the correctness and generality of the assumptions.

Dumbbell Topology — We assumed in Assumption 1 that any large network can be subdivided into dumbbell topologies. This assumption relies on the observation that in the Internet, few flows practically have more than one bottleneck, and that flows having more than one bottleneck actually mainly depend on the most congested one [6]. Thus, the assumption of a single point of congestion seems realistic enough. However, we also assumed that the congestion only affects packets, not ACKs. This assumption might be too restrictive

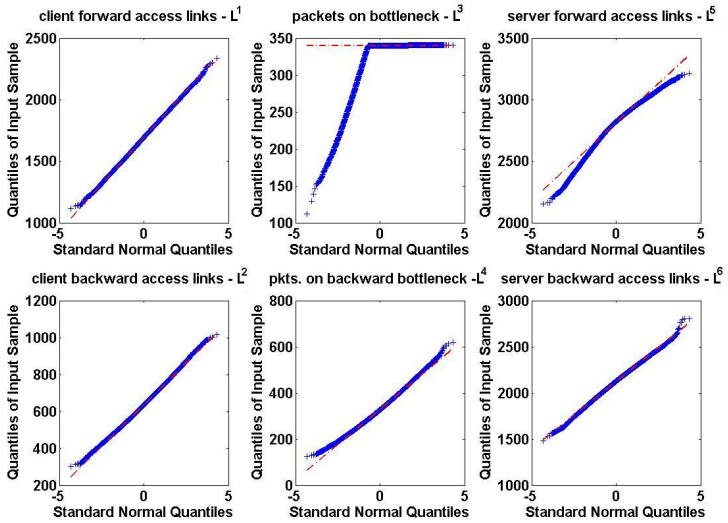

Fig. 11. *Normal Q-Q plots of measured link occupancies.*

– even though we found that our Gaussian-based models still held in various simulations using reverse-path ACK congestion.

Statistical Independence of $\{W_i\}_{i=1,\ldots,N}$ — It is obviously not correct that the congestion windows are completely independent, since they interact through the shared bottleneck queue. However, in order to study how far from reality the independence assumption is, we first checked how correlated the congestion windows are over time. We obtained the following correlation matrix for the congestion windows of five arbitrary flows, using 70,000 consecutive time samples in a simulation with 500 persistent TCP flows.

$$C = \begin{pmatrix} 1 & 0.066 & 0.14 & 0.058 & -0.025 \\ 0.066 & 1 & 0.054 & 0.0005 & -0.081 \\ 0.14 & 0.054 & 1 & 0.063 & 0.051 \\ 0.058 & 0.0005 & 0.063 & 1 & 0.003 \\ -0.025 & -0.081 & 0.051 & 0.003 & 1 \end{pmatrix} \tag{20}$$

It can be seen that the correlation coefficients between different flows are indeed low in front of the maximum absolute value of 1.

Of course, while independent random variables have zero correlation, the reverse is not necessarily true. Nevertheless, this low correlation would tend to indicate that the flows are indeed desynchronized, and therefore that the simplifying assumption of independence is not too far from reality. In fact, as the number of flows increases, we also found that a heuristic measure of the independence of their window sizes was decreasing. We took two arbitrary flows, and measured the symmetric form of the Kullback-Leibler distance between the joint distribution of their window sizes and the product of their respec-

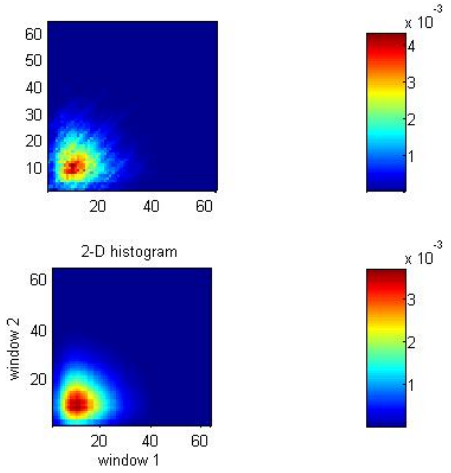

Fig. 12. *Comparison of the distributions of 2 congestion windows with their joint distribution.*

tive distributions. We found a measure of 20.03 for a network with 10 flows, 3.37 with 50 flows, and 2.65 with 200 flows, with a lower measure intuitively reflecting a higher independence between the window size distributions.

Fig. 12 illustrates such a comparison between the joint distribution and the distribution product of two arbitrary windows, as obtained in a simulation. The upper chart displays the joint distribution of the two window sizes, while the lower chart shows the product of the distributions of the two windows. By definition, the upper and lower charts for exactly independent windows would be identical.

Identical Distribution of $\{W_i\}_{i=1,\ldots,N}$ — The distributions of the congestion window sizes mainly depend on the loss rate p in the shared bottleneck buffer [22, 23]. In simulations, when there was no strong synchronization, the loss rate was indeed found to be nearly equal for all flows, and the window size distributions were nearly equal as well.

Burstiness — In Section 6, the comparison of the bursty model with the fluid model already strengthened the bursty assumption. More generally, this assumption needs to be used with care in networks without enough space on the links for a packet burst (extremely small link latencies or link capacities). In other cases, while not completely reflecting reality, this assumption seems close enough [6].

RTT Distribution and Queuing Delay — We assumed that the queueing delay can be neglected in front of the link propagation times. In fact, in the Stanford model, the worst-case queueing delay is $\frac{B}{C} = \frac{RTT}{\sqrt{n}}$, where RTT is the average round-trip propagation time. Consequently, the assumption seems reasonable when n is large, as long as there is no flow with a round-trip

propagation time significantly small in front of RTT.

Number of Packets — We assumed that the number of packets in the network is close to the total window size, as reflected by the definition of the window size. This assumption is especially justified when the loss rate is reasonably small, as seen in our simulations and in the literature [6].

8 Generality of Results

It is obviously impossible to consider every possible topology and every possible set of flows. We made hundreds of simulations for this paper, and still feel that there is much to research. Nevertheless, we can already discuss the scope of the results and their sensitivity to various topology parameters.

Buffer Size — Our paper is about models that are valid in networks with small buffers. In a network with large buffers, the flows might become synchronized, and our independence assumptions and the ensuing models might not be applicable. Likewise, we would not be able to neglect the queueing delay in our models.

Propagation Times — We chose the link propagation times using a uniform distribution with different parameters so as to reflect the diversity of real-life Internet flows. In simulations, our models were fairly insensitive to these propagation times, as long as there were many flows and there were no flows with near-zero round-trip-times.

Number of Flows — In simulations, we found that the desynchronization was already practically correct for several hundred flows, as previously stated in [6]. We thus believe that with the hundreds of thousands of flows present in a congested backbone router, the desynchronization will be even more correct. The same desynchronization effect is expected due to the growth of the Internet resources. The growing router service rates would increase the mean and variance of the accessing flows. However, we would expect that this increase would not affect the independence assumption in a significant way.

Protocols — We assumed that most of the traffic consists of persistent TCP flows, running our model in simulations with only a small fraction of short-term TCP flows. Considering a large fraction of them is out of scope of this paper. We also introduced a fraction of UDP flows, and noted that these flows only had a minimal impact on buffer sizing.

9 Conclusion

In this paper, we provided a complete statistical model for a large network with small buffers. We started with a model for the traffic on a single access line. Then, we modeled the arrival rates to the bottleneck queues. Later, we found a model for the queue size distribution and the loss rate. Finally, using these inter-related models, we showed how to solve a single fixed-point equation to obtain the full network statistical model, and how to use this fixed-point model to provide Gaussian models of link occupancies.

A router designer might directly use this network model for buffer sizing. Indeed, the designer might consider a set of possible benchmark parameters and model the behavior of the resulting network. Then, given target QoS requirements such as the required maximum packet loss rate or maximum expected packet delay, the router designer will be able to design the buffer size that satisfies these constraints.

Acknowledgments

This work was partly supported by the European Research Council Starting Grant n° 210389, the Alon Fellowship, the ATS-WD Career Development Chair, and the Loewengart Research Fund. We would like to thank Raphi Rom, Israel Cidon, Reuven Cohen, and the Technion Comnet lab staff for their insightful comments.

References

[1] N. McKeown, "Sizing router buffers," *EE384Y Course Talk*, Stanford, 2006.

[2] G. Shrimali, I. Keslassy, and N. McKeown, "Designing packet buffers with statistical guarantees," *IEEE Hot Interconnects XII*, Stanford, CA, 2004.

[3] J. Garcia-Vidal, M. March, L. Cerda, J. Corbal, and M. Valero "A DRAM/SRAM memory scheme for fast packet buffers," *IEEE Transactions on Computers*, May 2006.

[4] S. Iyer, R. R. Kompella, and N. McKeown, "Designing packet buffers for router line cards," to appear in *IEEE/ACM Transactions on Networking*, 2008.

[5] C. Villamizar and C. Song, "High performance tcp in ansnet," *ACM Computer Communications Review*, 1994.

[6] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," *ACM SIGCOMM,* Portland, OR, 2004.

[7] G. Appenzeller, N. McKeown, J. Sommers, and P. Barford, "Recent results on sizing router buffers," *Network Systems Design Conference*, San Jose, CA, 2004.

[8] D. Wischik and N. McKeown, "Part I: buffer sizes for core routers," *ACM Computer Communications Review*, 35(3): 75–78, 2005.

[9] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden, "Routers with very small buffers," *IEEE Infocom*, Barcelona, Spain, 2006.

[10] J. Cruise, "Poisson convergence, in large deviations, for the superposition of independent point processes," submitted to *Annals of Operations Research.*

[11] G. Raina, D. Towsley, and D. Wischik, "Part II: control theory for buffer sizing," *ACM Computer Communications Review*, 35(3): 79–82, 2005.

[12] G. Raina and D. Wischik, "Buffer sizes for large multiplexers: TCP queueing theory and instability analysis," *EuroNGI*, 2005.

[13] K. E. Avrachenkov, U. Ayesta, E. Altman, P. Nain, and C. Barakat, "The effect of router buffer size on the TCP performance," *LONIIS Workshop*, Saint Petersburg, Russia, 2002.

[14] K. Avrachenkov, U. Ayesta, A. Piunovskiy, "Optimal choice of the buffer size in the Internet routers," *IEEE CDC-ECC*, 2005.

[15] A. Dhamdhere, H. Jiang, and C. Dovrolis "Buffer sizing for congested Internet links," *IEEE Infocom*, Miami, FL, March 2005.

[16] G. Vu-Brugier, R. Stanojevic, D. Leith, and R. Shorten, "A critique of recently proposed buffer sizing strategies," *ACM Computer Communications Review*, 37(1): 43–48, 2007.

[17] R. S. Prasad, C. Dovrolis, and M. Thottan, "Router buffer sizing revisited: the role of the input/output capacity ratio," *ACM CoNext*, New York, 2007.

[18] R. S. Prasad and C. Dovrolis, "Beyond the model of persistent TCP flows: open-loop vs closed-loop arrivals of non-persistent flows," *41st Annual Simulation Symposium*, April 2008.

[19] T. Bu and D. F. Towsley, "A fixed point approximation of TCP behavior in a network," *ACM SIGMETRICS*, 2001.

[20] J.P. Hespanha, S. Bohacek, K. Obraczka, and J. Lee, "Hybrid modeling of TCP congestion control," *HSCC*, Rome, Italy, 2001.

[21] R. Shorten, F. Wirth, and D. Leith, "A positive systems model of TCP-like congestion control: asympototic results," *Tech. Rep. 2004-1*, Hamilton Institute, 2004.

[22] E. Altman, K. E. Avrachenkov, A. A. Kherani, and B. J. Prabhu, "Performance analysis and stochastic stability of congestion control protocols," *IEEE Infocom*, Miami, FL, 2005.

[23] M. Shifrin, "The Gaussian nature of TCP in large networks," *M.Sc. Research Thesis*, Technion, Israel, August 2007.

[24] S. L. Zabell, "Alan Turing and the central limit theorem," *American Mathematical Monthly*, 102: 483–494, 1995.

[25] P. Tran-Gia and H. Ahmadi, "Analysis of a discrete-time $G^{[X]}/D/1-S$ queueing system with applications in packet-switching systems," *IEEE Infocom*, New Orleans, LA, 1988.

[26] NS2 Network Simulator, available at `www.isi.edu/nsnam/ns/`

[27] J.J. Filliben, "The probability plot correlation coefficient test for normality," *Technometrics*, 17(1), 1975.