Anomalous Loss Performance for Mixed Real-Time and TCP Traffic in Routers with Very Small Buffers

Arun Vishwanath, *Student Member, IEEE,* Vijay Sivaraman, *Member, IEEE,* and George N. Rouskas, *Senior Member, IEEE*

***Abstract*—In the past few years there has been vigorous debate regarding the size of buffers required at core Internet routers. Recent arguments supported by theory and experimentation show that under certain conditions, core router buffer sizes of a few tens of packets suffice for realising acceptable end-to-end TCP throughputs. This is a significant step towards the realisation of optical packet switched (OPS) networks, which are inherently limited in their ability to buffer optical signals. However, prior studies have largely ignored the presence of real-time traffic, which is increasing in importance as a source of revenue for Internet service providers. In this paper we study the interaction that happens between real-time (open-loop) and TCP (closed-loop) traffic when they multiplex at buffers of very small size (few tens of packets), and make a significant discovery, namely that in a specific range of buffer size, real-time traffic losses increase as buffer size becomes larger. Our contributions pertaining to this anomalous behaviour are threefold. First, we exhibit this anomalous loss performance for real-time traffic via extensive simulations using synthetic traffic and real video traces. Second, we develop quantitative models that reveal the dynamics of buffer sharing between real-time and TCP traffic that lead to this behaviour. Third, we show how various factors such as the nature of real-time traffic, mixture of long-lived and short-lived TCP flows, and packet sizes impact the severity of the anomaly. Our study is the first to consider interactions between real-time and TCP traffic in very small (potentially all-optical) buffers, and informs router manufacturers and network operators of the factors to consider when dimensioning such small buffer sizes for desired performance balance between real-time and TCP traffic.**

***Index Terms*—routers with very small buffers, mixed TCP and real-time traffic, anomalous loss performance, optical packet switched networks**

I. Introduction

In recent years there has been vigorous debate on how large buffers at an Internet router should be. Conventional wisdom, attributed to [1], holds that a router should be able to store a round-trip-time worth of data so as to keep the output link fully utilised while TCP ramps up its window size after a loss event; equivalently, this rule-of-thumb mandates buffer size $B = RTT \times W$, where RTT is the average round-trip time of a TCP connection flowing through the router, and W the capacity of the bottleneck link. For typical $RTT = 250$ ms, a router with a $W = 40$ Gbps link would require 10 Gigabits of buffering, which poses a considerable challenge to router design.

This buffer sizing rule was first challenged in 2004 by researchers from Stanford University [2], [3] who showed that when a large number M of long-lived TCP flows multiplex at a bottleneck link, synchronisation does not arise, and near-100% utilisation of the bottleneck link can be achieved with only $B = RTT \times W/\sqrt{M}$ buffers. This means that a router carrying $10,000$ TCP flows need only buffer $10,000$ packets instead of the million packets required by the rule-of-thumb.

Since 2004, several new arguments on buffer sizing have been put forth. Stanford University researchers further proposed in [4], [5], [6] that under certain conditions, which they believe hold in today's Internet, as few as 20-50 packet buffers suffice for TCP traffic to realise acceptable link utilisation, a claim supported by their experimental results at Sprint ATL and Verizon Communications [7]. Measurement study on a Sprint backbone router also found the queue size to seldom exceed 10 packets [8], while the choice of 50 packet buffers is recommended in [9] to guarantee overall stability. These initial results show the feasibility of building all-optical networks that can be operated at 70-80$\%$ utilisation using routers having very small packet buffers. Clearly, the aforementioned results have significant implications from an all-optical router design point of view, where buffering presents a very important but difficult operation, since data is to be retained in the optical domain.

Researchers from Georgia Tech [10] revisited the ongoing buffer sizing debate from the perspective of average per-flow TCP throughput rather than focusing purely on link utilisation. The authors present evidence to suggest that the output/input capacity ratio at a router's interface largely governs the amount of buffering needed at that interface. If this ratio is greater than one, then the loss rate falls exponentially, and only a very small amount of buffering is needed, which corroborates with the results reported in [6]. However, the concern is that, if the output/input capacity ratio is lower than one, then the loss rate follows a power-law reduction and significant buffering is needed. Researchers from the University of Illinois at Urbana-Champaign also arrive at a similar conclusion in [11]. Other studies have considered factors such as application layer performance [12], [13] and fairness [14] influencing buffer sizing. In late 2008, the Stanford group presented experimental results validating the applicability of routers with very small buffers in the core of the Internet [15]. For a comprehensive survey on this topic of router buffer sizing, we refer the reader to our recent survey paper [16].

A. Vishwanath and V. Sivaraman are with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia. E-mails: {arunv@ee.unsw.edu.au, vijay@unsw.edu.au}.

G. N. Rouskas is with the Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206, USA. E-mail: {rouskas@ncsu.edu}.

This submission is an extended version of our papers presented at IEEE INFOCOM 2009 [39] and IEEE IWQoS 2008 [36] conferences.

A. Motivation

While most prior studies on buffer sizing have focused on electronic Internet routers, some earlier works such as [6], [17] have applied buffer sizing principles to optical switches. However, they focus entirely on TCP traffic performance, and ignore the performance implications for real-time traffic. From the observation of traffic in the Internet core, it is widely accepted that TCP constitutes nearly 85-90% of the traffic, while real-time (UDP) accounts for about 5-10%. This has led all previous work to largely ignore the impact of very small buffers on UDP's performance. In this paper, we focus our attention on what happens at a router with very small buffers when both open-loop UDP and closed-loop TCP traffic coexist, and show why it is important to address their joint performance. We use the term real-time, UDP, and open-loop traffic interchangeably.

To understand the dynamics of buffer occupancy at a bottleneck link router, we mixed a small fraction of UDP traffic with TCP traffic and measured the UDP packet loss and end-to-end TCP throughput. Before starting our simulations, our intuition was that in the regime of very small buffers (up to 50 KiloBytes):

1) UDP packet loss would fall *monotonically* with buffer size, and
2) End-to-end TCP throughput would increase with buffer size to saturation.

Surprisingly, our observation was contrary to our intuition. We found that there exists a certain continuous region of buffer sizes (typically starting from 8-10 KB or so) wherein the performance of real-time traffic degrades with increasing buffer size. In other words, packet loss for real-time traffic *increases* as the buffer size *increases* within this region. We call this region of buffer size an "anomalous region" with respect to real-time traffic. More surprisingly, we found that when there are a sufficiently large number of TCP flows, this performance degradation for UDP traffic does not come with any significant improvement in end-to-end TCP throughput, and in fact the inflection point occurs around the buffer size region corresponding to when TCP has nearly attained its saturation throughput.

This phenomenon is important for a number of reasons and forms the motivation for the study in this paper. Firstly, as real-time multimedia applications such as on-line gaming, interactive audio-video services, VoIP and IPTV proliferate in the Internet and become part of an ISP's revenue stream, router buffer sizing studies cannot afford to ignore the performance impact on real-time traffic when it is multiplexed with TCP traffic, which to the best of our knowledge has not been undertaken before in the context of very small buffers.

Secondly, in the regime of very small buffers, it is prudent to size router buffers so as to balance the performance of TCP and UDP traffic. Operating the router buffers in the "anomalous region" can result in increased UDP packet loss, with only a marginal improvement in end-to-end TCP throughput, which is undesirable from a network operator's point-of-view.

Thirdly, our results have important implications for emerging optical packet switched networks since buffering of packets in the optical domain remains a complex and expensive operation. Optical buffering using fibre delay lines (FDLs) [18] is impractical due to the bulk of large spools (1 km of fibre buffers light for only 5μsec) and the large optical crossbar sizes needed to switch packets in and out of the delay lines (see for example the shared memory architecture in [19]). Recent advances in integrated photonic circuits have enabled optical FIFO buffers [20], [21], [22], though their size is limited to a few tens of packets, and is not expected to increase dramatically in the near future. The anomaly revealed by our study shows that the investment made in deploying larger optical buffers has the potential to negatively impact quality of service and lead to worse performance, which could be a significant concern for the operator of the network.

Finally, recent work, such as those in [23] and [24], has proposed the use of adaptive strategies wherein routers adapt their buffer size on-the-fly (depending upon the prevailing network conditions) so as to achieve a desired loss rate and link utilisation. However, these studies rely on the existence of a monotonic relationship between packet loss and buffer size, i.e., that loss rate decreases as buffer size increases. Since our results indicate that this is not always the case, it is important that further studies be wary of the non-monotonicity that may exist when modelling the interaction between TCP and UDP traffic in the regime of very small buffers.

B. Contributions of This Work

In the context of loss performance for real-time (UDP) traffic multiplexed with TCP traffic in routers with very small buffers, our contributions are threefold:

1) We demonstrate via extensive simulations the existence of anomalous loss performance (namely an increase in loss with increasing router buffer size) for real-time traffic. Our simulations consider real video traces, synthetic short-range dependent Poisson and long-range dependent fractional Brownian motion models, and several thousand TCP flows.
2) We develop quantitative models that help explain the anomaly. Our first model intuitively captures the buffer sharing dynamics between real-time and TCP traffic and shows the impact of the latter's greedy nature on the effective buffers available to the former. Our second model takes a more rigorous Markov chain based approach and allows explicit numerical evaluation of packet loss. Both models validate the anomalous loss seen in simulations, and provide an analytical handle to explore the phenomenon in greater depth.
3) We illustrate the impact of several system parameters on the severity of the anomaly. Significant ones include the relative mix of short/long lived TCP flows, characteristics of real-time traffic and distribution of packet sizes.

We believe the phenomenon studied in this paper adds a new dimension to the ongoing debate on buffer sizing, including in the context of optical packet switches, and our results aid switch manufacturers and network operators in selecting small buffer sizes that achieve desired performance balance between TCP and real-time traffic.

The rest of this paper is organised as follows: In Section II, we introduce the anomalous loss behaviour using real video traces. In Section III, we develop an intuitive analytical model that captures this phenomenon succinctly. In Section IV we present a more rigorous analysis using a M/M/1/B queueing model, which further validates the anomaly observed in simulations. In Section V, we study how various network design factors such as real-time and TCP traffic characteristics affect the anomalous loss performance. In Section VI, we investigate the impact of varying UDP packet sizes on the anomaly. We conclude the paper in Section VII and point to directions for future work.

II. The Anomaly

To illustrate the anomalous loss behaviour, we consider a simple dumbbell topology (Fig. 1) that is commonly used to analyse the performance of various congestion control algorithms, including TCP. Such a topology captures packet queueing effects at the bottleneck link, which is the dominant factor in end-to-end loss and delay variation for each flow, while abstracting the rest of the flow path by an access link on either side of the bottleneck link. We use ns-2 [25] (version 2.30) for our simulations and consider 1000 TCP flows, corresponding to each source-destination pair *(s-tcp$_i$,d-tcp$_i$)*, $1 \leq i \leq 1000$. Further, we use TCP-Reno in all our simulations, consistent with the TCP version used in previous related work on buffer sizing, and employ FIFO queue with drop-tail queue management, which is commonly used in most routers today.

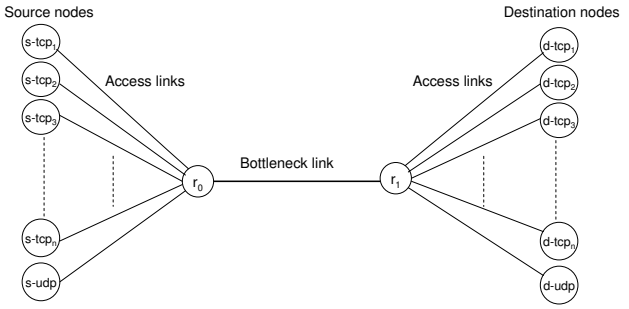


Fig. 1. ns-2 simulation topology

UDP traffic is generated between nodes *(s-udp,d-udp)*. It suffices to have a single UDP flow since open-loop traffic can, without loss of generality, be aggregated. Multiple UDP flows traversing the bottleneck link can thus be modelled as a single UDP flow that represents the aggregate of all individual UDP flows passing through that bottleneck link. However, we need multiple TCP flows since they each react independently to the prevailing network condition and the state of the buffers.

The propagation delay on the UDP access link is chosen at 5 ms, while it is uniformly distributed between $[1, 25]$ ms on the TCP access links. The propagation delay on the bottleneck link (r_0, r_1) is 50 ms; thus round-trip times vary between 102 ms and 150 ms. All TCP sources start at random times between $[0, 10]$ s. UDP source starts at time 0 s. The simulation duration is 800 s and performance measurements are recorded after 200 s, to allow for the stabilisation of all TCP flows.

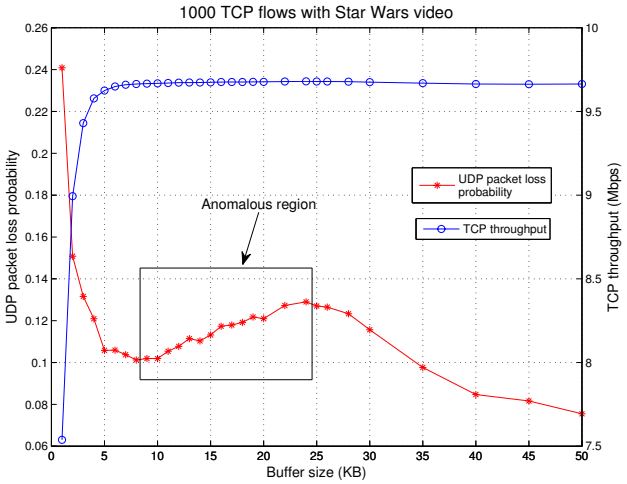


Fig. 2. Starwars 200 Byte packets: UDP loss and TCP throughput

Buffer size at the bottleneck router r_0 is varied in terms of KiloBytes. To set the packet sizes, we draw on the fact that several real-time applications, for e.g. on-line gaming [26], [27] use small UDP packets since they require extremely low latencies. The study showed that almost all packets were under 200 Bytes. Our experiments using Skype and Yahoo Messenger showed that for interactive voice chat, UDP packet sizes were between 150-200 Bytes. Also, traces obtained at a trans-Pacific 150 Mbps link [28] suggests that average UDP packet sizes are smaller than average TCP packet sizes. Therefore, in all our simulations, we fix the TCP packet size at 1000 Bytes and simulate fixed and variable size UDP packets in the range of $[150, 300]$ Bytes.

Akin to the traffic in the Internet core, we want to keep the fraction of UDP traffic to within 3-10% as well. We performed simulations using various movie traces such as Star Wars, Jurassic Park I, Diehard III, Silence of the lambs, Aladdin etc. For brevity, we present results from only a subset of the movies mentioned above. Results for the movies not described here closely follow the ones described. All the movie traces have been profiled and are known to exhibit self-similar and long-range-dependent traffic characteristics.

We first illustrate the phenomenon using the video traffic trace from the movie *Star Wars*, obtained from [29] and references therein. The mean rate is 374.4 Kbps and the peak rate is 4.446 Mbps; the peak rate to mean rate ratio being nearly 12. The packet size is fixed at 200 Bytes. We set the bottleneck link at 10 Mbps and the TCP access links at 1 Mbps, while the UDP access link is kept at 100 Mbps. The bottleneck link was only 10 Mbps because the mean rate of the video trace (UDP) is low (374.4 Kbps), and we want to keep the fraction of UDP traffic feeding into the core to within 3-10% of the bottleneck link rate (to be consistent with the nature of Internet traffic today). In this example, the video traffic constitutes $\approx 3.75\%$ of the bottleneck link rate. Subsequent sections of the paper will present results considering higher bottleneck link rates as well.

We have a high-speed access link for UDP since UDP traffic feeding into the core can be an aggregate of many individual UDP streams. TCP traffic on the 1 Mbps access link models

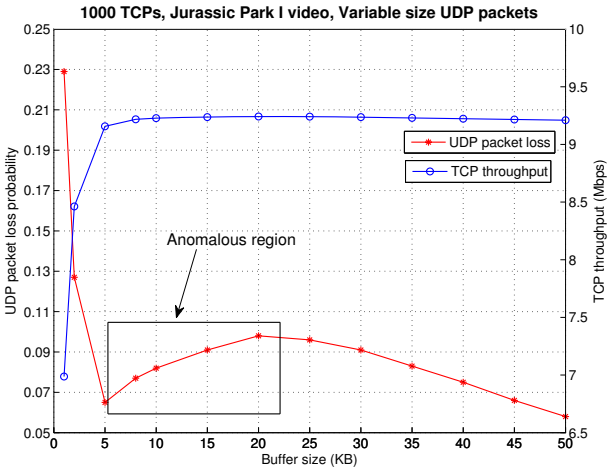


Fig. 3. Jurassic Park [150,300] Byte packets: UDP loss and TCP throughput

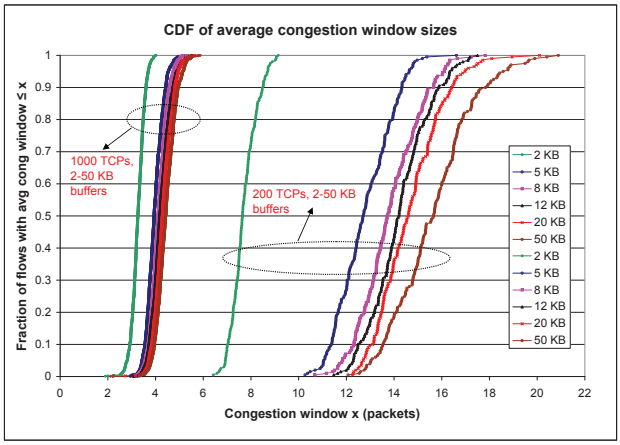


Fig. 4. CDF of average congestion window size for 1000 and 200 TCP flows

traffic from a typical home user. Fig. 2 shows the UDP packet loss and TCP throughput curves as a function of buffer size at the bottleneck router in the range of 1 KB to 50 KB. We see that TCP quickly ramps up to nearly 9.6 Mbps with only about 8 KB of buffering, reaching close to its saturation throughput. Simultaneously, UDP packet loss falls rapidly as well. Up to this point, both TCP and UDP behave as expected. However, as the buffer size increases beyond 8 KB till about 24 KB, UDP performance degrades as its packet loss increases with buffer size in this region. The loss at 24 KB of buffering is approximately 30% more than the loss at 8 KB of buffering. There is however no appreciable increase in TCP throughput.

To verify that the anomalous UDP loss was not unique to the packet trace we had chosen, we performed our simulations using several other traces. For instance, we used video traces from the movie *Jurassic Park I* obtained from [30], with packet sizes uniformly distributed in the range $[150, 300]$ Bytes. The video traffic contributes 7.7% of the bottleneck link rate. Fig. 3 shows the corresponding UDP packet loss curves as a function of buffer size corresponding to this scenario, and clearly indicates the presence of a region of buffer size in which loss increases with buffer size. We also observed this behaviour when UDP traffic is generated from synthetic models (discussed in Section V), suggesting that the anomaly is fundamental and not just a coincidence. Moreover, in the absence of TCP traffic, UDP loss was observed to drop monotonically with buffer size, confirming that the anomaly arises due to the interaction of open and closed loop traffic at the bottleneck router with very small buffers.

Through our results in this study, we hope to bring the anomaly to the attention of optical switch vendors and service providers who could make considerable investment in incorporating optical buffering in their packet switches, only to obtain potentially worse performance if they inadvertently operate their buffer sizes in this anomalous region. Given that each extra KiloByte of optical buffering can add significantly to the cost of the optical switch (this was confirmed in our interactions with Bell Labs researchers who are prototyping an all-optical packet switch called IRIS [31]), manufacturers and operators should be wary of the potential for negative returns on this investment.

Having observed the anomalous loss behaviour in several simulation scenarios, we now develop quantitative models that can help explain this phenomenon.

III. An Intuitive Model of the Anomaly

We begin with an intuitive explanation of why we think the anomaly happens, and this helps us develop a simplistic yet effective model that quantifies the buffer sharing dynamics between TCP and real-time traffic.

To explain the anomaly we start by characterising the dynamics of TCP congestion control as a function of bottleneck link buffer size. We show in Fig. 4 the CDF of the average window size of the TCP flows (obtained from simulation) for various buffer sizes at the bottleneck link, under two different flow settings (both of which carry 5% Poisson UDP traffic): one in which 1000 flows share a 200 Mbps link, and the other in which 200 flows share the 200 Mbps bottleneck link (the TCP per-flow rates are 200 Kbps and 1 Mbps respectively for the two scenarios, representative of different provisioning scenarios, and we note that the anomaly is seen in both the settings). We observe from the figure that when buffers at the bottleneck link are extremely small (say in the range 2 to 5 KB) the congestion window size of the TCP flows are also small – for example, we see that when bottleneck buffers are 2 KB, average congestion window size is 1.8-4 packets when there are 1000 flows, and 6-9 packets when there are 200 flows. As the bottleneck buffer grows larger, say to 10-25 KB, TCP average congestion window sizes also increase – for example, the figure shows that when bottleneck buffers are say 20 KB, average congestion window sizes are in the range 2.1-5.6 packets for the first scenario with 1000 TCP flows, and in the range 12-20 packets for the second scenario with 200 TCP flows. This represents a factor 1.2 to 2 increase in average flow congestion window size as buffers increase from 2 to 20 KB.

The significance of TCP's average congestion window size on the anomaly is as follows: when TCP congestion window is very small (due to very small bottleneck buffers of say 1-5 KB), each TCP flow transmits only a few packets in each round-trip time, and is therefore mostly idle. Consequently, the buffers at the bottleneck link are often devoid of TCP packets, allowing UDP packets to enjoy use of these buffers

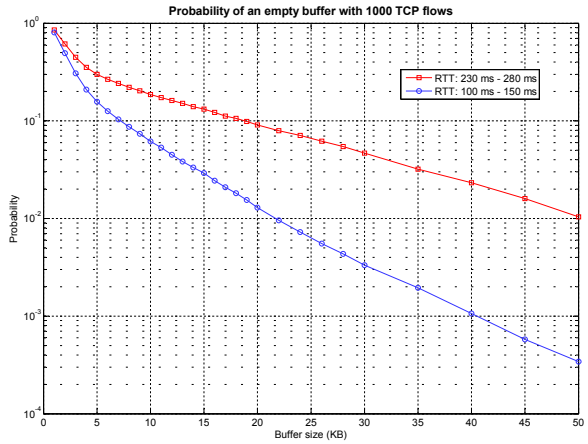


Fig. 5. Probability of idle buffer vs. buffer size for TCP traffic

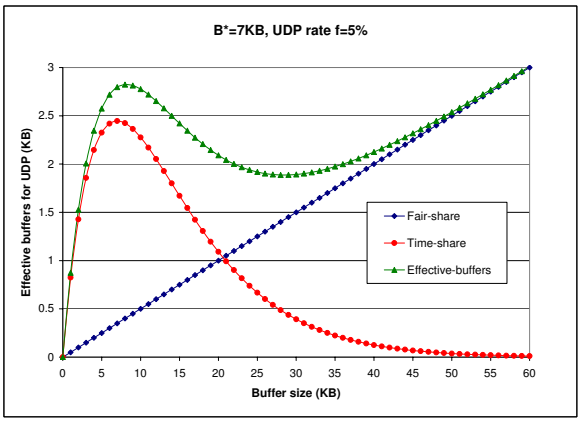


Fig. 6. Effective buffers for UDP traffic

for the most part. In this region, therefore, TCP and UDP predominantly "time-share" the buffers, and UDP loss decreases with buffer size, much like it would if TCP traffic were non-existent. On the other hand, when a larger fraction of the TCP flows are able to increase their congestion window (equivalently a smaller fraction of the TCP flows remain idle), due to bottleneck buffers being larger (say in the range 10-25 KB corresponding to the anomaly), TCP traffic makes more use of the buffers at the bottleneck link, leaving a smaller fraction of the buffers for UDP traffic to use. The aggressive nature of TCP in increasing its congestion window to probe for additional bandwidth, causes the "space-sharing" of bottleneck-link buffers between TCP and UDP in this region to be skewed in favour of TCP, leaving lesser buffers available to UDP traffic even as buffer size increases.

We now try to quantify the above intuition via a simple analytical model that captures the transition from *time-sharing* to *space-sharing* of the bottleneck-link buffers between TCP and real-time traffic. We make the assumption that there is a sufficiently large number of TCP flows sharing the bottleneck link, and that they have sufficiently large round-trip times such that the delay-bandwidth product is larger than the buffering available at the bottleneck link. Moreover, TCP is assumed to contribute a vast majority of the overall traffic on the link (this is consistent with observations that nearly 85-90% of today's Internet traffic is carried by TCP). Under such circumstances, we first make the following observation:

Observation: TCP's usage of the bottleneck buffers increases exponentially with the size of the buffer. More formally, let B denote the buffer size (in KB) at the bottleneck link, and $P_I(B)$ the probability that at an arbitrary instant of time the buffers at the bottleneck link are devoid of TCP traffic. Then,

$$P_I(B) \approx e^{-B/B^*} \quad (1)$$

where B^* is a constant (with same unit as B) dependent on system parameters such as link capacity, number of TCP flows, round-trip times, ratio of long-lived to short-lived TCP flows, etc. The constant B^* can be inferred from the plot of the natural logarithm of $P_I(B)$ as a function of B, which yields a straight line. The slope of the line corresponds to $-1/B^*$.

This behaviour has been observed in the past by various researchers: by direct measurement of idle buffer probabilities [32, Sec. III], as well as indirectly via measurement of TCP throughput [6, Fig. 1]: the latter has shown roughly exponential rise in TCP throughput with bottleneck buffer size, confirming that TCP's loss in throughput (which arises from an idle buffer) falls exponentially with buffer size. We also validated this via extensive simulations (shown in Fig. 5 and in various other TCP plots in later sections) in *ns-2*. 1000 TCP flows with random round-trip times from a chosen range were multiplexed at a bottleneck link, and the idle buffer probability was measured as a function of bottleneck link buffer size. The large number of flows, coupled with randomness in their round-trip times, ensures that the TCP flows do not synchronise their congestion windows. Fig. 5 plots on log-scale the idle buffer probability as a function of bottleneck buffer size for two ranges of round-trip times, and show fairly linear behaviour in the range of 5 to 50 packets (each packet was 1 KiloByte), confirming the exponential fall as per Eq. (1).

Having quantified TCP's usage of the bottleneck buffers, we now consider a small fraction f (say 5 to 10%) of real-time (UDP) traffic multiplexed with TCP traffic at the bottleneck link. The small volume of UDP traffic does not alter TCP behaviour significantly; however, TCP's usage of the buffer does significantly impact loss for UDP traffic. If we assume the buffer is very small (a few tens of KiloBytes), we can approximate the buffer as being in one of two states: idle (empty) or busy (full). With the objective of estimating the "effective" buffers space available to UDP traffic, we identify the following two components:

- **Fair-share:** During periods of time when TCP and UDP packets co-exist in the buffer, the buffer capacity B is shared by them in proportion to their respective rates. The first-in-first-out nature of service implies that the average time spent by a packet in the system is independent of whether the packet is UDP or TCP, and Little's law can be invoked to infer that the average number of waiting packets of a class is proportional to the arrival rate of that class. UDP packets therefore have on average access to a "fair share" of the buffers, namely fB, where f denotes the fraction of total traffic that is UDP.
- **Time-share:** Whenever the buffer is devoid of TCP traffic (i.e. with probability $P_I(B)$), UDP packets have access to the remaining buffer space $(1-f)B$ as well. We call

this the "time share" portion, since this portion of the buffer is shared in time between UDP and TCP traffic. The time-share portion of buffers available to UDP is therefore $P_I(B)(1-f)B$.

Combining the fair-share and time-share components, and invoking Eq. (1) gives us an estimate of the total "effective" buffers $\bar{B}^{udp}$ available to UDP traffic:

$$\bar{B}^{udp} = fB + (1-f)Be^{-B/B^*} \quad (2)$$

To illustrate the significance of this equation we plot it for $f = 0.05$ (i.e. 5% UDP traffic) and $B^* = 7$ KB (consistent from Fig. 5). Fig. 6 shows the total effective buffers for UDP, as well as the fair-share and time-share components. The fair-share component fB increases linearly with buffer size, while the time-share component $(1-f)Be^{-B/B^*}$ rises to a peak and then falls again (readers may notice a shape similar to the Aloha protocol's throughput curve): this happens because smaller buffers are more available for UDP to time-share, but as buffers get larger TCP permits exponentially diminishing opportunity for time-sharing. The total effective buffers for UDP, being the sum of the above two components, can therefore show anomalous behaviour, i.e., a region where larger real buffers can yield smaller effective buffers for UDP. For any realistic UDP traffic model (note that our analytical model does not make any specific assumption about the UDP traffic model), the smaller effective buffers will result in higher loss, which is of serious concern to network designers and operators who operate their router buffer sizes in this region.

The model presented above is highly simplified and ignores several aspects of TCP dynamics as well as real-time traffic characteristics. It nevertheless provides valuable insight into the anomaly, and will be used in later sections for a quantitative understanding of the impact of various parameters on the severity of the anomaly.

IV. A Markov model of the anomaly

It is in general challenging to mathematically analyse finite buffer systems in which several thousand feedback-based adaptive TCP flows interact with stochastic real-time traffic. In what follows we develop a realistic yet rigorous Markov model based on some simplifications:

Assumption: *TCP packet arrivals are Poisson.* If a large number (potentially thousands) of long-lived TCP flows multiplex at a bottleneck link, it is believed [2] they do not synchronise their window dynamics behaviour, and can thus be treated as independent flows. Combined with the fact that each TCP flow's window will be quite small (since bottleneck buffers are small), implying that each flow will only generate a small amount of traffic per RTT, the aggregation of a large number of such independent flows can reasonably be assumed to yield Poisson traffic. Prior studies on buffer sizing have also employed this assumption [11]. We further validate this assumption using *ns-2* simulations on the dumbbell topology shown in Fig. 1.

2000 TCP flows with random RTTs and start-times are multiplexed at the bottleneck link, which operates at 200 Mbps. TCP packet size is fixed at 1 KB. Two sets of buffer sizes are chosen at the bottleneck router r_0 - 50 KB representing the very small buffer case, and 6250 KB $(250ms \times 200Mbps)$ representing the traditional delay-bandwidth product rule. The simulation is run for 151 seconds and data in the range 15-150 seconds is used in all the computations. Our objective is to test if TCP packet arrivals to the bottleneck link buffer are near-Poisson. To do so we measure the burstiness of the arrival traffic at various time-scales. Burstiness at time-sale s is quantified by $\beta(s)$, the coefficient of variation (i.e. ratio of standard deviation to mean) of traffic volume measured over time intervals of size s. Log-log plots of $\beta(s)$ versus s are routinely used in the literature to depict traffic burstiness over various time-scales as an indicator of self-similarity of traffic traces, and to show the influence of the Hurst parameter H. We will use estimates of H obtained from the burstiness plots as a measure of how close the traffic is to being short-range dependent (i.e. Poisson).

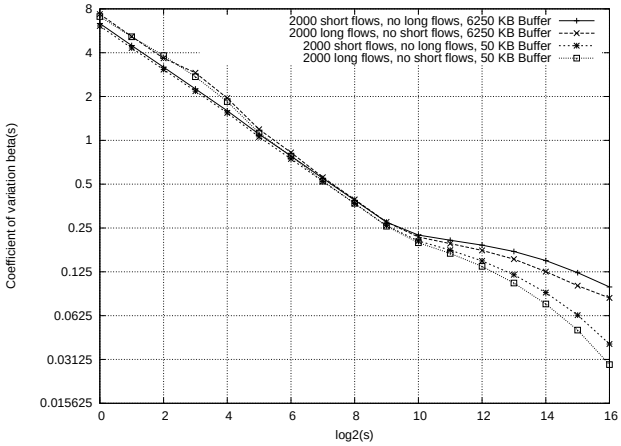


Fig. 7. TCP traffic burstiness at different time-scales for buffer sizes of 6250 KB and 50 KB

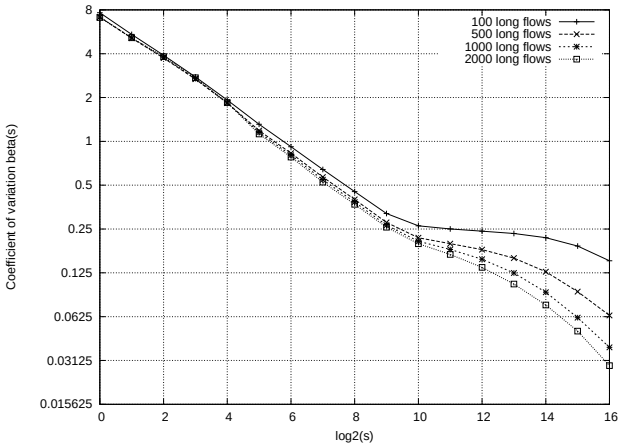


Fig. 8. TCP traffic burstiness at different time-scales with varying number of flows and 50 KB buffer size

Fig. 7 shows the traffic burstiness $\beta(s)$ as a function of the time-scale (s) (in μsec) for four different combinations of TCP flows and buffer size. The top two curves correspond to large (delay-bandwidth) buffer size of 6250 KB, while the bottom two are for small buffers of 50 KB (the two curves in each set correspond respectively to short and long lived flows, with the former generated using a model described in Section VI). Both sets of curves are fairly similar till time-scale of $2^{10}\mu$s $\approx$ 1 ms. However, at time-scales beyond a millisecond, the

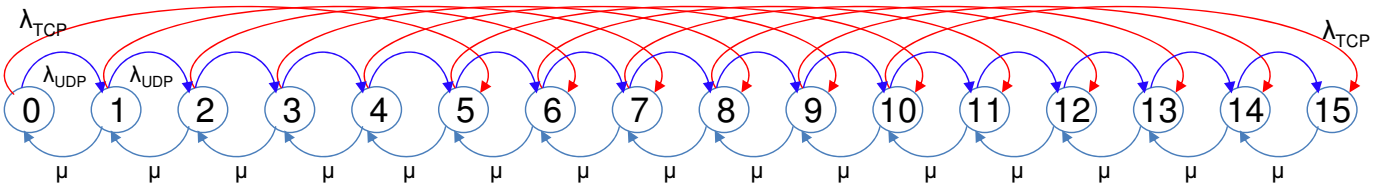


Fig. 9. Markov chain state transition diagram for buffer occupancy with buffer size = 3000 Bytes

curves for large buffers flatten significantly, indicating onset of long range dependence with a Hurst parameter estimated at approximately 0.8. This shows that when buffers are large, TCP traffic is significantly bursty at time-scales commensurate with the time it takes to empty/fill the buffer, and the buffer dynamics cannot be modeled with a Poisson assumption (due to the exhibited long range dependence).

By contrast, the bottom two curves (corresponding to small buffers) in the plot do not change slope significantly at time-scales beyond 1 ms, and retain a slope close to that of short range dependent traffic. This lends credence to our assumption that small bottleneck buffers help reduce synchronisation among flows and make the traffic akin to Poisson.

To further affirm the Poisson nature of aggregated TCP traffic in the core, we investigated the impact of increasing the number of TCP flows. Fig. 8 plots the TCP burstiness for 100, 500, 1000 and 2000 long-lived flows for fixed buffer size of 50 KB at the bottleneck router in the simulation topology of Fig. 1. It can be clearly noted that at timescales beyond a millisecond, the burstiness curve falls more steeply (and hence the traffic exhibits reducing long range dependence) as the number of TCP flows increases. This shows that the TCP arrival process becomes more Poisson as the number of flows increases (due to scalability and memory issues we were unable to scale our simulations to a larger number of flows).

In summary, if bottleneck buffers are small, and if the number of TCP flows multiplexed is large, the assumption that aggregate TCP traffic arrivals to the bottleneck link are Poisson is well justified.

Assumption: *UDP packet arrivals are also Poisson.* Stochastic studies such as [33], [34], [35] have shown that the aggregation of traffic from a large number of independent flows (as can be expected at a core link) converges to Poisson. This important result makes the analysis tractable (though the phenomenon of anomalous loss is observed even if UDP arrivals are non-Poisson).

Claim: *UDP packets are on average smaller in size than TCP packets*, as discussed in Section II, and as reported in several measurements of traffic in the Internet core [28]. Consistent with our example presented in Fig. 2, we choose average TCP and UDP packet sizes to be 1000 and 200 Bytes.

Claim: *The aggregate TCP rate increases exponentially with bottleneck link buffer size*, as demonstrated in Fig. 5 and discussed in the previous section. Denoting the bottleneck buffer size as B (in KiloBytes), the TCP throughput λ_{TCP} is given by:

$$\lambda_{TCP} \approx \{1 - (e^{-B/B^*})\} \times \lambda_{TCP}^{sat} \tag{3}$$

where λ_{TCP}^{sat} denotes the saturation throughput of TCP (when buffer size is large).

In order to construct a Markov chain model we make the further assumption that packet transmission times are exponentially distributed (we will relax this assumption in the next subsection). We can then model the FIFO queue at the bottleneck link router as an M/M/1 system with finite buffer B and with two classes of customers:

1) UDP arrivals are Poisson at fixed rate (denoted by λ_{UDP}), and require exponential service time with unit mean (the service rate is normalised to average UDP packet size), and
2) TCP arrivals are Poisson at rate λ_{TCP} derived from Eq. (3), where each TCP packet arrival brings a bulk of 5 customers (corresponding to the packet size ratio $1000/200$), each requiring exponential service time with unit average.

For illustrative purposes, let us consider the buffer size B to be 3 KiloBytes. Then, we can model the state of the system as the number of customers in the FIFO queue. Fig. 9 shows the resulting Markov chain. A transition from state j to state $j+5$ corresponds to the arrival of a TCP packet, whereas a transition from state j to state $j+1$ corresponds to the arrival of a UDP packet.

Denoting $B_{bytes} = B \times 1000 = 3000$ to be the corresponding buffer size in Bytes, and N the number of states in the Markov chain, then

$$N = \frac{B_{bytes}}{UDP\,packet\,size} + 1 = \frac{3000}{200} + 1 = 16. \tag{4}$$

If p_j represents the steady state probability of the queue being in state j (i.e., the probability that the queue contains j customers), then we can write the global balance equations as follows:

$$p_0\,(\lambda_{UDP} + \lambda_{TCP}) = p_1\,\mu \tag{5}$$

$$p_i\,(\lambda_{UDP} + \lambda_{TCP} + \mu) = p_{i-1}\,\lambda_{UDP} + p_{i+1}\,\mu \qquad (1 \le i \le 4) \tag{6}$$

$$p_i\,(\lambda_{UDP} + \lambda_{TCP} + \mu) = p_{i-1}\,\lambda_{UDP} + p_{i+1}\,\mu + p_{i-5}\,\lambda_{TCP} \qquad (5 \le i \le 10) \tag{7}$$

$$p_i\,(\lambda_{UDP} + \mu) = p_{i-1}\,\lambda_{UDP} + p_{i+1}\,\mu + p_{i-5}\,\lambda_{TCP} \qquad (11 \le i \le 14) \tag{8}$$

$$p_{15}\,\mu = p_{14}\,\lambda_{UDP} + p_{10}\,\lambda_{TCP} \tag{9}$$

The above equations and the normalising constraint $\sum_{i=0}^{15} p_i = 1$ form a set of linear equations that can be solved to compute the probability that an incoming UDP packet will be dropped, which in this example is p_{15}. Obtaining balance equations as the buffer size B increases is straightforward, and the resulting set of linear equations is easily solvable numerically (in MATLAB) to get the UDP packet loss probability.

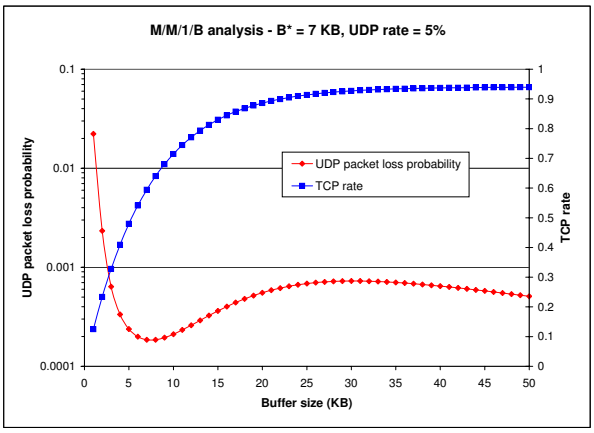


Fig. 10. Anomalous UDP loss results from the M/M/1/B analytical model

The analytical result shown in this paper chooses model parameters to match the simulation setting as closely as possible: the normalised UDP rate is set to $\lambda_{UDP} = 0.05$ (i.e. 5% of link capacity), and the TCP saturation throughput $\lambda_{TCP}^{sat} = 0.94/5$ (so that TCP and UDP customers have a combined maximum rate less than the service rate of $\mu = 1$ in order to guarantee stability). The constant $B^* = 7$ KB is consistent with what is obtained from simulations in Fig. 5.

Fig. 10 plots the UDP loss (on log scale) obtained from solving the M/M/1 chain with bulk arrivals and finite buffers, as well as the TCP rate in Eq. (3), as a function of buffer size B. It can be observed that in the region of 1-8 KB of buffering, UDP loss falls monotonically with buffer size. However, in the buffer size region between 9-30 KB, UDP packet loss increases with increasing buffer size, showing that the model is able to predict the anomaly found in simulations.

A. M/D/1/B Analysis

We refine the M/M/1/B model by relaxing the assumption that packet sizes are exponentially distributed. It has been recently observed that Internet packet sizes have a bi-modal distribution [37], [38]: with peaks at large packets (1500 Byte TCP data) and small packets (typically 40 Byte TCP ACK). Real-time and other streams generate intermediate packet sizes (200-500 Bytes). To develop a model that is tractable yet reflective of these dominant modes, we employ an M/D/1/B model in which packet sizes are bi-modal: large for TCP packets (1000 Bytes) and small for UDP packets (say 200 Bytes). For brevity, we do not explain the derivation of this model here and instead refer the interested reader to our paper [39] for the detailed analysis. It suffices to state here that the results from the M/D/1/B analysis are qualitatively similar to the M/M/1/B results presented above, and both chains predict the inflection point to occur at around 8 KB. Moreover, the M/D/1/B model validates the anomaly from the realistic scenario of having bi-modal packet sizes and not relying on exponential service times.

V. Impact of Real-Time Traffic Characteristics

Using simulations (all performed on the dumbbell topology shown in Fig. 1) as well as the analytical models developed above, we now investigate the impact of various system parameters on the nature of the observed anomalous loss performance. This section studies how real-time traffic characteristics such as traffic model and intensity affect the anomaly, while subsequent sections consider the effect of TCP traffic characteristics and packet size distributions.

A. Traffic Model

It may be recalled that our intuitive model in Section III did not make any specific assumptions about the stochastics of UDP traffic. We now validate via simulation that the anomaly is agnostic to whether the UDP traffic exhibits short or long range dependence. The two models are described next.

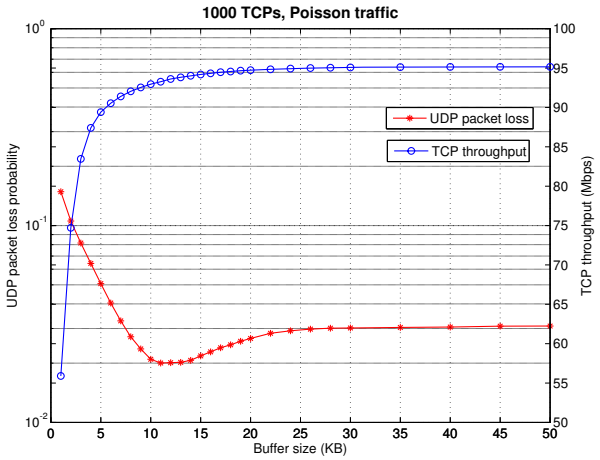


Fig. 11. Poisson: UDP packet loss and TCP throughput from simulation

1) **Poisson:** We start with the well-known Poisson model as the UDP traffic source. The core link (we use the term core link to refer to the bottleneck link and vice-versa) bandwidth is set at 100 Mbps. TCP access links are at 10 Mbps each, while the UDP access link operates at 100 Mbps. The average rate of Poisson traffic is 5 Mbps, constituting about 5% of the total bottleneck link bandwidth. Fig. 11 shows the UDP packet loss (on log-scale) and the corresponding TCP throughput curves when the buffer size at the bottleneck router is varied from 1 KB to 50 KB. TCP is able to quickly ramp up to nearly 93 Mbps with just about 11 KB of buffering, corresponding to nearly 98% of its saturation throughput. We note from the figure that up to 11 KB, UDP packet loss falls with increasing buffer size. In addition, further increase in buffer size leads to an increase in UDP packet loss. The loss at 30 KB of buffering is 50% more than the loss at 11 KB of buffering, while there is only a negligible increase in TCP throughput.

2) **fBm:** It is widely believed that Internet traffic is not Poisson in nature but tends to exhibit self-similar and long-range dependent properties. To see if the phenomenon also occurs under this scenario, we generated fBm traffic at the same average rate of 5 Mbps. Other parameters are the same as before. The fBm model used is similar to our previous work in [40], [41]. The traffic model combines a constant mean arrival rate with fractional Gaussian noise (fGn) characterised by zero mean, variance σ^2 and Hurst parameter $H \in [1/2, 1)$. We use our filtering method in [42] to generate, for a chosen H, a sequence x_i of normalised fGn (zero mean and unit variance). A discretisation interval Δt is chosen, and each x_i then denotes the amount of traffic, in addition to the constant rate stream that arrives in the i-th interval. Specifically, the

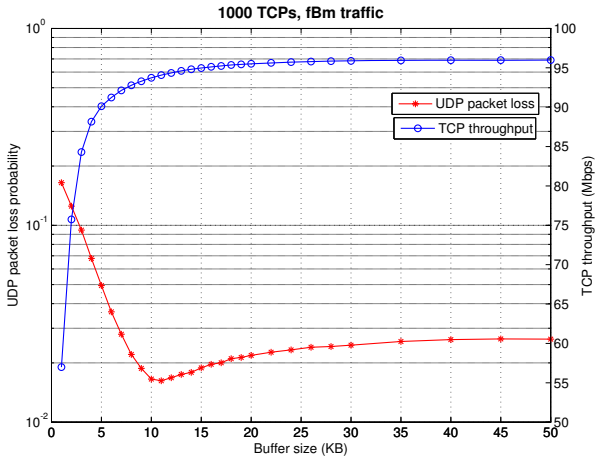


Fig. 12. fBm: UDP packet loss and TCP throughput from simulation

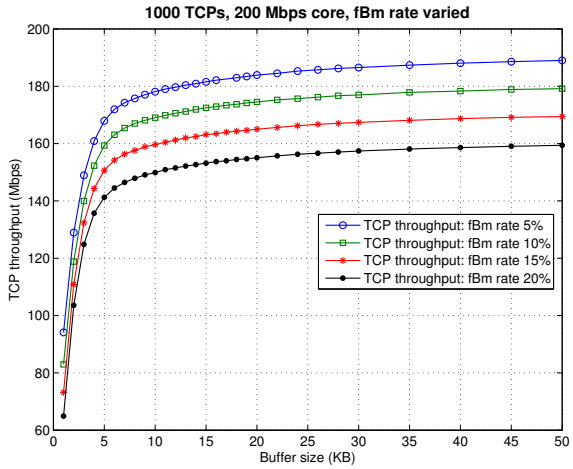


Fig. 14. TCP throughput for varying fBm rates

traffic y_i (in bits) arriving in the i-th interval of length Δt seconds is computed using $y_i = max\{0, \rho_c \Delta t + s x_i\}$ where ρ_c denotes in bits-per-second the constant rate stream, and s is a scaling factor that determines the instantaneous burstiness. For this work we set the Hurst parameter at $H = 0.85$ and the discretisation interval $\Delta t = 1.0s$. The scaling factor s is chosen to satisfy $\rho_c \Delta t / s = 1.0$, which corresponds to moderate burstiness (around 16% of the samples are truncated), and ρ_c is then adjusted to give the desired mean traffic rate. The fluid traffic is then packetised into fixed-length packets (of size 200 Bytes) before being fed into the simulations.

We plot the UDP packet loss (on log-scale) and the TCP throughput curves as a function of buffer size in Fig. 12. Here too, as in the case of the Poisson traffic model, TCP attains 98% of its saturation throughput with only about 11 KB of buffering. UDP packet loss is the lowest at this point. An increase in buffer size negatively affects UDP packet loss, but results in only a marginal improvement in TCP throughput. The loss at 30 KB of buffering is nearly 50% more than the loss at 11 KB of buffering.

B. Traffic Intensity

Having observed the anomalous loss phenomenon for both short and long range dependent UDP traffic models, we now explore if the anomaly is affected by the relative volumes of real-time and TCP traffic. This helps us understand if ISPs who carry a relatively larger fraction of real-time traffic are more susceptible to this anomalous loss performance, and also if the anomaly would change if real-time applications were to become more widespread thereby changing the traffic mix in the Internet. With this aim, we simulated 1000 TCP flows on a 200 Mbps core link with (long range dependent) UDP traffic rate set at 5%, 10%, 15% and 20% of the core link capacity. The resulting plots of UDP loss and TCP throughput are shown in Figures 13 and 14. We observe from Fig. 13 that when the UDP rate is 5%, the inflection point is clearly seen to exist at about 9 KB. However, as the fraction of UDP traffic increases, the inflection point gradually shifts to the left, while the magnitude of the anomaly seems to diminish.

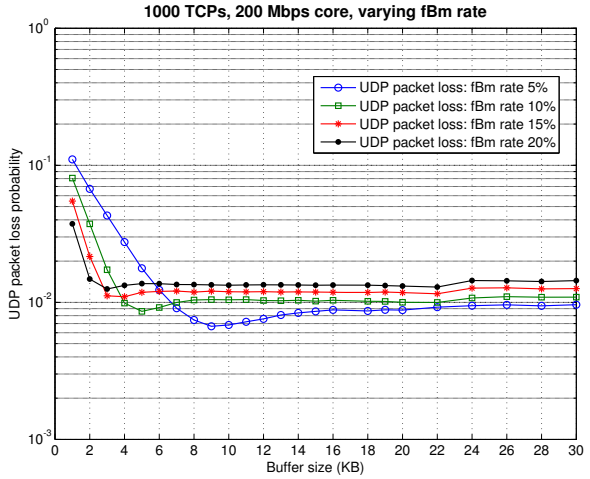


Fig. 13. UDP loss with varying fBm rate

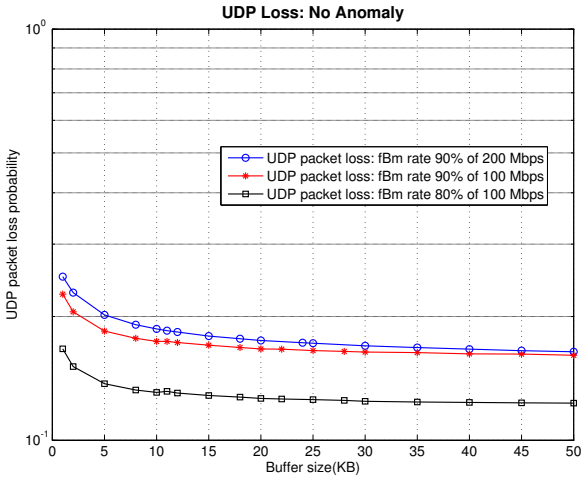
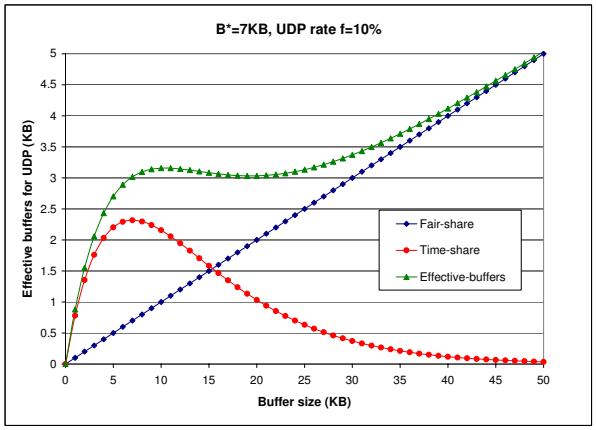
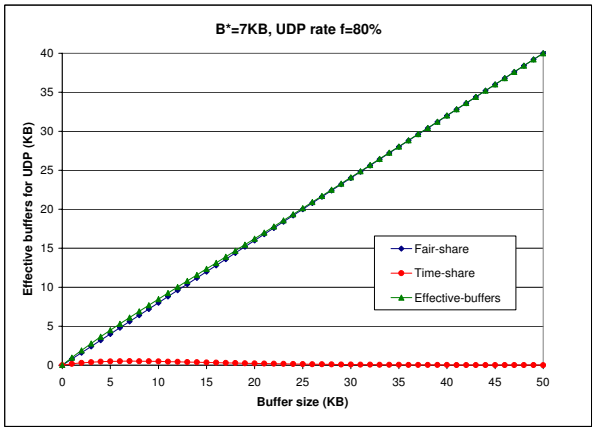


Fig. 15. UDP loss falling monotonically with buffer size by simulation

To see if the anomaly vanishes entirely at high UDP rates, we simulated three scenarios, corresponding to 80 and 90 Mbps average UDP rates on a 100 Mbps core link, and 180 Mbps average UDP rate on a 200 Mbps core link, each with 1000 TCP flows. The fraction of UDP traffic is thus very high at 80-90%. The resulting UDP loss curves are plotted in Fig. 15. Clearly, we can see that the UDP loss curves do not exhibit a point of inflection, i.e. UDP loss falls monotonically with increasing buffer size and there is no anomalous loss.



(a) $\bar{B}^{udp}$ at 10% rate



(b) $\bar{B}^{udp}$ at 80% rate

Fig. 16. Effective buffers for UDP (from analysis) when the fraction of UDP traffic is 10% (low) and 80% (high)

A qualitative explanation for why the anomaly vanishes at high UDP rates follows. Referring back to the case when UDP rates are low, increasing buffers in the anomalous region gave TCP an exponentially larger opportunity to use the overall buffers, while giving UDP only a minimal fair-share of extra buffering; the net effect being a reduction in the effective buffers available to UDP. Now, when UDP rates are high, increasing the buffers at the bottleneck link gives UDP substantially more buffers as its fair-share (in proportion to its rate), while diminishing the opportunity for TCP to time-share the buffers with UDP. This results in a net positive gain in the effective buffers available to UDP. As a result, the UDP packet loss falls monotonically with increasing buffer size.

The above intuition can be quantified using our buffer sharing model developed in Section III. Referring to Eq. (2), recall that f represents the fraction of UDP traffic, and the two terms in the summation represent respectively the fair-share and time-share components on the effective-buffers available to UDP. We plot in Figures 16(a) and 16(b) the fair-share component, the time-share component, and the effective-buffers when $B^* = 7$ KB, for two sets of UDP rates, namely, 10% (low) and 80% (high). From the figures, and also from Fig. 6 that plots these values for $f = 0.05$ (5% UDP traffic), we note that the shape of the curves corresponding to the time-share component and the effective-buffers available to UDP changes as the UDP rate increases. The presence of the time-share component is less pronounced, while the effective buffers approaches a straight line at higher rates. To explain the change in the nature of these curves we note that from Eq. (2), as f increases, the fair-share component fB begins to dominate over the time-share component, since $(1-f)Be^{-B/B^*}$ becomes negligible (tends towards 0) at large f. This implies that the effect of the time-share component on the effective buffers available to UDP falls with increasing UDP rate (as seen in the figures). Consequently, $\bar{B}^{udp}$ increases linearly with buffer size B, which implies that the effective buffers available to UDP increases as the real buffer size increases, thus yielding a straight line with slope f. This helps explain why at high UDP rates, simulations show that the packet loss falls monotonically with buffer size.

VI. Impact of TCP Characteristics

Our Markov chain analysis relied on the assumption that TCP's usage of buffers increases exponentially with buffer size. Though this has been observed when all TCP flows are long-lived ([32, Sec. III], [6, Fig. 1], Fig. 5 above), the reader may wonder if similar behaviour is seen when many of the TCP flows are short-lived (or equivalently, the number of TCP flows is time-varying). This is an important consideration since measurement based studies at the core of the Internet suggest that a large number of TCP flows (e.g. HTTP requests) are short-lived ("mice") and carry only a small volume of traffic, while a small number of TCP flows (e.g. FTP) are long-lived ("elephants") and carry a large volume of traffic.

We studied UDP loss for such TCP traffic mixes by simulating them in *ns-2* over the dumbbell topology shown in Fig. 1. In Fig. 17, we plot the TCP empty buffer probability on a 200 Mbps core link for four different ratios of long-lived to short-lived flows. The total number of TCP flows is kept constant at 2000. In order to incorporate realistic TCP traffic, we consider the closed-loop flow arrival model described in [43] and [10], operating as follows. A given number of users (up to a maximum of 2000 in our example) perform successive file transfers to their respective destination nodes. The size of the file to be transferred follows a Pareto distribution with mean 100 KB and shape parameter 1.5. These chosen values are representative of Internet traffic, and comparable with measurement data. After each file transfer, the user transitions into an idle or off state, or as the authors of [10] suggest, a "thinking period". The duration of the thinking period is exponentially distributed with mean 1 second. It is widely believed that Internet traffic exhibits self-similar and long-range dependent characteristics. It can be noted that the above traffic generation mechanism, which is a combination of several ON-OFF sources with Pareto-distributed ON periods, is in fact long-range dependent [44].

Fig. 17 plots the empty buffer probability for the four different ratios of long-lived to short-lived flows. Our first observation is that the empty buffer probability falls fairly linearly (on log-scale) with buffer size (in the region of 8-50

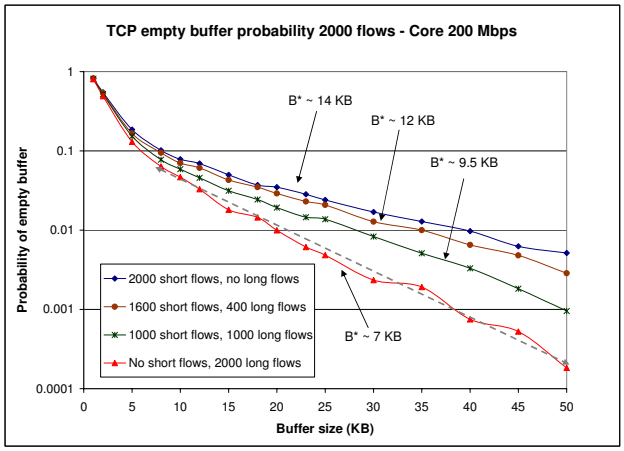


Fig. 17. Empty buffer probability with varying number of short and long lived TCP flows

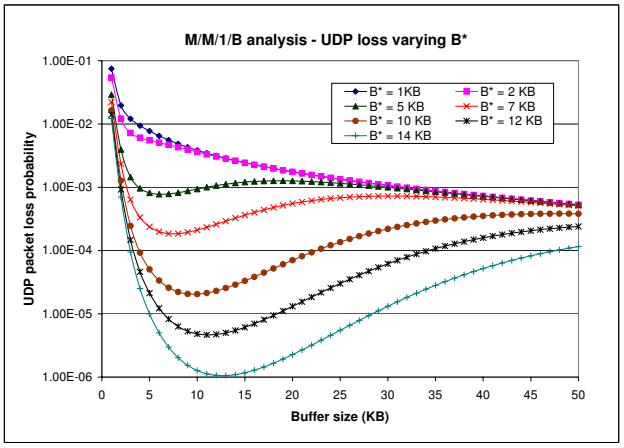


Fig. 19. UDP loss from M/M/1/B analysis

KiloBytes) irrespective of the traffic mix. This satisfies one of the important assumptions required by our analytical models shown in Eq. (1) and Eq. (3), and thus renders our models valid for mixes of short and long TCP flows. However, the slope of the linear region, which in turn determines the B^* required by the model, does seem to depend on the relative fractions of short and long lived flows. The figure shows that as the fraction of long-lived flows increases from 0 to 1, the value of B^* decreases correspondingly from 14 KB down to 7 KB. Intuitively, this is because short-lived flows do not generate sufficient traffic to continuously saturate the link and most of them remain in the slow-start phase without entering into the congestion avoidance mode during the entire file transfer process. However, with the increase in the number of long-lived flows, there is a corresponding increase in the buffer occupancy, since long-lived flows always have data to send and are more likely to be in the congestion avoidance mode. This results in the core link being saturated, and reduces the probability of the buffer being empty, explaining why B^* reduces as the number of long-lived flows increases.

Having observed how B^* changes with the long-short TCP flow mix, we study the corresponding impact on the performance predicted by our analytical model. Fig. 18 shows the

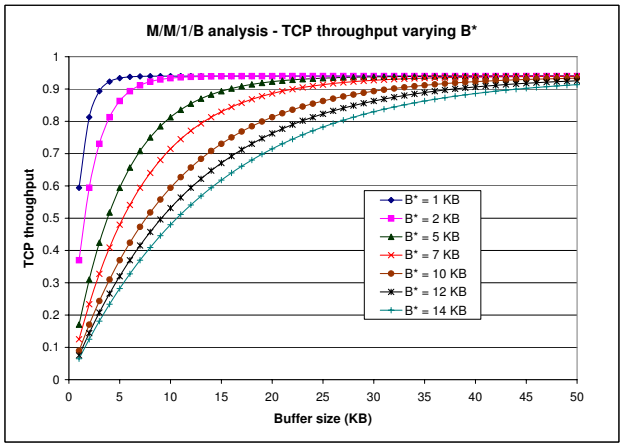


Fig. 18. TCP throughput from M/M/1/B analysis

TCP throughput curves obtained from the M/M/1/B analysis as a function of core link buffer size for different values of B^*, ranging from 1 KB to 14 KB. The key point to note from the figure is that as B^* increases, TCP requires bigger buffers to attain saturation throughput, which is 0.94 or 94% of the core link rate since this analysis plot considers the presence of 0.05 or 5% UDP traffic. In other words, the smaller the B^*, the faster TCP rises, thus needing fewer buffers to attain saturation throughput.

Fig. 19 shows the impact of B^* on the UDP loss prediction from our model. We observe that the UDP packet loss (and hence the severity of the anomaly) is more pronounced when B^* is larger. Moreover, the inflection point, i.e., the point at which UDP packet loss begins to increase, shifts slightly to the right as B^* increases.

To verify if these observations are corroborated in simulation, we multiplex 10 Mbps Poisson traffic with the above four mixes of short and long lived TCP flows, and record the UDP packet loss as a function of buffer size. Fig. 20 shows that as the number of long-lived flows increases, there is a corresponding increase in losses for UDP traffic. Referring back to our analytical model, this can be argued as follows: when the fraction of long-lived TCP flows increases from 0 to 1, the core link buffer occupancy due to TCP traffic alone increases, reducing B^* from 14 KiloBytes to 7 KiloBytes. This in turn permits lesser opportunity for UDP traffic to access the buffers, leading to higher UDP losses. Though the prediction from analysis qualitatively matches simulation, and even predicts the inflection point with reasonable accuracy, we notice a quantitative discrepancy between them, particularly when short TCP flows dominate. This is because our assumption that TCP arrivals are Poisson is not very accurate when we do not have a large number of TCP flows (refer to Fig. 8), and when the volume of short-lived TCP traffic is significant (refer to Fig. 7), since the short-lived TCP flows used in simulation have long range dependent characteristics. We believe that if we have potentially tens of thousands of TCP flows multiplexing at a router with very small buffers, the loss estimates obtained from the model will be more accurate. Nevertheless, our model validates the anomaly even in the presence of short/long-lived TCP mixes, and gives a good indication on how the severity

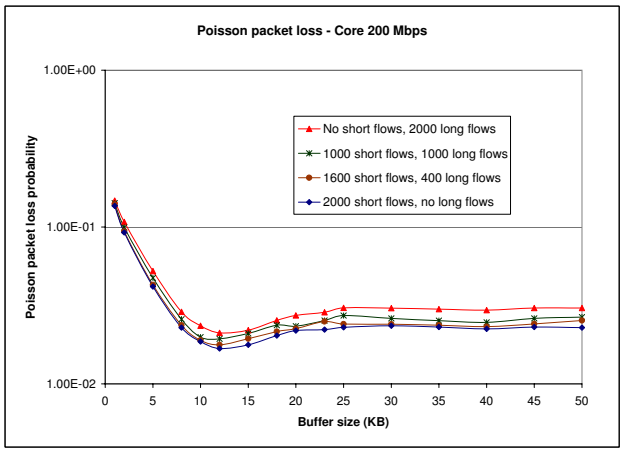


Fig. 20. UDP loss for varying number of short and long TCP flows

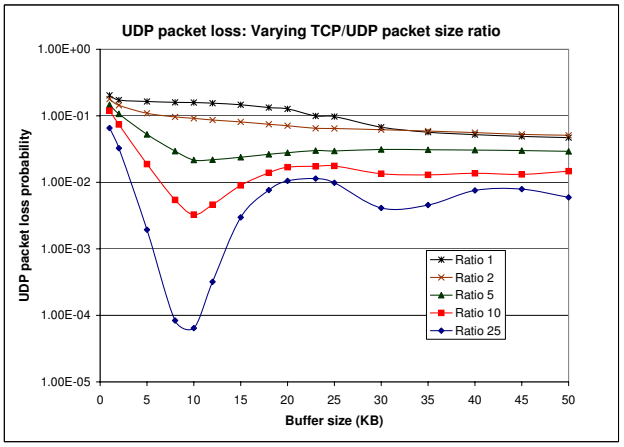


Fig. 21. UDP packet loss from simulation with varying packet size ratios

of the anomaly changes with the TCP traffic mix. Given that TCP traffic is by itself notoriously difficult to analyse, let alone thousands of TCP flows interacting with real-time traffic, we believe our model offers sufficient accuracy to be a valuable aid in sizing routers (potentially all-optical) with very small buffers.

Our study also evaluates the impact of several other TCP parameters such as round-trip times, number of flows, and core link capacity on the severity of the anomaly. With large enough number of flows, these parameters did not seem to have a very dominant effect on the anomaly, and hence we do not discuss them here. The interested reader is referred to our paper [36] for a more detailed discussion of the impact of these parameters.

VII. Impact of Packet Size Distribution

In this section, we investigate the impact of varying TCP/UDP packet size ratios on the anomalous loss performance, and also point out the implications for TCP ACK (acknowledgement) packet losses.

We consider UDP packet sizes of 40, 100, 200, 500 and 1000 Bytes, while fixing TCP packets at 1000 Bytes. TCP flows, 1000 in number, along with 5% (i.e. 5 Mbps) Poisson UDP traffic, are multiplexed on the dumbbell topology with a 100 Mbps core link. Fig. 21 shows the UDP packet loss observed in simulation as a function of core link buffer size for different packet size ratios.

The figure indicates that UDP losses are higher when UDP packets are larger (i.e. ratio of TCP to UDP mean packet size gets smaller). This by itself is not surprising, since for a given mean bit-rate bigger packets constitute burstier arrivals (they are equivalent to a bulk arrival of several smaller packets), and moreover larger packets are dropped in their entirety even if a large part (but not whole) of the packet can be accommodated in the buffer. What is however interesting to note in the simulation plot is that reducing the UDP packet size makes the anomaly more pronounced: when TCP and UDP packets have identical sizes (1000 Bytes), the anomaly is not witnessed in simulation, but when UDP packets are only 40 Bytes long, the anomaly is quite severe.

We compare the simulation results against prediction from our Markov chain model. We use the same $B^* = 7$ KB that was observed in simulation of this scenario (already shown in Fig. 5), and employ in our M/M/1/B chain a bulk arrival size equal to the TCP/UDP packet size ratio. Fig. 22 shows the UDP loss curves obtained via analysis for different packet size ratios. Although the analysis and simulation curves are not a perfect match, it nevertheless corroborates with the results obtained from simulation, and indicates that:

1) First, the analysis predicts correctly that as the TCP/UDP packet size ratio gets larger, losses for UDP traffic become smaller. The bottom curve in Fig. 22 depicts losses for 40 Byte UDP packets (i.e., ratio 25) and clearly shows that loss corresponding to this packet size is significantly lower than loss for larger UDP packets.
2) Second, the analysis seems to be fairly accurate in predicting the inflection point. While the simulations suggest that the inflection point occurs at 10 KB, the analysis predicts it to happen at around 8 KB. Further, that losses (for all the packet size ratios except 25) fall again beyond 30 KB of buffers is predicted successfully by analysis.
3) Third, when TCP and UDP packets are of equal or near-equal size, both simulation and analysis show that the anomaly is insignificant, and the anomaly increases in severity as the disparity in packet sizes increases.

The importance of packet size to the anomalous loss performance also has an implication for TCP ACK packets that are typically 40 Bytes long. We therefore undertook a simulation study of whether TCP ACK packets will also exhibit similar anomalous behaviour. We simulated 1000 bidirectional TCP flows (without UDP) on the dumbbell topology and recorded the ACK packet drops at routers r_0 and r_1 (see Fig. 1). The simulation parameters are identical to the setup described earlier. In Fig. 23 we plot the ACK packet loss probability as a function of core link buffer size. ACK drops in the forward direction correspond to losses at r_0, while the losses in the reverse direction correspond to losses at r_1. Clearly, ACK packets also suffer from the anomaly, and indeed match well with the analytical estimate plotted in Fig. 22.

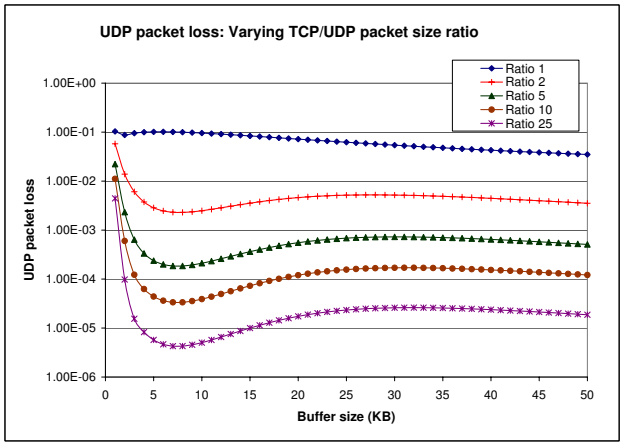


Fig. 22. M/M/1/B UDP packet loss results with varying packet size ratios

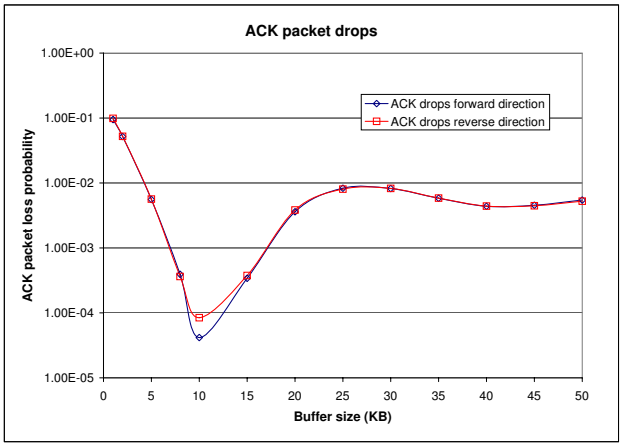


Fig. 23. Losses for TCP ACK packets from simulation

VIII. Conclusions and Future Work

The subject of router buffer sizing has received considerable attention over the past few years. Researchers have questioned the use of the rule-of-thumb, and have argued that few tens of packets of buffering suffice at core Internet routers for TCP traffic to realise acceptable link utilisation. However, the research has been primarily TCP centric, since over 85-90% of today's Internet traffic is carried by TCP. Although real-time (UDP) traffic accounts for only about 5-10%, we note that its popularity, through the prolific use of on-line gaming, real-time video conferencing, VoIP, and many other multimedia applications, is growing in the Internet. As such, we believe that the study of router buffer sizing should not focus on TCP alone, but should consider its impact on the performance of real-time traffic also.

As a first step, in this paper, we examined the dynamics of UDP and TCP interaction at a bottleneck link router equipped with very small buffers. We observed a curious phenomenon - losses for real-time traffic do not fall monotonically with buffer size as one would expect. Instead, there exists an inflection point beyond which loss increases with increasing buffer size. We showed the existence of the anomalous loss behaviour using real video traffic traces, short-range dependent Poisson traffic, and long-range dependent fBm traffic models. Further, we developed simple analytical models that gave insights into why the anomaly exists under certain circumstances. We also presented scenarios describing when the anomaly does not exist. Through extensive simulations, we investigated the impact of various factors such as fraction of UDP traffic, ratio of long-lived to short-lived TCP flows, and packet sizes on the anomaly. The effect of these factors on the inflection point were studied in conjunction with the analytical models.

It is apparent that emerging optical packet switched networks will be capable of buffering no more than a few tens of KiloBytes of data. Given this stringent constraint and the fact that adding extra buffering adds significantly to the cost of the optical switch, the anomalous behaviour revealed in our study can be of serious concern to optical switch manufacturers and service providers who could make considerable investment in these optical packet switches, only to realise worse performance if they inadvertently operate their buffer sizes in this anomalous region.

Several aspects of the problem require further investigation. In our analytical models, we have assumed TCP arrivals to be Poisson. There is scope to refine this assumption, particularly in the presence of short-lived flows that are known to have some long-range dependent characteristics in their activity durations. We can also seek to refine our Markov chain based model to incorporate other UDP traffic models. The impact of the ACK drop anomaly on TCP throughput and average flow completion times warrants a deeper understanding. Finally, we are in the process of developing an experimental testbed on which we can reproduce the anomaly seen in simulations and analysis. Commercial switches and routers today have hidden buffers that do not permit buffer sizes to be configured as low as a few tens of packets. For our experimentation we are therefore using the FPGA-based routing platform called NetFPGA (refer web-site) recently developed at Stanford University that allows fine-grained control of buffer sizes, and has been used successfully in their buffer sizing studies [15]. We hope to be able to report results from such a test-bed in the near future.

References

[1] C. Villamizar and C. Song, "High performance TCP in ANSNet," *ACM SIGCOMM Computer Communications Review*, vol. 24, no. 5, pp. 45-60, 1994.

[2] G. Appenzeller, I. Keslassy and N. McKeown, "Sizing router buffers," *Proc. ACM SIGCOMM*, Oregon, USA, Aug/Sep 2004.

[3] G. Appenzeller, "Sizing router buffers," *PhD Thesis, Stanford University"*, Mar 2005.

[4] D. Wischik and N. McKeown, "Part I: Buffer sizes for core routers," *ACM SIGCOMM Computer Communications Review*, vol. 35, no. 3, pp. 75-78, Jul 2005.

[5] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown and T. Roughgarden, "Part III: Routers with very small buffers," *ACM SIGCOMM Computer Communications Review*, vol. 35, no. 3, pp. 83-90, Jul 2005.

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

[7] Y. Ganjali and N. McKeown, "Update on buffer sizing in Internet routers", *ACM SIGCOMM Computer Communications Review*, vol. 36, no. 5, pp. 67-70, Oct 2006.

[8] N. Hohn, D. Veitch, K. Papagiannaki and C. Diot, "Bridging router performance and queuing theory," *Proc. ACM SIGMETRICS*, New York, USA, Jun 2004.

[9] G. Raina, D. Towsley, and D. Wischik, "Part II: Control theory for buffer sizing," *ACM SIGCOMM Computer Communications Review*, vol. 35, no. 2, pp. 79-82, Jul 2005.
[10] R. S. Prasad, C. Dovrolis and M. Thottan, "Router buffer sizing revisited: The role of the output/input capacity ratio," *Proc. ACM CoNEXT*, New York, USA, Dec 2007.
[11] A. Lakshmikantha, R. Srikant and C. Beck, "Impact of file arrivals and departures on buffer sizing in core routers," *Proc. IEEE INFOCOM*, Phoenix, Arizona, USA, Apr 2008.
[12] A. Dhamdhere and C. Dovrolis, "Open issues in router buffer sizing," *ACM SIGCOMM Computer Communications Review*, vol. 36, no. 1, pp. 87-92, Jan 2006.
[13] G. Vu-Brugier, R. S. Stanojevic, D. J. Leith and R. N. Shorten, "A critique of recently proposed buffer-sizing strategies," *ACM SIGCOMM Computer Communications Review*, vol. 37, no. 1, pp. 43-47, Jan 2007.
[14] M. Wang and Y. Ganjali, "The effects of fairness in buffer sizing," *Proc. IFIP NETWORKING*, Atlanta, USA, May 2007.
[15] N. Beheshti, Y. Ganjali, M. Ghobadi, N. McKeown, and G. Salmon, "Experimental study of router buffer sizing," *Proc. ACM/USENIX Internet Measurement Conference*, Greece, Oct 2008.
[16] A. Vishwanath, V. Sivaraman and M. Thottan, "Perspectives on router buffer sizing: Recent results and open problems." *ACM SIGCOMM Computer Communications Review Editorial Zone* vol. 39, no. 2, pp. 34-39, Apr 2009.
[17] N. Beheshti, Y. Ganjali, R. Rajaduray, D. Blumenthal and N. McKeown, "Buffer sizing in all-optical packet switches," *IEEE/OSA Optical Fiber Communication (OFC) Conference*, California, USA, Mar 2006.
[18] D. Hunter, M. Chia, and I. Andonovic, "Buffering in optical packet switches," *IEEE/OSA Journal of Lightwave Technology*, vol. 16, no. 12, pp. 2081–2094, Dec 1998.
[19] S. Yao, S. Dixit and B. Mukherjee, "Advances in photonic packet switching: An overview," *IEEE Communications Magazine*, vol. 38, no. 2, pp. 84-94, Feb 2000.
[20] H. Park, E. F. Burmeister, S. Bjorlin and J. E. Bowers, "40-Gb/s optical buffer design and simulations," *Proc. Numerical simulation of optoelectronic devices (NUSOD)*, California, USA, Aug 2004.
[21] E. F. Burmeister et al., "SOA gate array recirculating buffer for optical packet switching," *Proc. IEEE/OSA Optical Fiber Communication (OFC) Conference*, California, USA, Feb 2008.
[22] J. D. LeGrange et al., "Demonstration of an integrated buffer for an all-optical packet router," *Proc. IEEE/OSA Optical Fiber Communication (OFC) Conference*, California, USA, Mar 2009.
[23] Y. Zhang and D. Loguinov, "ABS: Adaptive buffer sizing for heterogeneous networks," *Proc. IEEE International Workshop on Quality of Service (IWQoS)*, Enschede, The Netherlands, Jun 2008.
[24] R. Stanojevic, R. Shorten and C. Kellet, "Adaptive tuning of drop-tail buffers for reducing queueing delays," *IEEE Communications Letters*, vol. 10, no. 7, pp. 570-572, Jul 2006.
[25] ns-2 network simulator - http://www.isi.edu/nsnam/ns/
[26] W. Feng, F. Chang, W. Feng and J. Walpole, "Provisioning on-line games: A traffic analysis of a busy counter-strike server," *Proc. ACM SIGCOMM Internet Measurement Workshop*, France, Nov 2002.
[27] W. Feng, F. Chang, W. Feng and J. Walpole,"A traffic characterization of popular on-line games," *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, Jun 2005.
[28] Packet traces from measurement and analysis on the WIDE Internet backbone. http://tracer.csl.sony.co.jp/mawi
[29] V. Markovski, F. Xue and L. Trajkovic, "Simulation and analysis of packet loss in video transfers using user datagram protocol," *The Journal of Supercomputing*, vol. 20, no. 2, pp. 175-196, 2001.
[30] Video traffic traces for performance evaluation: http://trace.eas.asu.edu/TRACE/ltvt.html
[31] P. Bernasconi et al., "Architecture of an integrated router interconnected spectrally (IRIS)", *Proc. IEEE High Performance Switching and Routing*, Poland, Jun 2006.
[32] L. Andrew, T. Cui, J. Sun, M. Zukerman, K. Ho, and S. Chan, "Buffer sizing for nonhomogeneous TCP sources," *IEEE Communications Letters*, vol. 9, no. 6, pp. 567-569, Jun 2005.
[33] G. Raina and D. Wischik, "Buffer sizes for large multiplexers: TCP queueing theory and instability analysis," *Proc. Next Generation Internet Networks*, Rome, Apr 2005.
[34] J. Cao and K. Ramanan, "A Poisson limit for buffer overflow probabilities," *Proc. IEEE INFOCOM*, New York, USA, Jun 2002.
[35] J. Cruise, "Poisson convergence, in large deviations, for the superposition of independent point processes," *Annals of Operations Research*, vol. 170, no. 1, pp. 79-94, Sep 2009.
[36] A. Vishwanath and V. Sivaraman, "Routers with very small buffers: Anomalous loss performance for mixed real-time and TCP traffic," *Proc. IEEE International Workshop on Quality of Service (IWQoS)*, Enschede, The Netherlands, Jun 2008.
[37] R. Sinha, C. Papadopoulos and J. Heidemann, "Internet packet size distributions: Some observations," *Technical report ISI-TR-2007-643, USC/Information Sciences Institute,* May 2007, http://www.isi.edu/~johnh/PAPERS/Sinha07a.html
[38] Caida Internet packet size distribution, http://www.caida.org/research/traffic-analysis/pkt_size_distribution/graphs.xml
[39] A. Vishwanath, V. Sivaraman and G. N. Rouskas, "Considerations for sizing buffers in optical packet switched networks," *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr 2009.
[40] V. Sivaraman, H. ElGindy, D. Moreland and D. Ostry, "Packet pacing in short buffer optical packet switched networks," *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr 2006.
[41] V. Sivaraman, H. ElGindy, D. Moreland and D. Ostry, "Packet pacing in small buffer optical packet switched networks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 4, pp. 1066-1079, Aug 2009.
[42] D. Ostry, "Synthesis of accurate fractional gaussian noise by filtering," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1609-1623, Apr 2006.
[43] B. Schroeder, A. Wierman, and M. Harchol-Balter, "Closed versus open: A cautionary tale," *Proc. USENIX NSDI*, USA, May 2006.
[44] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level," *Proc. ACM SIGCOMM*, USA, Aug/Sep 1995.

Arun Vishwanath (S '04) is currently a Ph.D. candidate in the School of Electrical Engineering and Telecommunications at the University of New South Wales in Sydney, Australia. He was a visiting Ph.D. student in the Department of Computer Science at North Carolina State University in 2008. His research interests include energy-efficient network architecture and design, optical networks, and router buffer sizing.

Vijay Sivaraman (M '94) received his B. Tech. degree from the Indian Institute of Technology in Delhi, India, in 1994, his M.S. from North Carolina State University in 1996, and his Ph.D. from the University of California at Los Angeles in 2000, all in Computer Science. He has worked at Bell-Labs and a silicon valley startup manufacturing optical switch-routers. He is now a Senior Lecturer at the University of New South Wales in Sydney, Australia. His research interests include optical networking, packet switching and QoS routing.

George N. Rouskas (S '92, M '95, SM '01) is a Professor of Computer Science at North Carolina State University. He received the Diploma in Computer Engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1989, and the M.S. and Ph.D. degrees in Computer Science from the College of Computing, Georgia Institute of Technology, Atlanta, GA, in 1991 and 1994, respectively. His research interests include network architectures and protocols, optical networks, multicast communication, and performance evaluation. He is coeditor of the book Traffic Grooming for Optical Networks: Foundations, Techniques and Frontiers (Springer 2008). He is a recipient of a 1997 NSF Faculty Early Career Development (CAREER) Award. Dr. Rouskas is a member of the ACM and of the Technical Chamber of Greece.