2013

Sizing Router Buffer for the Internet with Heterogeneous TCP

Peng Yang
University of Nebraska-Lincoln, pyang@cse.unl.edu

Ertong Zhang
University of Nebraska-Lincoln, ezhang@cse.unl.edu

Lisong Xu
University of Nebraska-Lincoln, xu@cse.unl.edu

Performance Computing and Communications Conference (IPCCC), 2013 IEEE
32nd International
Year: 2013
Pages: 1 - 9, DOI: 10.1109/PCCC.2013.6742797

Sizing Router Buffer for the Internet with Heterogeneous TCP

Peng Yang, Ertong Zhang, Lisong Xu
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0115
Email: {pyang, ezhang, xu}@cse.unl.edu

***Abstract*—The router buffer sizing problem is a vital problem to the performance of the Internet. The traditional rule-of-thumb is that the router buffer size should be equal to the bandwidth-delay product (BDP) of a link. Recent studies show that the router buffer size can be significantly smaller than the BDP without causing negative impact on the TCP performance in the Internet. But a fundamental assumption of all those studies is that all the TCP traffic in the Internet is generated by the traditional RENO protocol, which, however, is no longer true as the current Internet is dominated by multiple different TCP protocols, such as RENO, CUBIC and Compound TCP (CTCP). Thus, it is imperative that we revisit the router buffer sizing problem for the Internet with heterogeneous TCP. In this paper, we propose methods to determine the router buffer size requirements under various constraints for the Internet with heterogeneous TCP and discuss the tradeoff among the constraints. The constraints considered include the link utilization constraint, the packet drop rate constraint, and the queuing delay constraint. Our study shows that the required router buffer size can be significantly smaller than the BDP but also demonstrates that it is dependent on the protocol mix of the heterogeneous TCP flows.**

I. Introduction

The router buffer sizing problem is an important problem in the Internet resource planning. Simply put, the router buffer sizing problem is to set an appropriate router buffer size for a link so that both link utilization and TCP performance are maximized. We focus on TCP traffic, since a significant amount of Internet traffic is controlled by TCP.

Maximizing link utilization: To maximize the utilization of a link, the router buffer size should be large enough to absorb traffic fluctuation and keep the link fully utilized. The appropriate router buffer size depends on specific traffic characteristics (e.g., mean and variance). We want the router buffer size as small as possible, but still can maintain a certain percentage of the link utilization. The minimum router buffer size is important because it helps to avoid the bufferbloat problem where the excessive router buffer size results in large queuing delay and large delay variation.

The traditional rule-of-thumb is to set the minimum router buffer size of a link to at least the bandwidth-delay product (BDP) of the link. However, this rule is based on the traffic characteristics of the RENO protocol [1], [2], and considers only one or a few concurrent TCP flows. Recent studies [3] show that the minimum router buffer size of a link with a large number of concurrent TCP flows can be significantly smaller than the BDP without causing negative impact on the TCP performance. But a fundamental assumption of all those studies is that all the TCP traffic in the Internet is generated by the traditional RENO protocol. As more and more different TCP protocols are proposed and deployed in the Internet, these rules and studies cannot adapt to the current Internet where TCP flows are controlled by multiple different TCP protocols.

Maximizing TCP performance: The performance of a TCP flow can be measured by the flow completion time, which is the the time taken for the TCP flow to complete the transmission. To shorten the flow completion time, the TCP flow should experience a low packet drop rate and short queuing delay, both of which highly depend on the router buffer sizes along the path of the TCP flow. These two constraints usually contradict with each other. On the one hand, a low packet drop rate requires large router buffer sizes. On the other hand, short queuing delay requires small router buffer sizes.

Heterogeneous TCP: A recent Internet measurement result [4] shows that multiple different TCP protocols have been deployed in the Internet, and there are 5 widely deployed TCP protocols: RENO, BIC, CUBIC', CUBIC", and CTCP. RENO [1] is the traditional TCP protocol, and in this paper we use RENO to refer to the traditional Additive-Increase-Multiplicative-Decrease congestion control algorithm used in both Reno [5], NewReno [6], and SACK [7]. BIC [8] is the default TCP protocol of Linux kernel 2.6.19 and before. After that, BIC was replaced by CUBIC [9]. There are two major CUBIC versions: the one implemented in Linux kernel 2.6.25 and before is referred to as CUBIC', and the one implemented in Linux kernel 2.6.26 and after is referred to as CUBIC". CTCP [10] is the default TCP protocol in Windows. Among the 30,000 web servers in the measurement, only 3.31%∼14.47% of the web servers still use RENO, 20.45% use BIC, 12.81% use CUBIC', 13.66% use CUBIC", and 14.5%∼25.66% use CTCP. We can clearly see that today's Internet is very different from the Internet several years ago when all web servers used RENO. This motivates our work presented in this paper.

Our contribution: The problem that we address in this paper is: *What's the minimum router buffer size to maintain a certain percentage of the link utilization, and/or to control the packet drop rate, and/or to control the queuing delay, in cases where a large number of long TCP flows controlled by the 5 widely*

978-1-4799-3214-6/13/$31.00 ©2013 IEEE

deployed TCP protocols traverse a single bottleneck link? By "long" TCP flows, we mean that these TCP flows transmit large files and have long durations. These flows spend most of their lifetime in the congestion avoidance state.

Our main results are:

- The minimum router buffer size that satisfies the link utilization constraint is considerably smaller than the BDP of a link when a large number of TCP flows traverse a single bottleneck link. Roughly speaking, it is inversely proportional to the square root of the number of the TCP flows. This means that less router buffer is needed to maintain the link utilization as the number of TCP flows increases. The minimum router buffer size also depends on the TCP protocol mix, i.e. the percentage of the TCP flows using each TCP protocol. Different TCP protocols require different minimum buffer sizes to maintain the same link utilization. For example, The traditional RENO requires a larger router buffer size than BIC.
- Adding the packet drop rate constraint to the link utilization constraint may or may not increase the router buffer size significantly, depending on the TCP protocol mix and the packet drop rate threshold. Some TCP protocols, like RENO, require more router buffer even if we want to control the packet drop rate below a moderate threshold. However, other TCP protocols, such as BIC, only need a small increase of the router buffer size in order to control the packet drop rate below a very low threshold.
- The queuing delay constraint can compromise both the link utilization constraint and the packet drop constraint. The compromise is dependent on the number of TCP flows and the TCP protocol mix.

Paper organization: The organization of the paper is as follows: in Section II, we discuss the background and the related work; in Section III, we formulate the router buffer sizing as a statistical problem; in Sections IV, V, and VI, we analyze the minimum router buffer size that satisfies the link utilization constraint, the packet drop rate constraint, and the queuing delay constraint, respectively; in Section VII, we discuss the tradeoff when we combine all three constraints together; finally, we present the evaluation of our results in Section VIII and conclude the paper in Section IX.

II. Background and Related Work

The TCP traffic has several unique characteristics making the router buffer sizing an important problem for the performance of the Internet. Firstly, the TCP traffic has a bandwidth probing process where a TCP flow increases its traffic to probe the available bandwidth. Secondly, the TCP traffic has a backoff process where a TCP flow reduces its traffic when it detects a congestion event.

Taking the RENO protocol for example, the bandwidth probing process consists of two parts: (1) the slow start stage where a TCP flow doubles its traffic every round-trip time (RTT) and (2) the congestion avoidance stage where a TCP flow increases its traffic by one data packet every RTT. The backoff process consists of one part, i.e. the loss recovery stage where a TCP flow halves its traffic rate if it receives at least 3 duplicated ACKs, or resets its traffic rate to an initial value when the retransmit timer expires.

In a bandwidth probing process, a TCP flow experiences a congestion event when its traffic exceeds the available bandwidth. The congestion triggers the backoff process that reduces the TCP traffic. This can possibly cause a link to be underutilized. This characteristic makes the router buffer sizing very challenging. On the one hand, a small router buffer may cause a link underutilized. On the other hand, a large router buffer can cause the bufferbloat problem [11].

Previous studies such as [12] tried to achieve the tradeoff by calculating the minimum router buffer size to maintain the link utilization. The rule-of-thumb is to set the router buffer size to the BDP of a link. However, this rule may overprovision the router buffer size. Recent studies, such as [3], showed that the router buffer of a link with a large number of TCP flows can be significantly smaller than the BDP of the link and still be able to maintain the link utilization. Specifically, Appenzeller and Keslassy [3] proved that the minimum router buffer size to maintain the link utilization is inversely proportional to the square root of the number of the TCP flows. Jiang and Dovrolis [13] imposed a constraint on the packet drop rate when calculating the minimum router buffer size. Although the resulting minimum router buffer size is larger in [13] than in [3], the former can lead to a lower packet drop rate than the latter, and thus improves the TCP performance.

Our study differs from the previous studies in that we consider the TCP flows using multiple different TCP protocols, whereas the previous studies still assume that all the TCP flows use the same RENO protocol. Different TCP protocols have different traffic characteristics. Some may require larger router buffer sizes than the other protocols. Thus, the results from the previous studies cannot directly apply to today's Internet where the TCP flows are controlled by heterogeneous TCP protocols.

III. Router Buffer Sizing - A Statistical View

In this section, we discuss the minimum router buffer size that can satisfy the link utilization constraint, and/or the packet drop constraint, and/or the queuing delay constraint. The importation notation used in the paper is summarized in Table I.

For a bottleneck link shared by n long TCP flows with the same RTT, let's denote its bandwidth-delay product by BDP, the corresponding router buffer size by B, and the aggregate traffic sent by the n TCP flows in one RTT by W. The aggregate traffic W is nothing but the sum of the congestion windows of all the TCP flows. Different TCP flows may change their congestion windows differently. For example, some may increase their congestion window sizes, some may decrease their congestion window sizes, and the amount of increase and decrease may be different too. As a result, the aggregate traffic W is random.

For the link utilization constraint, the aggregate traffic W should fill up the BDP to keep the link fully utilized, i.e.

TABLE I
NOTATION USED IN THE PAPER

Symbol	Description
n	The number of long TCP flows sharing the bottleneck link
BDP	The bandwidth-delay product of a link
B	The router buffer size
W	The aggregate traffic sent by n TCP flows in one RTT
μ	The mean of the aggregate traffic W
δ	The standard deviation of the aggregate traffic W
τ	The threshold for the link utilization guarantee probability
φ	The upper bound of the packet drop rate
ρ	the threshold for the packet drop rate guarantee probability
λ	The upper bound of the packet queueing delay
C	The capacity of the bottleneck link
T	The duration of a TCP congestion epoch
a	The characteristic value of a TCP protocol

$W \geq BDP$. Due to randomness of the aggregate traffic W, we can not guarantee that $W \geq BDP$ all the time. Instead, we require that the probability of $W \geq BDP$, denoted by $P(W \geq BDP)$, is above a threshold τ. For example, a τ of 0.99 means $W \geq BDP$ with a probability of at least 0.99, which is equivalent to a link utilization of at least 99% in a statistical sense.

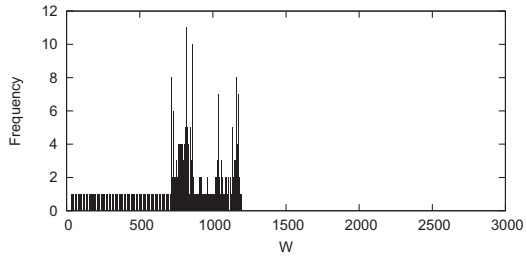
For the packet drop constraint, we bound the packet drop rate by φ. This means that the number of packets dropped in one RTT cannot exceed φW. Since the aggregate traffic W contains the traffic in the link (i.e., BDP), the traffic in the router buffer (i.e., $\leq B$), and the traffic dropped (i.e., φW), the aggregate traffic W should satisfy $W \leq BDP + B + \varphi W$ in order to control the packet drop rate below φ. Due to the randomness of the aggregate traffic W, we can only guarantee that the probability $P(W \leq BDP + B + \varphi W)$ is above a threshold ρ. A ρ close to 1 means that the packet drop rate is below φ with a high probability.

For the queuing delay constraint, we bound the queuing delay by λ seconds. This puts an upper bound on the router buffer size, which is $C\lambda$, where C is the bandwidth of the bottleneck link.

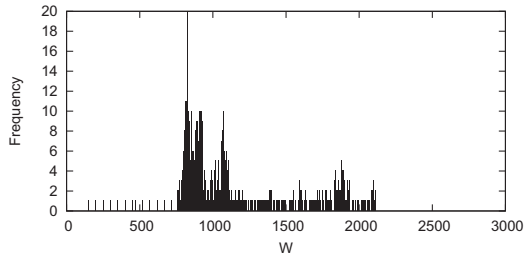
Considering the randomness of the aggregate traffic W, we restate the problem we address below:

What's the minimum router buffer size so that the probability $P(W \geq BDP)$ is above the threshold τ, and/or the probability $P(W \leq BDP + B + \varphi W)$ is above the threshold ρ, and/or the queuing delay is below λ seconds?

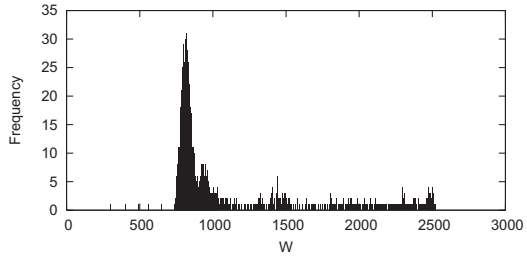
In order to answer the above problem, we need to know the distribution of the aggregate traffic W. The aggregate traffic W, being the sum of the congestion windows of the n TCP flows, roughly follows the normal distribution if n is sufficiently large according to the Central Limit Theorem. We provide an empirical evaluation of this result using the following example. Let n TCP flows traverse an OC-12 link (about 622 Mbps) with a delay of 100 ms. We present the histograms of the aggregate traffic W as n grows from 10 to 200 in Figure 1. As the figure shows, the aggregate traffic W approaches the normal distribution as the number of the TCP flows increases.



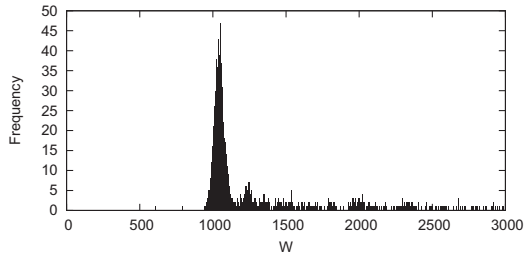
(a) 10 TCP flows



(b) 50 TCP flows



(c) 100 TCP flows



(d) 200 TCP flows

Fig. 1. As the number of TCP flows increases, the aggregate traffic W approaches the normal distribution.

Based on the normal distribution of the aggregate traffic W, we use Figure 2 to demonstrate the relationship among the key variables in the above problem. To satisfy the link utilization constraint, the router buffer size should be chosen so that the grey area on the left side is at most $1-\tau$. To satisfy the packet drop constraints, the router buffer size should be chosen so that the grep area on the right side is at most $1-\rho$. To satisfy the queuing delay constraint, the router buffer size should be no more than $C\lambda$.

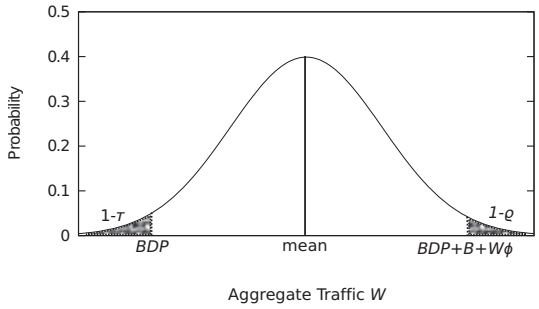


Fig. 2. Relationship among key variables

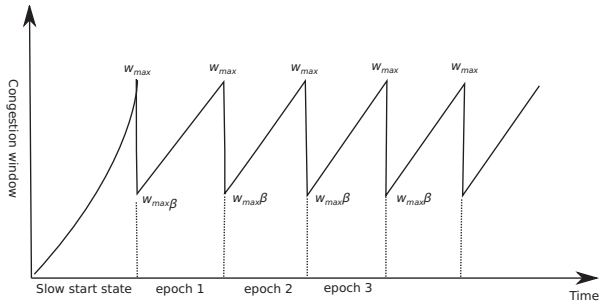


Fig. 3. The congestion window evolution of the RENO protocol. The other TCP protocols have different growth curves. However, the overall pattern is similar for all the TCP protocols.

IV. Link Utilization Constraint

In this section, we analyze the first part of the above problem, i.e. what's the minimum router buffer size so that the probability $P(W \geq BDP)$ is above the threshold τ. In the following discussion, we use μ to denote the mean of the aggregate traffic W, and use δ to denote its standard deviation. We have the following theorem:

Theorem 1: The mean and the standard deviation of the aggregate traffic must satisfy the following inequality:

$$\mu \geq z_\tau \delta + BDP \tag{1}$$

so that the probability $P(W \geq BDP)$ is above the threshold τ. z_τ is the z-score corresponding to the probability $2\tau - 1$.

Proof: Standardizing the aggregate traffic W yields the following equation:

$$P(W \geq BDP) = P(\frac{W-\mu}{\delta} \geq \frac{BDP-\mu}{\delta})$$

$\frac{BDP-\mu}{\delta}$ must be smaller than or equal to z-score $-z_\tau$ so that the probability $P(\frac{W-\mu}{\delta} \geq \frac{BDP-\mu}{\delta})$ is above the threshold τ. This gives us:

$$\mu \geq z_\tau \delta + BDP$$

This proves the theorem. ■

Note that, as the standard deviation δ increases, the aggregate traffic mean μ must also increase so that Inequality (1) is satisfied. A large aggregate traffic mean requires a large router buffer size to accommodate. In this sense, Inequality (1) is consistent with our common understanding that a large router buffer size is helpful for maintaining the link utilization by absorbing large traffic variation. We further verify Inequality (1) in Section VIII.

A. Analyzing each individual TCP flow

In order to obtain the minimum router buffer size that satisfies Inequality (1), we study the mean and the standard deviation of the aggregate traffic. According to the Central Limit Theorem:

$$\mu = \sum_{i=1}^{n} \mu_i \tag{2}$$

$$\delta = \sqrt{\sum_{i=1}^{n} \delta_i^2} \tag{3}$$

where μ_i and $\delta_i, i = 1, 2, ..., n$ are the congestion window mean and the standard deviation of the TCP flow $i, i = 1, 2, ..., n$, respectively. Given that all the TCP protocols considered in this paper are TCP-friendly, we assume that the congestion window mean of each TCP flow is roughly the same, that is, $\mu_i = \frac{\mu}{n}$. However, different TCP flows can have different congestion window standard deviations depending on the TCP protocols they use.

In the following discussion, we analyze the congestion window standard deviation of all the TCP protocols considered in this paper. We adopt the following conventions to simplify the analysis:

- We consider the congestion window standard deviation only in the congestion avoidance state. This simplification can provide good results for the long TCP flows.
- We denote the maximum congestion window size in the congestion avoidance state by w_{max}, and assume that a TCP flow can always increase its congestion window to w_{max} before it detects a congestion event.

In a TCP protocol, the congestion window standard deviation depends on the congestion window evolution in the congestion avoidance state. Specifically, it is determined by a characteristic of a TCP protocol, called congestion window growth function. At the beginning of the evolution, the congestion window starts to grow from a value referred to as the slow start threshold ($ssthresh$). $ssthresh$ is set according to w_{max} and another multiplicative decreasing parameter, denoted by β. The relationship is $ssthresh = \beta w_{max}$. The growth stops when the congestion window reaches w_{max} and a congestion event is detected. After recovering from the congestion, the congestion window starts to grow again from $ssthresh$. The congestion window evolution pattern is illustrated in Figure 3.

The congestion window growth function can be described by the curve in one epoch. Let $w_i(t)$ denote the congestion window growth function of TCP flow i. Let T denote the epoch duration that is the time taken for the congestion window to grow from $ssthresh$ to w_{max}. The congestion window mean μ_i in an epoch can be obtained as:

$$\mu_i = \frac{\int_0^T w_i(t)dt}{T} \tag{4}$$

TABLE II
CONGESTION WINDOW GROWTH FUNCTION AND THE RELATED PARAMETERS (R IN THE TABLE IS THE RTT)

TCP protocol	Growth function $w(t)$	β	Epoch duration T
RENO	$\beta w_{max} + \frac{1}{R}t$	0.5	$(1-\beta)w_{max}R$
BIC	$\beta w_{max} + 0.1 w_{max}\frac{t}{R}$	0.8	R
CUBIC'	$0.4(t-T)^3 + w_{max}$	0.8	$\sqrt[3]{w_{max} \times \beta / 0.4}$
CUBIC"	$0.4(t-T)^3 + w_{max}$	0.7	$\sqrt[3]{w_{max} \times \beta / 0.4}$
CTCP	$(\frac{t}{32} + (\beta w_{max})^{0.25})^4$	0.5	$5.1 \times \sqrt[4]{w_{max}}$

TABLE III
CONGESTION WINDOW MEAN AND STANDARD DEVIATION

TCP protocol	Mean μ_i	Standard deviation δ_i
RENO	$\frac{3}{4}w_{max}$	$\frac{1}{3\sqrt{3}}\mu_i$
BIC	$0.85 w_{max}$	$0.0241\mu_i$
CUBIC'	$0.95 w_{max}$	$0.06\mu_i$
CUBIC"	$0.925 w_{max}$	$0.09\mu_i$
CTCP	$0.73 w_{max}$	$0.19\mu_i$

Once we obtain the congestion window mean, the congestion window standard deviation δ_i can be calculated by:

$$\delta_i = \sqrt{\frac{\int_0^T (w_i(t) - \mu_i)^2 dt}{T}} \tag{5}$$

We list the congestion window growth functions and the related parameters for all the five TCP protocols in Table II. According to this table and Equations (4) and (5), we calculate the congestion window mean and the standard deviation, and the results are summarized in Table III.

Table III associates the congestion window standard deviation with its mean. In general, the congestion window standard deviation δ_i can be expressed by $a_i\mu_i$. We refer to a_i as the characteristic value of the TCP flow i, and is determined by the TCP protocol used by the flow. For example, if the TCP flow i uses the RENO protocol, a_i is equal to $\frac{1}{3\sqrt{3}}$. The characteristic value reflects the congestion window variation of a TCP protocol, and is an important factor in determining the minimum router buffer size.

B. Analyzing the aggregate TCP flows

Using the characteristic values of TCP flows, we obtain the relationship between the standard deviation and the mean of the aggregate traffic:

$$\delta = \sqrt{\sum_{i=1}^{n} \delta_i^2} = \sqrt{\sum_{i=1}^{n} (a_i\mu_i)^2} = \frac{a}{\sqrt{n}}\mu \tag{6}$$

where $a = \sqrt{(\sum_{i=1}^{n} a_i^2)/n}$. Inequality (1) and Equation (6) together provide us a way to calculate the minimum router buffer size B that satisfies the link utilization constraint, as shown in the following theorem:

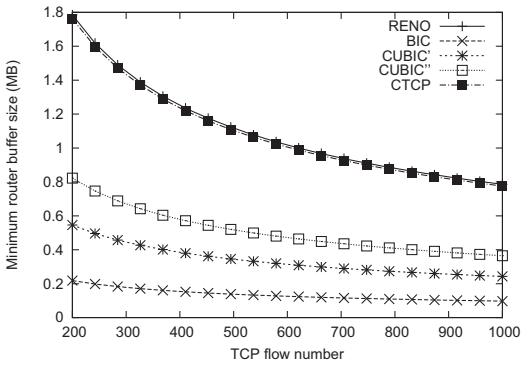


Fig. 4. Minimum router buffer size as n TCP flows of the same TCP protocol traverse a bottleneck link.

Theorem 2: The minimum router buffer size B that satisfies the link utilization constraint is

$$B \geq \frac{BDP}{1 - \frac{z_\tau a}{\sqrt{n}}} - BDP \tag{7}$$

Proof: Substituting the equation $\delta = \frac{a}{\sqrt{n}}\mu$ into Inequality (1), we get:

$$\mu \geq \frac{BDP}{1 - \frac{z_\tau a}{\sqrt{n}}} \tag{8}$$

Since the aggregate traffic mean μ has to be accommodated by the BDP and the router buffer size B, we have:

$$B + BDP \geq \mu \geq \frac{BDP}{1 - \frac{z_\tau a}{\sqrt{n}}} \tag{9}$$

i.e.

$$B \geq \frac{BDP}{1 - \frac{z_\tau a}{\sqrt{n}}} - BDP$$

This proves the theorem. ■

According to Inequality (7), we plot the minimum router buffer size as $n \in [200, 1000]$ TCP flows of the same TCP protocol traverse a bottleneck link with a capacity of 622Mbps and a delay of 100ms in Figure 4.

We can see that the minimum router buffer size is a decreasing function of $\sqrt{n}$, i.e., the square root of the number of TCP flows. This means that as the number of TCP flows increases, we need less router buffer size to satisfy the link utilization constraint.

We can also see that the TCP protocols with larger characteristic values need a larger router buffer size to satisfy the link utilization constraint. Therefore, the characteristic value can reflect the congestion window variation of a TCP flow. Moreover, the parameter a can reflect the variation of the aggregate traffic. If the TCP flows use more than one TCP protocol, the corresponding curve in Figure 4 should be between the curve for the RENO protocol and the curve for the BIC protocol. If more TCP flows use the TCP protocols

with larger characteristic values, it is closer to the curve for the RENO protocol. Otherwise, it is closer to the curve for the BIC protocol.

V. Packet Drop Rate Constraint

In this section, we analyze the second part of the problem, i.e. what's the minimum router buffer size so that the probability $P(W \leq BDP + B + \varphi W)$ is above the threshold ρ? We have the following theorem.

Theorem 3: The router buffer size B should be at least $(\frac{z_\rho a}{\sqrt{n}} + 1)(1-\varphi)\mu - BDP$ so that the packet drop rate is below the threshold φ. z_ρ is the z-score corresponding to the probability $2\rho - 1$.

Proof: We standardize the aggregate traffic W. This yields:

$$\begin{aligned} P(W \leq BDP + B + \varphi W) &= P(W \leq \frac{BDP + B}{1-\varphi}) \\ &= P(\frac{W-\mu}{\delta} \leq \frac{\frac{BDP+B}{(1-\varphi)} - \mu}{\delta}) \end{aligned}$$

$\frac{\frac{BDP+B}{(1-\varphi)} - \mu}{\delta}$ should be bigger than or equal to the z-score z_ρ so that the probability $P(\frac{W-\mu}{\delta} \leq \frac{\frac{BDP+B}{(1-\varphi)} - \mu}{\delta})$ is above the threshold ρ. This yields:

$$B \geq (z_\rho \delta + \mu)(1-\varphi) - BDP \tag{10}$$

Substituting the equation $\delta = \frac{a}{\sqrt{n}}\mu$ into Inequality (10), we get:

$$B \geq (\frac{z_\rho a}{\sqrt{n}} + 1)(1-\varphi)\mu - BDP \tag{11}$$

This proves the theorem. ∎

We can see that the router buffer size is an increasing function of the characteristic value a. A large characteristic value a causes large variation of the aggregate traffic. This needs a large router buffer size to limit the number of the packets dropped. Inequality (11) is verified in Section VIII.

Surprisingly, Inequality (11) is satisfied if

$$(\frac{z_\rho a}{\sqrt{n}} + 1)(1-\varphi) \leq 1 \tag{12}$$

regardless of the router buffer size due to the fact that $B + BDP \geq \mu$. Using the threshold $\rho = 0.99$, we plot the parameter φ and the number of TCP flows n that satisfy Inequality (12) in Figure 5. The bottleneck link in this case is a link with a capacity of 622Mbps and a delay of 100ms. For each curve in the figure, the area above the curve contains the (φ, n) pairs that satisfy Inequality (12). For example, to only control the packet drop rate below 1%, we need at least 24 TCP flows for the BIC protocol, or 1526 TCP flows for the RENO protocol to traverse the bottleneck link. The required number of TCP flows for other TCP protocols and for TCP protocol mixes are between these two numbers.

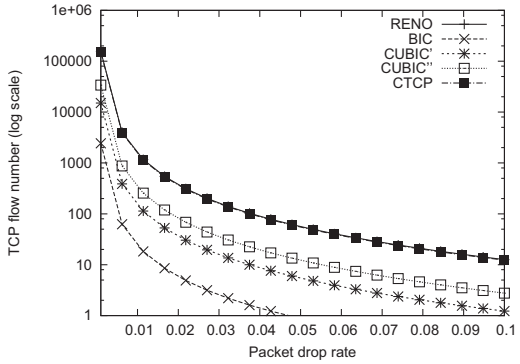


Fig. 5. Packet drop rate bound φ and TCP flow number n that satisfy Inequality (12).

VI. Queuing Delay Constraint

The third part of the problem considers what the maximum router buffer size is so that the queuing delay is below λ seconds. Let C denote the capacity of the bottleneck link, and then a maximum queuing delay of λ seconds corresponds to a maximum router buffer size of $C\lambda$. We will discuss the impact of queueing delay constraint in the next section.

VII. Combining All the Constraints

In this section, we analyze the feasible region of the router buffer size when combining all three constraints together.

A. Combining the Link utilization and the Packet Drop Rate Constraints

In order to satisfy both the link utilization constraint and the packet drop rate constraint, the router buffer size must satisfy Inequality (7) and the aggregate traffic mean must satisfy Inequality (8). Thus, we have:

$$\begin{aligned} B &\geq (\frac{z_\rho a}{\sqrt{n}} + 1)(1-\varphi)\mu - BDP \\ &\geq (\frac{z_\rho a}{\sqrt{n}} + 1)(1-\varphi)\frac{BDP}{1 - \frac{z_\tau a}{\sqrt{n}}} - BDP \end{aligned} \tag{13}$$

When $(\frac{z_\rho a}{\sqrt{n}} + 1)(1-\varphi)$ is smaller than 1, the router buffer size in Inequality (13) cannot satisfy the link utilization constraint, i.e. Inequality (7). To solve this problem, we bound the router buffer size by both Inequality (7) and Inequality (13):

$$\begin{aligned} B \geq \max\{&\frac{BDP}{1 - \frac{z_\tau a}{\sqrt{n}}} - BDP, \\ &(\frac{z_\rho a}{\sqrt{n}} + 1)(1-\varphi)\frac{BDP}{1 - \frac{z_\tau a}{\sqrt{n}}} - BDP\} \end{aligned} \tag{14}$$

Let's use an example to demonstrate how much extra router buffer size we need to satisfy both the link utilization constraint and the packet drop constraint. Setting $\rho = 0.99$ and $n = 200$, we plot the router buffer size increase as the packet drop rate threshold φ grows from 0.001 to 0.01, as shown in Figure 6 . The bottleneck link in this case is a link with a capacity of 622Mbps and a delay of 100ms.

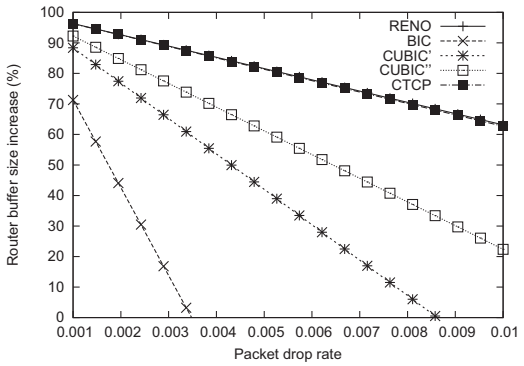


Fig. 6. Router buffer size increase by considering the packet drop rate constraint in addition to the link utilization constraint.

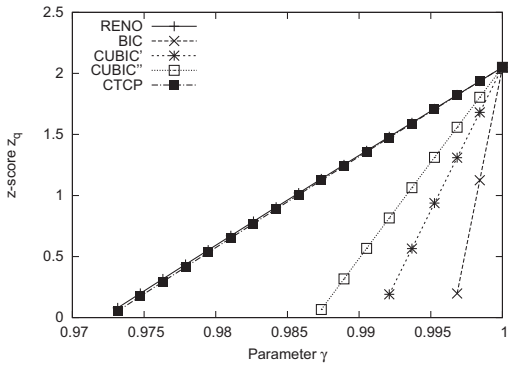


Fig. 7. Lower bound of the z-score z_q as the parameter γ grows from $\frac{BDP}{BDP+B_l}$ to 1.

The overall observation is that it depends on the TCP protocols mix and the packet drop rate threshold φ. For example, when the TCP flows use the BIC protocol, no router buffer size increase is required to satisfy the packet drop rate around 0.0035. However, when the TCP flows use the RENO protocol, an 87% increase of the router buffer size is required to keep the packet drop rate around 0.0035. The router buffer size increase for the other TCP protocols and TCP protocol mixes are between these two extremes.

B. Adding the Queuing Delay Constraint

The queuing delay constraint puts an upper bound on the router buffer size, i.e. $C\lambda$. If $C\lambda$ is bigger than or equal to the bound in Inequality (14), all three constraints can be satisfied at the same time. Otherwise, we have to compromise the other two constraints. In the following discussion, we analyze how much compromise we should make in order to satisfy the queuing delay constraint. We define the compromise by the reduction of the threshold τ for the link utilization constraint, and by the increase of the packet drop rate for the packet drop rate constraint.

1) Compromise on the Link Utilization Constraint: Let B_l denote the router buffer size, μ_l denote the aggregate traffic mean, and δ_l denote the standard deviation when the link utilization constraint is satisfied. Further, let B_l and μ_l satisfy the boundary condition of Inequality (9), i.e.,

$$B_l + BDP = \mu_l = \frac{BDP}{1 - \frac{z_\tau a}{\sqrt{n}}} \tag{15}$$

Now, we set the router buffer size to $C\lambda$, and denote the the aggregate traffic mean, and the standard deviation by μ_q and δ_q, respectively in this case. We have the following relationship:

$$\mu_q \leq BDP + C\lambda = \gamma(BDP + B_l) = \gamma\mu_l$$

and

$$\delta_q = \frac{a}{\sqrt{n}}\mu_q \leq \gamma\frac{a}{\sqrt{n}}\mu_l = \gamma\delta_l$$

where γ is a parameter between 0 and 1 that represents the compromise on the router buffer size.

In order to calculate the reduction on the threshold τ, we need to compare the z-scores in these two cases: (1) z_τ when the link utilization constraint is satisfied, and corresponds to the threshold τ, and (2) z_q when the router buffer size is set to $C\lambda$, and corresponds to a threshold q. Then, the reduction can be calculated as $\tau - q$. The z-score z_τ can be expressed as:

$$z_\tau = \frac{\mu_l - BDP}{\delta_l} = \frac{\sqrt{n}}{a} - \frac{BDP}{\delta_l} \tag{16}$$

The z-score z_q can be expressed as:

$$z_q = \frac{\mu_q - BDP}{\delta_q} \leq \frac{\gamma\mu_l - BDP}{\gamma\delta_l} = \frac{\sqrt{n}}{a} - \frac{BDP}{\gamma\delta_l} \tag{17}$$

Based on Equations (16) and (17), we have the following relationship:

$$z_q \leq \frac{z_\tau}{\gamma} - (\frac{1}{\gamma} - 1)\frac{\sqrt{n}}{a} \tag{18}$$

Inequality (18) provides us with a way to calculate the upper bound of the threshold q by looking up the probability $2q-1$ corresponding to $\frac{z_\tau}{\gamma} - (\frac{1}{\gamma} - 1)\frac{\sqrt{n}}{a}$ in the z-score table.

Let's use an example to demonstrate the reduction on the threshold τ. Suppose there are 200 TCP flows of the same TCP protocol traversing a bottleneck link with a capacity of 622Mbps and a delay of 100ms. The threshold τ is set to 0.99. We plot the lower bound of the z-score z_q as the parameter γ grows from $\frac{BDP}{BDP+B_l}$ to 1 in Figure 7.

On the one hand, if $\gamma = 1.0$, there is no compromise on the router buffer size, and the z-score z_q equals the z-score z_τ. This is equivalent to $q = \tau$, i.e., no reduction on the threshold τ. On the other hand, when $\lambda = 0$ (i.e., we do not want any queuing delay), the z-score z_q almost equals 0. This is equivalent to $2q - 1 = 0.5$, i.e., $q = 0.75$. The reduction on the threshold τ is $\tau - q = 0.99 - 0.75 = 0.24$. Other values of the parameter γ lead to a reduction on the threshold τ between these two extremes.

2) Compromise on the Packet Drop Rate Constraint: Let's denote the router buffer size that satisfy the packet drop rate constraint by B_p, and denote the decrease on the router buffer

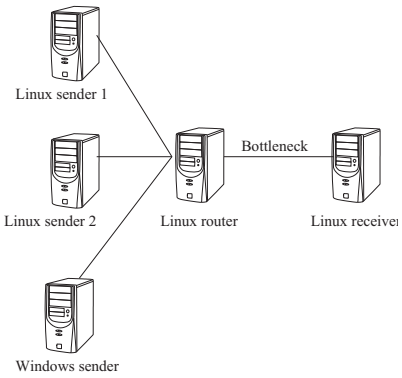


Fig. 8. Testbed

TABLE IV
NETEM PARAMETERS

Link / RTT	OC-12 (622Mbps)
28ms	2.3 MBytes (BDP1)
67ms	5.5 MBytes (BDP2)
90ms	18 MBytes (BDP3)

size by $D = B_p - C\lambda$, if the router buffer size is set to $C\lambda$. According to the packet drop rate constraint:

$$\begin{aligned} &P(W \leq BDP + B_p + \varphi W) \\ =&P(W \leq BDP + C\lambda + D + \varphi W) \\ \geq&\rho \end{aligned}$$

With the threshold ρ, at most φW packets out of W packets are dropped if the router buffer size is B_p. Under the the same threshold ρ, if the router buffer size is $C\lambda$, at most $D + \varphi W$ packets out of W packets are dropped. The packet drop rate is increased by

$$\frac{D + \varphi W}{W} - \frac{\varphi W}{W} = \frac{D}{W}$$

Since the aggregate traffic W is no larger than $BDP + B_p$, we obtain a lower bound on the increase of the packet drop rate:

$$\frac{D}{W} \geq \frac{D}{BDP + B_p} \tag{19}$$

VIII. EVALUATION

We evaluate our result using a testbed, which is is shown in Figure 8: two Linux senders, one Windows sender, and one linux receiver, all connected to a Linux router. On the linux router, we use Netem [14] to emulate 3 different BDPs, referred to as BDP1, BDP2 and BDP3, respectively, as shown in Table IV. The 3 RTT values in Table IV are the 10th percentile, the 50th percentile and the 54th percentile of the web server RTT samples measured in 2011 [4][1].

[1]Due to the testbed limitation, we cannot test larger delay. The resulting BDPs cannot be fully utilized

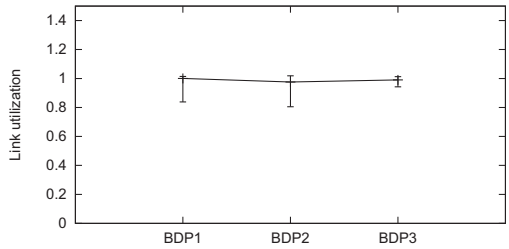


Fig. 9. Router buffer sizes set according to Inequality (7) are able to maintain a high link utilization.

The Linux receiver and router are running CentOS 5.7 (kernel version 2.6.18); the first Linux sender is running CentOS 5.7 with a custom 2.6.27 kernel capable of sending TCP flows with different TCP protocols simultaneously; the second Linux sender is running CentOS 5.7 with a 2.6.25 kernel for sending CUBIC' traffic; the Windows sender is running Windows 7 for sending CTCP traffic. The TCP flows between the receiver and the senders are HTTP traffic generated by the Apache web server running on the Linux senders and the IIS web server running on the Windows sender, respectively. We use the following TCP protocol mix for the TCP flows: 42 RENO flows, 51 BIC flows, 30 CUBIC' flows, 41 CUBIC" flows, and 34 CTCP flows, to simulate their actual deployment [4] in the current Internet.

Firstly, we set the router buffer size according to Inequality (7) to satisfy the link utilization constraint. We require that the probability $P(W \geq BDP) \geq 0.99$, i.e, the link is utilized at least 99%. The actual link utilization is shown in Figure 9. For each BDP, we plot three related values using an error bar: the point in the middle indicates the average link utilization throughout the test; the upper bar indicates the highest link utilization; and the lower bar indicates the lowest link utilization. The figure shows that the router buffer sizes set according to Inequality (7) are able to maintain a high link utilization for the 3 different BDPs.

The router buffer size that satisfies the link utilization constraint is based on Inequality (1). We verified this inequality in Figure 10. The "+" points represent the measurements of the aggregate traffic mean under the three BDP settings. The "$\times$" represent the lower bounds calculated according to Inequality (1). Within reasonable errors, the aggregate traffic mean closely matches the lower bound.

Secondly, we add the packet drop rate constraint and set the router buffer size according to Inequality (14). We require that the probability $P(W \geq BDP) \geq 0.99$, and the probability $P(W \leq BDP + B + \varphi W) \geq 0.99$ with $\varphi = 0.01$. The actual link utilization and the packet drop rates are shown in Figure 11. The figure shows that the router buffer sizes according to Inequality (14) are able to maintain a low packet drop rate and a high link utilization at the same time: the lowest packet drop rate achieved is 0.005, which is below the specified packet drop rate $\varphi = 0.01$; the lowest average link utilization is 0.9739. The packet drop rate constraint adds a

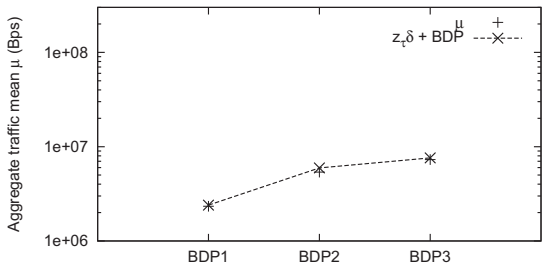


Fig. 10. Inequality (1) is satisfied by the aggregate traffic.

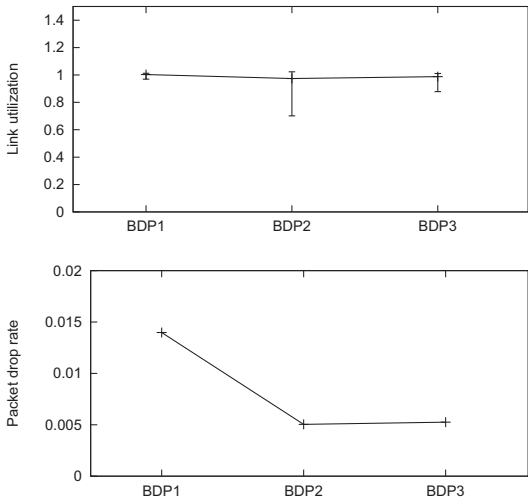


Fig. 11. Router buffer sizes set according to Inequality (14) are able to maintain a high link utilization and a low packet drop rate.

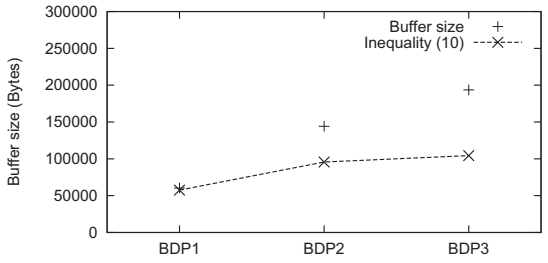


Fig. 12. Inequality (10) is satisfied by the aggregate traffic and the router buffer size.

significant increase to the router buffer size. The router buffer size is increased by 45.45% under all three BDP settings.

The router buffer size that satisfies the packet drop rate constraint is based on Inequality (10). We verified this inequality in Figure 12. The "+" points represent the router buffer sizes under the three BDP settings. The "×" points represent the lower bounds calculated according to Inequality (10). In all three BDP settings, the router buffer sizes are bigger than the lower bounds, which indicates that Inequality (10) is satisfied.

IX. Conclusion

We investigated the router buffer sizing problem for the heterogenous TCP protocols under the link utilization, the packet drop rate, and the queuing delay constraints and discussed the tradeoffs among them. We confirmed the previous studies that the router buffer size for maintaining the link utilization can be significantly lower than the traditional rule-of-thumb, i.e. the BDP. As the number of TCP flows traversing the bottleneck link increases, the router buffer size can be further reduced without compromising the link utilization. Roughly speaking, the required router buffer size is inversely proportional to the square root of the number of TCP flows. Adding packet drop rate constraint to the link utilization constraint can potentially increase the required router buffer size. Enforcing the queuing delay constraint can compromise both the link utilization and the packet drop rate, and we analyzed the amount of the compromise. Besides, the router buffer size also depends on the TCP protocol mix. Different TCP protocols have different traffic characteristics, and require different router buffer sizes to maintain the same link utilization and the packet drop rate. We defined the characteristic value for a certain TCP protocol mix. The router buffer size that satisfies the link utilization constraint and the packet drop rate constraint is a function of this characteristic value.

References

[1] V. Jacobson, "Congestion avoidance and control," in *Proceedings of ACM SIGCOMM*, Stanford, CA, August 1988.

[2] D. Chiu and R. Jain, "Analysis of the increase/decrease algorithms for congestion avoidance in computer networks," *Journal of Computer Networks and ISDN*, vol. 17, no. 1, pp. 1–14, June 1989.

[3] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," in *Proceedings of ACM SIGCOMM*, Portland, Oregon, August 2004.

[4] P. Yang, J. Shao, W. Luo, L. Xu, J. Deogun, and Y. Lu, "TCP congestion avoidance algorithm identification," *IEEE Transactions on Networking*, 2013 (to appear), http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6594906.

[5] M. Allman, V. Paxson, and E. Blanton, "TCP congestion control," *RFC 5681*, September 2009.

[6] S. Floyd, T. Henderson, and A. Gurtov, "The NewReno modification to TCP's fast recovery algorithm," *RFC 3782*, April 2004.

[7] M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, "TCP selective acknowledgment options," *RFC 2018*, October 1996.

[8] L. Xu, K. Harfoush, and I. Rhee, "Binary increase congestion control for fast long-distance networks," in *Proceedings of IEEE INFOCOM*, Hong Kong, March 2004.

[9] S. Ha, I. Rhee, and L. Xu, "CUBIC: A new TCP-friendly high-speed TCP variant," *ACM SIGOPS Operating System Review*, vol. 42, no. 5, pp. 64–74, July 2008.

[10] K. Tan, J. Song, Q. Zhang, and M. Sridharan, "A compound TCP approach for high-speed and long distance networks," in *Proceedings of IEEE INFOCOM*, Barcelona, Spain, April 2006.

[11] J. Gettys and K. Nichols, "Bufferbloat: Dark buffers in the Internet," *Communications of the ACM*, vol. 55, no. 1, pp. 57–65, January 2011.

[12] C. Villamizar and C. Song, "High performance TCP in ANSNET," *ACM Computer Communications Review*, vol. 24, no. 5, pp. 45–60, 1994.

[13] H. Jiang and C. Dovrolis, "Buffer sizing for congested Internet links," in *Proceedings of IEEE INFOCOM*, Miami, FL, March 2005, pp. 1072–1083.

[14] S. Hemminger, "Network emulation with NetEm," in *Proceedings of the 6th Australia's National Linux Conference*, Australia, April 2005.