Are Bigger Optical Buffers Necessarily Better?

Arun Vishwanath*, Vijay Sivaraman* and George N. Rouskas+
*School of EE&T, University of New South Wales, Sydney, NSW 2052, Australia
Emails: {arunv@ee.unsw.edu.au, vijay@unsw.edu.au}
+Department of Computer Science, North Carolina State University
Raleigh, NC 27695-8206, USA, Email: {rouskas@ncsu.edu}

***Abstract*— Conventional wisdom suggests that bigger switch buffers translate to lower packet loss. However, we have observed in simulations (using *ns2*) that buffer sizes in the range of interest for optical packet switched networks show unexpected behaviour: larger buffers can cause higher losses for open-loop (real-time) traffic when it multiplexes with closed-loop (TCP) traffic. In this short paper we develop a simplified Markov Chain model that helps explain this anomalous behaviour. The phenomenon observed in this paper can be of serious concern to all-optical packet switch designers and network service providers, who make huge investment in setting up the network infrastructure, but only to realise potentially degraded performance if appropriate care is not taken when dimensioning their router buffer sizes.**

I. Introduction

A fundamental concern in optical packet switched (OPS) networks is packet loss due to contention at output links of switching nodes in the network. Unlike in electronic switches, where as many as a million contending packets can easily be stored in RAM, buffering in optical switches still remains a very complex and expensive operation. Until recently, it was widely believed that a core Internet router mandates $B = T \times C$ of buffering to maintain 100% link utilisation, where T denotes the average round-trip time (RTT) of a TCP flow through the router, and C the capacity of the bottleneck link interface. This rule is commonly referred to as the rule-of-thumb [1]. In 2004, researchers from Stanford University first challenged the rule by showing that if a large number N of long-lived TCP flows multiplex at a bottleneck link router (as is common in today's backbone links), then the lack of synchronisation amongst the flows permits a near-100% utilisation of the bottleneck link with only $B = T \times C/\sqrt{N}$ of buffering [2].

While this drastically reduces the buffer size required at core Internet routers, the amount of buffering needed is still prohibitively large to move to an all-optical packet switched Internet core. OPS node architectures employing fibre delay lines (FDLs) [3] can buffer packets in the optical domain by circulating it within spools of fibre (thus delaying it before sending it out of the output interface). However, the high speed of light warrants large fibre spools for even minimal amount buffering (for e.g., 1 km of fibre can buffer light for only 5μsec). Further, incorporating FDLs into a typical optical switch design (such as the shared memory architecture [4]) requires larger optical crossbars, making them bulky and adding significantly to the cost of the optical switch as the amount of buffering needed increases. Meanwhile, there has been advances in the design and prototyping of on-chip optical memory devices. It is shown in [5] that it is possible for emerging integrated photonic circuits to be able to buffer packets in the optical domain. However, they are expected to buffer at most a few dozen packets, making it practically difficult to build on-chip optical memory circuits that can buffer hundreds of packets. All these research efforts suggest that if OPS networks are to be realised in practice in the foreseeable future, then we have to make do with only very limited buffering.

II. Observation and Motivation

Given the severely constrained buffer capacity at OPS nodes, a worthwhile question to ask is whether packet loss rates will be prohibitively high to permit a move to an OPS Internet core. Recently, researchers from Stanford further argued in [6] that if TCP flows were to space out the packets they send into the core, as few as 20-50 packet buffers suffice at core nodes to realise high (over 80%) link utilisation. This claim was supported by their experimental results in Sprint ATL, and being verified by groups at Bell Labs and Georgia Tech [7]. If true, the small sacrifice in link capacity in order to facilitate all-optical buffering and switching seems worthwhile, particularly since most operators over-provision capacity in their backbone networks anyway.

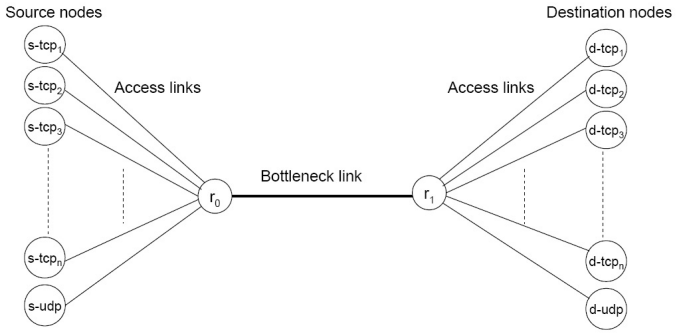

Fig. 1. ns2 simulation topology

The arguments on the feasibility of such small (optical) buffers have largely focused on TCP traffic, which accounts for 90-95$\%$ of traffic on the Internet. As real-time applications such as audio/video, gaming, etc. become more widespread in the Internet, it is also important to consider loss performance for open-loop UDP traffic. A natural intuition is that larger optical buffers would benefit not only closed-loop TCP traffic,

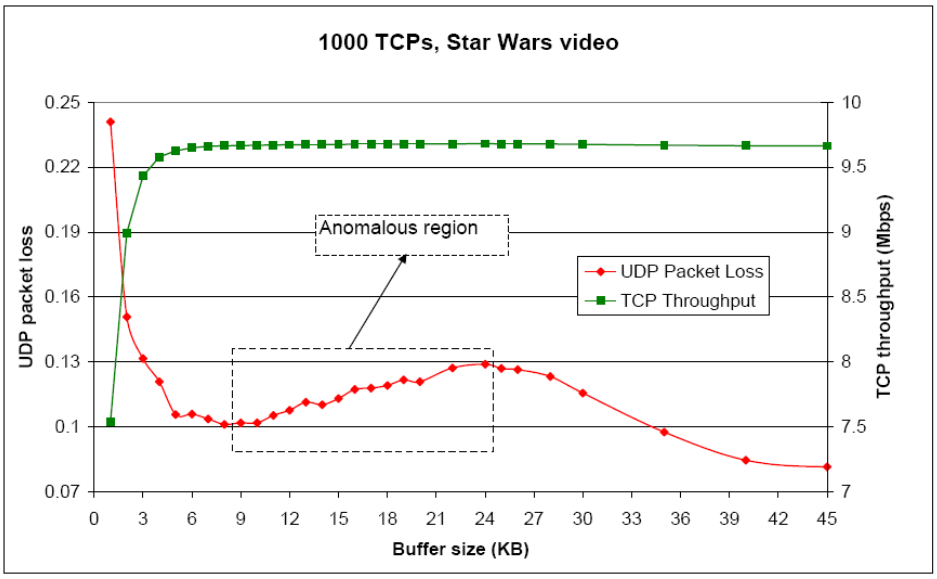

Fig. 2. Starwars video: UDP packet loss and TCP throughput

but also open-loop UDP traffic, by absorbing bursts and reducing loss probability. It was therefore surprising when our observations in simulations (reported extensively in [8]) were contrary to intuition – we describe here one scenario that illustrates the phenomenon.

A dumb-bell shaped network topology (Fig. 1) with a single bottleneck link is simulated with 1000 TCP flows having random round-trip and start times, together with a real UDP traffic trace from the movie *Star Wars* (obtained from [9]). The UDP traffic constituted $\approx 4\%$ of the bottleneck link rate (consistent with the UDP traffic volume in the Internet core). TCP and UDP packet sizes were set at 1000 Bytes and 200 Bytes respectively, which is consistent with observations in the Internet that UDP packet sizes are smaller as they often require low latencies. Fig. 2 shows the UDP packet loss and TCP throughput curves as a function of buffer size, and indicates the presence of an "anomalous region", i.e., a continuous region of buffer sizes wherein the packet loss for real-time traffic increases with increasing buffer size.

This phenomenon has serious implications for optical packet switch designers. It suggests that when TCP and UDP traffic interact, there is a regime of buffer sizes (8-24 KB in our example) in which larger buffers give **worse** performance for real-time traffic while only marginally benefiting TCP throughput. Given that each extra KiloByte of optical buffering can add significantly to the cost of the optical switch, manufacturers and operators should be wary of the potential for **negative** returns on this investment.

The anomalous loss performance was observed in simulation under a wide range of settings for UDP (short-range and long-range models) and TCP (round-trip times, number of flows, etc.), as reported in [8]. In this paper we develop an original Markov-chain based analytical model that attempts to explain this anomalous behaviour in a simplified way.

III. A Markov Chain Model

Our objective is to develop a highly simplified model that provides some analytical insight into why real-time traffic shows anomalous loss behaviour when multiplexed with TCP traffic in the regime of optical buffer sizes. To this end we make the following assumptions:

Assumption 1: UDP packets are on average smaller in size than TCP packets. This has been reported in several measurements of traffic in the Internet core [10], and is attributed to the stricter latency requirements of real-time applications such as video, and on-line gaming applications that use UDP [11]. Consistent with our example presented in Fig. 2 above, we choose average TCP and UDP packet sizes to be 1000 and 200 Bytes respectively.

Assumption 2: UDP and TCP packet arrivals are Poisson. If the number of TCP flows is large (say 1000 or more), it is believed they do not synchronise their window dynamics behaviour, and can 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 be Poisson.

Assumption 3: The aggregate TCP rate increases exponentially with bottleneck link buffer size. If B denotes the bottleneck buffer size (in KB), then the TCP throughput λ_{TCP} is given by:

$$\lambda_{TCP} = \{1 - (e^{-B/B^*})\} * \lambda_{TCP}^{sat} \quad (1)$$

where λ_{TCP}^{sat} denotes the saturation throughput of TCP for very large buffer size, and B^* is a constant (with same unit as B) that depends on system parameters such as link capacity, round-trip times, etc. The exponential rise in TCP throughput with buffer size has been reported by previous researchers [12, Sec. III], [6, Fig. 1]. We have also observed this in

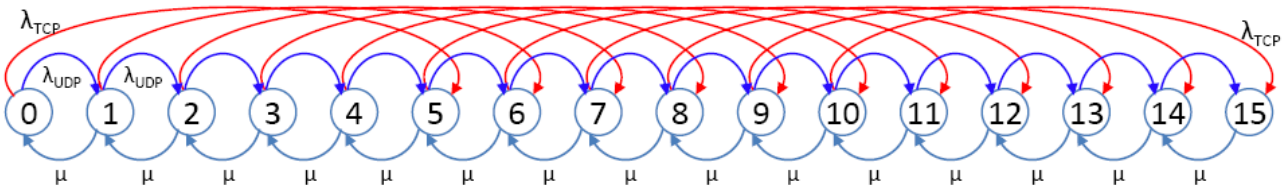

Fig. 3. Markov chain state transition diagram for buffer occupancy with buffer size = 3000 Bytes

simulation; as an illustration Fig. 4 shows on log scale the idle buffer probability as a function of buffer size when 1000 TCP flows multiplex at the bottleneck link. The linear behaviour in the range 5-50 KB demonstrates that TCP leaves the buffers idle exponentially less often as buffer size grows, implying that its throughput rises exponentially with buffer size. This plot also allows us to estimate B^* (the slope of the log-linear curve being $-1/B^*$) to be 6 KB, which we will use in our analysis below.

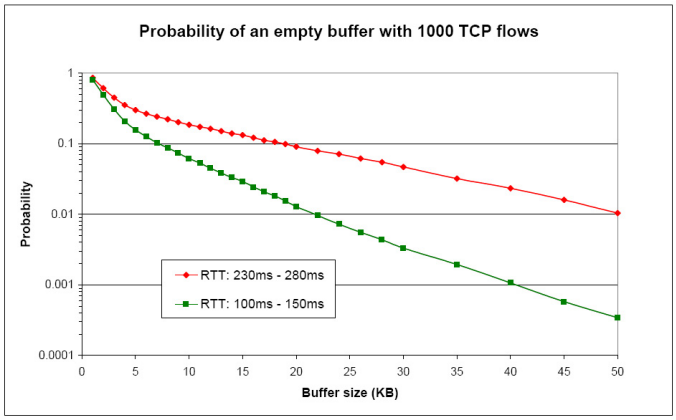

Fig. 4. Probability of idle buffer vs. buffer size for TCP traffic

With the above assumptions, we can model the FIFO queue at the bottleneck link as an M/M/1 system with finite buffer B and with two classes of customers:

1) UDP arrivals are Poisson at fixed rate λ_{UDP} and require exponential service time with unit mean (the service rate is normalised to average UDP packet size), and
2) TCP arrivals (denoted by λ_{TCP}) are Poisson at rate derived from Equation 1, 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. 3 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 $B_{bytes} = B * 1000 = 3000$ to be the corresponding buffer size in Bytes, and N the number of states in the Markov chain, then

$$N = \frac{B_{bytes}}{UDP\,packet\,size} + 1 = \frac{3000}{200} + 1 = 16. \quad (2)$$

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:

$$p_0\,(\lambda_{UDP} + \lambda_{TCP}) = p_1\,\mu \quad (3)$$

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

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

$$p_i\,(\lambda_{UDP} + \mu) = p_{i-1}\,\lambda_{UDP} + p_{i+1}\,\mu + p_{i-5}\,\lambda_{TCP} \quad (11 \le i \le 14) \quad (6)$$

$$p_{15}\,\mu = p_{14}\,\lambda_{UDP} + p_{10}\,\lambda_{TCP} \quad (7)$$

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 obtain the UDP packet loss probability.

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 $\lambda_{TCP}^{sat} = 0.94$ (so that TCP and UDP have a combined maximum rate less than the service rate of $\mu = 1$ in order to guarantee stability). The constant $B^* = 6$ KB, is consistent with what is obtained from Fig. 4.

Fig. 5 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 Equation 1, as a function of buffer size B. It can be observed that in the region of 1-7 KB of buffering, UDP loss falls monotonically with buffer size. However, in the buffer size region between 8-25 KB, UDP packet loss increases with increasing buffer size, showing that

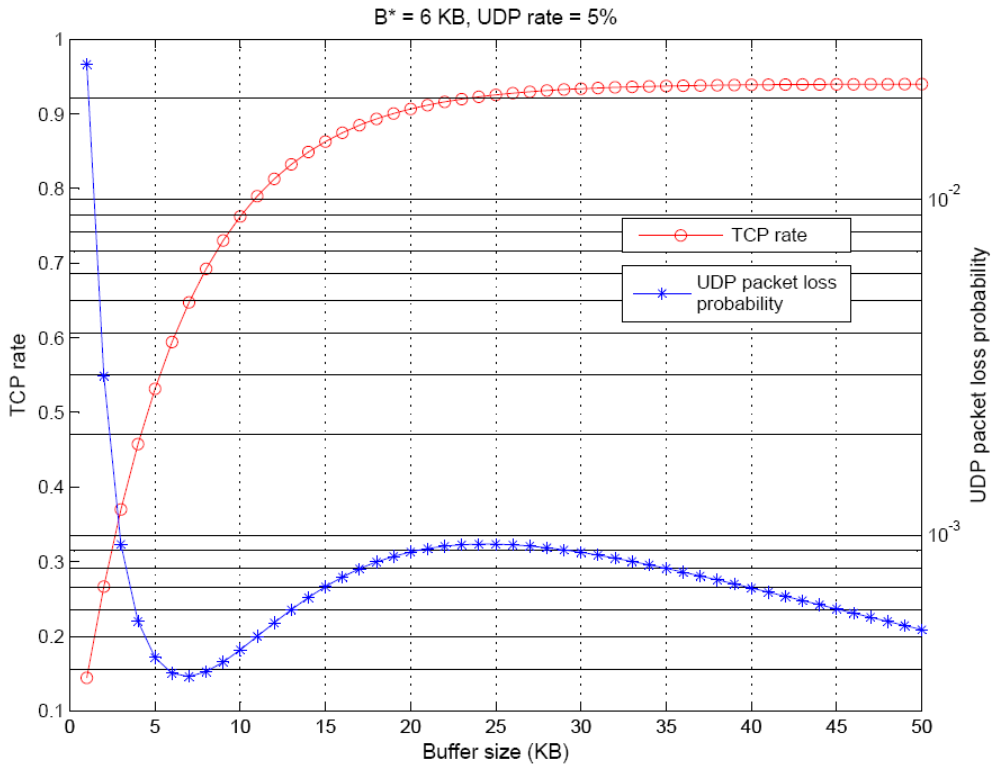

Fig. 5. Results from the Markov chain model for the anomalous loss of UDP traffic

the model is able to predict the anomaly found in simulations. A qualitative explanation of the anomaly is as follows: When the buffers are extremely small (say 1-7 KB), the congestion window of each TCP source remains extremely small as well. This results in each TCP flow transmitting only a few packets during an RTT, thus remaining idle for the most part. As a result, UDP gets exclusive access of the buffers resulting in its packet loss falling monotonically. However, in the range of about 8-25 KB, a larger fraction of TCP flows are able to increase their congestion windows, thereby leaving only a smaller fraction of the buffers for UDP traffic to use. This results in losses for UDP traffic going up in this region.

IV. Conclusions

In this short paper, we addressed the joint performance of TCP and UDP traffic at a bottleneck link optical switch equipped with very small buffers. We presented a simplified Markov Chain model to gain insights as to why real-time traffic can show counterintuitive loss behaviour when mixed with TCP traffic. It is apparent that emerging OPS networks are capable of buffering only a few dozen packets. Given this stringent constraint and the fact that adding extra buffering adds significantly to the cost of the optical switch, the anomalous behaviour studied here can be of serious concern to network service providers who make considerable investment in deploying these all-optical routers, but only to result in worse performance if they inadvertently operate their router buffer sizes in this anomalous region. Our work hopes to bring this to their attention.

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] D. Hunter, M. Chia, and I. Andonovic, "Buffering in optical packet switches," *IEEE Journal of Lightwave Technology*, vol. 16, no. 12, pp. 2081–2094, Dec 1998.

[4] 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.

[5] 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.

[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] 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.

[8] 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)*, Netherlands, June 2008.

[9] 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.

[10] Packet traces from measurement and analysis on the WIDE Internet backbone. *http://tracer.csl.sony.co.jp/mawi*

[11] 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*, Nov 2002.

[12] 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.