A Critique of Recently Proposed Buffer-Sizing Strategies

G.Vu-Brugier, R. S. Stanojevic, D.J.Leith, R.N.Shorten*
Hamilton Institute, NUI Maynooth

ABSTRACT
Internet router buffers are used to accommodate packets that arrive in bursts and to maintain high utilization of the egress link. Such buffers can lead to large queueing delays. Recently, several papers have suggested that it may, under general circumstances, be possible to achieve high utilisation with small network buffers. In this paper we review these recommendations. A number of issues are reported that question the utility of these recommendations.

Categories and Subject Descriptors
C.2.3 [**Computer-communication networks**]: Network operations

General Terms
Management, Measurement, Performance

Keywords
Link utilisation, Buffer provisioning, TCP

1. INTRODUCTION
Buffers are used at network routers to temporarily store incoming packets when the arrival of packets received exceeds the capacity of the egress link. This is done to maintain a high-level of utilization of link capacity and to accommodate bursty traffic. Traditionally, router buffers have been provisioned according to the bandwidth-delay product (BDP) rule: namely, one chooses the buffer size as $B \times T$, where B is the rate of the link served by the router, and T is the "typical" round trip time experienced by connections utilizing the link. Building upon the basic observation that, under some circumstances, only a fraction of TCP flows reduce their sending rates in response to a single congestive event, a number of recent papers have suggested the possibility of deploying significantly smaller buffers without compromising utilisation of a congested link [3, 4, 5]. For related work on determining the buffer size to achieve a desired level of utilisation for given network conditions see also [8, 7, 6, 9, 10] and the references therein.

While this work is clearly of scientific merit, in this paper we highlight a number of fundamental issues that arise in applying these results to guide buffer sizing in real networks. In particular, we make the following observations.

*This work was supported by Science Foundation Ireland grants 00/PI.1/C067 and 04/IN3/I460.

(1) The buffer sizing strategies being proposed in the literature depend crucially on the nature of the arriving traffic, with the majority of existing results related to links with a fixed number n of long-lived tcp flows. Real links almost always contain a complex mix of flow connection lengths, round-trip times, UDP traffic etc. For a given traffic mix we can try to determine an *effective* number n – the recently proposed ADT algorithm [9, 10] is one (measurement-based) way to do this for example.

(2) Even when such a refinement is used, however, the traffic mix on a link may be time-varying. Our measurements on a production link confirm that traffic patterns do indeed change significantly (here "significantly" is with respect to buffer sizing requirements) over time. This immediately calls into question the utility of fixed buffer sizing strategies in real communication networks, and potentially motivates adaptive approaches to buffer sizing. Again the ADT algorithm [9, 10] can be used to adaptively tune the required buffer size.

(3) In the context of buffer sizing it is essential to distinguish between links at which TCP flows experience packet loss and those where they do not. Roughly speaking, we can classify links as core links or access links[1]. Core network links are over-provisioned, experience essentially no queueing and generate essentially no packet loss. We contrast this with access links where significant queueing and packet loss occurs. It is these latter links that are clearly of primary relevance for analysis and discussion of the interaction between buffer sizing and elastic traffic behaviour. To illustrate the lack of clarity in the literature on this point, we note, for example, that [11] cite measurements on a backbone link where no loss occurs (for buffer sizes above 2.5ms) as qualitative evidence in support for the long-lived TCP analysis of [3]. Since the analysis of [3] relates to a link which is the bottleneck for many long-lived TCP flows and necessarily suffers from significant packet loss regardless of buffer size, the inference here is clearly questionable.

(4) As we have already mentioned, link utilisation and queueing delay do not by themselves fully capture the quality of service perceived by TCP flows. As noted by [6], packet loss is also important. We illustrate this in the context of

[1]We emphasise that these names are suggestive only. A heavily loaded link in the network core at which significant packet loss occurs would be classed as an "access" link for our purposes

the use of a very small queue on a live production link – as discussed later, we had in fact to end our test prematurely (after 2.5 hours) owing to the high number of user complaints regarding link quality when using a small queue.

With hindsight, many of these points are perhaps unsurprising. Their fundamental relevance to the recent literature on buffer sizing is nevertheless self-evident and arguably points to the need for an expanded research agenda in this area.

2. LAB TESTBED MEASUREMENTS

We begin by presenting a number of lab testbed measurements illustrating the dependence of buffer sizing on traffic conditions. Following previous work, we consider the size of buffer required to achieve a target level of link utilisation – we use a 95% target here, but obtain broadly similar results with other target values.

2.1 Testbed setup

The testbed consisted of commodity PCs connected to gigabit switches to form the branches of a dumbbell topology. All sender and receiver machines used in the tests have identical hardware and software configurations as shown in Table 1 and are connected to the switches at 1Gb/sec. The router, running the FreeBSD dummynet software, can be configured with various bottleneck queue-sizes, capacities and round trip propagation delays to emulate a wide range network conditions. Flows are injected into the testbed using `iperf`. Web traffic sessions are generated by dedicated client and server PCs, with exponentially distributed intervals between requests and Pareto distributed page sizes. This is implemented using a client side script and custom CGI script running on an Apache server.

We employed the ADT algorithm [9] to determine the buffer size required to achieve the target link utilisation. This is an online, measurement-based algorithm. It is implemented as a user space perl script running on the FreeBSD router, with pseudo-code shown in Algorithm 1.

	Description
CPU	Intel Xeon CPU 2.80GHz
Memory	256 Mbytes
Motherboard	Dell PowerEdge 1600SC
Kernel	Linux 2.6.6
txqueuelen	1,000
max_backlog	300
NIC	Intel 82540EM
NIC Driver	e1000 5.2.39-k2
TX & RX Descriptors	4096

Table 1: Hardware and Software Configuration.

2.2 Experimental measurements

.

2.2.1 *Impact of mix of RTTs on buffer provisioning*

Figure 1 plots the measured buffer provisioning required to ensure 95% link utilisation when all flows have 120ms RTT and also when the flow RTTs are uniformly distributed between 20 and 270ms. Also marked on Figure 1 is $B \times T/\sqrt{n}$.

Algorithm 1 Pseudo-code for ADT algorithm

```
dT=3000 # update interval in seconds (5 minutes)
c = 1.1 # update gain
qmin specifies minimum buffer size in packets
qmax specifies maximum buffer size
while 1 do
  old ← number of bytes sent on link
  wait for time dT
  new ← number of bytes sent on link
  throughput ← (new − old)/dT
  if throughput < 0.95 ∗ capacity then
    qadt ← int(qadt ∗ c)
  else
    qadt ← int(qadt/c)
  end if
  qadt ← max(min(qadt, qmax), qmin)
  buffer size ← qadt
end while
```

These results are similar to those reported elsewhere and serve as a validation check on our experimental setup and the operation of the ADT algorithm.

2.2.2 *Impact of mix of connection lengths on buffer provisioning*

Figure 2 shows the corresponding results obtained when the number of long-lived flows is held constant and the number of competing web sessions is varied. For the web sessions we used standard values: namely, a mean time between requests of 1 second and a Pareto shape parameter of 1.2. It can be seen that even small numbers of web sessions are sufficient to have a significant impact on the choice of buffer here. This is perhaps unsurprising as the web sessions increase the number of flows that are in slow-start at any given time and so can be expected to increase the burstiness of the packet arrivals at the router. However, since web traffic is ubiquitous in modern networks, this observation has direct implications for the utility of analytic results focussed purely on links carrying long-lived flows.

Motivated by this observation, it seems important to investigate the buffer sizing requirements of a realistic traffic mix. Since the types of traffic mix encountered on real access links remains relatively poorly characterised, rather than pursuing further lab testing we next consider test results from a production link.

3. MEASUREMENTS FROM A PRODUCTION LINK

Following initial testing in our lab testbed, we investigated the buffer sizing requirements on a production link.

3.1 Test setup

For our tests we used the gateway connecting the Hamilton Institute to the wider campus network, which in turn is connected to the public internet via a 64Mbps link. The Hamilton Institute native link capacity is 100Mbps and the local network contains around 100 networked computers. Since this link is normally uncongested (no packet losses), for testing purposes we used a FreeBSD Dummynet box inserted

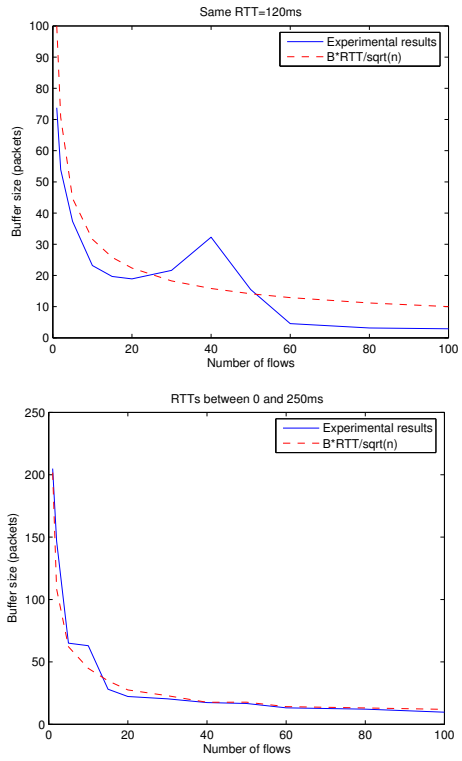

Figure 1: Buffer size for 95% link utilisation versus number of flows. In top plot all flows are long-lived and have the same RTT of 120ms, in lower plot flow RTTs are uniformly distributed between 20 and 270ms. 10Mbps bottleneck link.

between the local network and the gateway to throttle the link speed to 1Mbps – this link speed was selected based on the measured 5 minute average traffic load on the uncongested link.

Traffic on the gateway was measured using a combination of `tcpdump, tstat, snmp` (packet transmissions, packet loss etc). Link quality was also measured by active probing using `ping` to measure delay and verify the reported loss rate. Statistics on the download completion times for three selected web sites were also collected.

The measured distribution of TCP flow connection lengths and round-trip times on the link over a 24 hour period on 24th September 2006 (the mid-point in our tests) are shown in Figures 3 and 4.

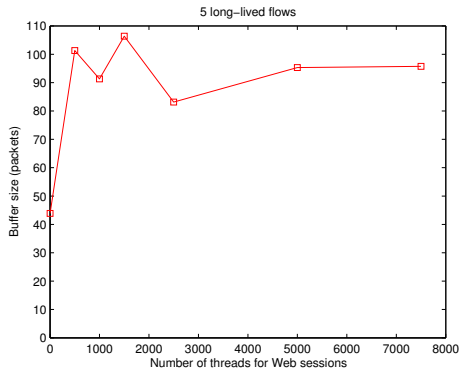

Figure 2: Buffer size for an increasing number of web sessions (5 long-lived flows)

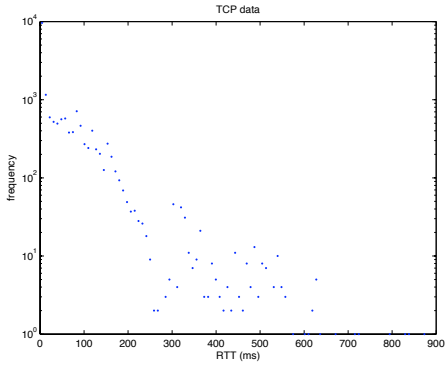

Figure 3: Distribution of round-trip times (taken from time-stamps in TCP data and ack packets).

3.2 Buffer sizing measurements

We used the ADT algorithm, running on the gateway, to estimate the buffer size required to achieve a target link utilisation of 95%. The maximum buffer size is capped at the bandwidth-delay product $B \times T$, where B is the link bandwidth (1Mbps) and T is 250ms (which corresponds to 31.25KB or 21 1500 byte packets). See Figure 5 for the data collected over one day, starting at 6am on the morning of 27th Sept 2006. Measurements are shown of the five minute average link utilisation, buffer size (with ping data overplotted to indicate level of queue occupancy) and packet loss rate. We make the following observations.

1. The traffic load, and associated buffer requirement, is strongly time-varying. To maximise utilisation during periods of light load requires the use of a large buffer and it can be seen that ADT selects the largest admissible buffer size during such periods. This requirement for large buffers is also evident from the earlier testbed measurements when the number of flows n is small. During periods of heavy load, smaller buffers can be used without compromising the target of 95% link utilisation – see for example the measurements during 12:00-14:00 in Figure 5.

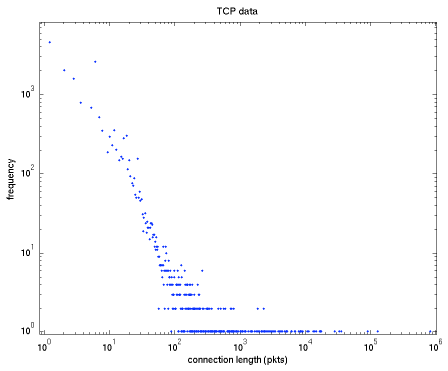

Figure 4: Distribution of connection sizes (number of packets containing payload data) .

2. The loss rate is high during periods of heavy load – reaching nearly 25% of packets at some points. It is unclear whether this high loss rate is directly linked ot the queue size or whether it is associated with the heavy load. We can see, for example, that the loss rate is high around 12:30 hours even though the buffer size is relatively large at this time. Although we could not reproduce identical traffic conditions as this was a live production link, baseline measurements taken with a fixed buffer size (specifically, a bandwidth-delay product corresponding to delay of 250ms) also show similar levels of packet loss. Of course, this might indicate that larger buffers are needed, but we leave detailed exploration of this to future work as the impact on user quality of service of queueing delay with buffer sizes greater than 250ms is felt to require detailed consideration. The impact of small buffer sizes on loss rate is, however, discussed further below.

3.3 Impact of very small buffers

It might be tempting to simply choose a small size of buffer and accept the cost of reduced utilisation during periods with few flows. We therefore also investigated the impact on link utilisation and loss of a range of choices of fixed buffer size. Of particular note was the negative impact of very small buffers on link quality. Figure 6 shows measurements obtained with a buffer size of only 3KB (or two 1500 byte packets). It can be seen that the packet loss rate remains persistently high (at around 20%) while link utilisation remains consistently below 60%. This test was performed during the day and can be compared to hours 12-14 in Figure 5. Due to the large number of user complaints concerning link quality during this test (recall that the link was in production use), the test was terminated after 2.5 hours. This cautionary experience strongly mitigates against the use in practice of very small buffers on a heavily loaded link.

3.4 Uncongested operation

In the foregoing tests, the network gateway was throttled from 100Mbps to a link speed of 1Mbps as our interest was in buffer sizing on links where queueing and packet loss occur. For comparison, we also carried out measurements with the link operating at its native speed. Under these conditions, the link experiences no packet loss (over the period of 7 days when measurements were collected) and essentially no queueing (sub-millisecond delay). This behaviour was observed to be insensitive to the buffer size used. This is unsurprising, but serves to emphasise the fact that the buffer sizing question on over-provisioned links is very different from that on heavily congested links.

4. CONCLUSIONS

We make the following observations from our measurements.

1. Real links contain a complex mix of flow connection lengths and round-trip times.

2. Traffic patterns change significantly (here "significantly" is with respect to buffer sizing requirements) over time.

3. In the context of buffer sizing it is essential to distinguish between links at which TCP flows experience packet loss and those where they do not. On over-provisioned links that experience essentially no queueing and generate essentially no packet loss the choice of buffer size has little impact on performance. We contrast this with access links where significant queueing and packet loss occurs.

4. Packet loss is an important aspect of link quality in practice. We illustrate this in the context of the use of a very small queue on a live production link – we had in fact to end our test prematurely (after 2.5 hours) owing to the high number of user complaints regarding link quality when using a very small buffer.

The fundamental relevance of these observations to the recent literature on buffer sizing is self-evident and points to the need for an expanded research agenda in this area. Our immediate conclusions point to the utility of adaptive buffer sizing solutions, provided that one accounts for both utilisation and loss rate in the adaptation strategy [12].

5. REFERENCES

[1] Z. Zhao, S. Darbha, and A. L. N. Reddy, "A method for estimating the proportion of nonresponsive traffic at a router," *IEEE Trans on Networking*, vol. 12, no. 4, pp. 708–718, 2004.

[2] J. Padhye and S. Floyd, "Identifying the TCP behavior of web servers," in *Proceedings of SIGCOMM*, 2001.

[3] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," in *SIGCOMM '04*, Portland, Oregon, USA, 2004.

[4] A. Dhamdhere, H. Jiang, and C. Dovrolis, "Buffer sizing for congested internet links," in *Proceedings of INFOCOM*, Miami, FL, March 2005.

[5] D. Wischik and N. McKeown. Part I: Buffer sizes for core routers. *Computer Comms Review*, 35(3), 2005.

[6] A. Dhamdhere, C. Dovrolis, "Open issues in router buffer sizing". *Computer Comms Review*, 36(1), pp. 87-92, 2006.

[7] D. Wischik, "Fairness, QoS and buffer sizing". *Computer Comms Review*, 36(1), pp. 93-95, 2006.

[8] K. Avrachenkov, U. Ayesta, A. Piunovskiy, "Optimal choice of buffer sizes in the internet". *Proceedings of IEEE Conference on Decision and Control*, 2005.

[9] R. Stanojevic, R. Shorten, C. Kellett. "Adaptive tuning of Drop-Tail buffers for reducing queueing delays". IEEE Comms Letters, vol. 10 (7), July, 2006.

[10] C. Kellett, R. Shorten, D. Leith. "Sizing internet buffers, Active queue management, and the Lur'e problem". *Proceedings of IEEE Conf on Decision and Control*, 2006.

[11] Y. Ganjali, N. McKeown. "Update on Buffer Sizing in Internet Routers". *Computer Communications Review*, October 2006.

[12] R. Stanojevic, R. Shorten. "How expensive is link utilization". Preprint, available online http://www.hamilton.ie/person/rade/QP.pdf.

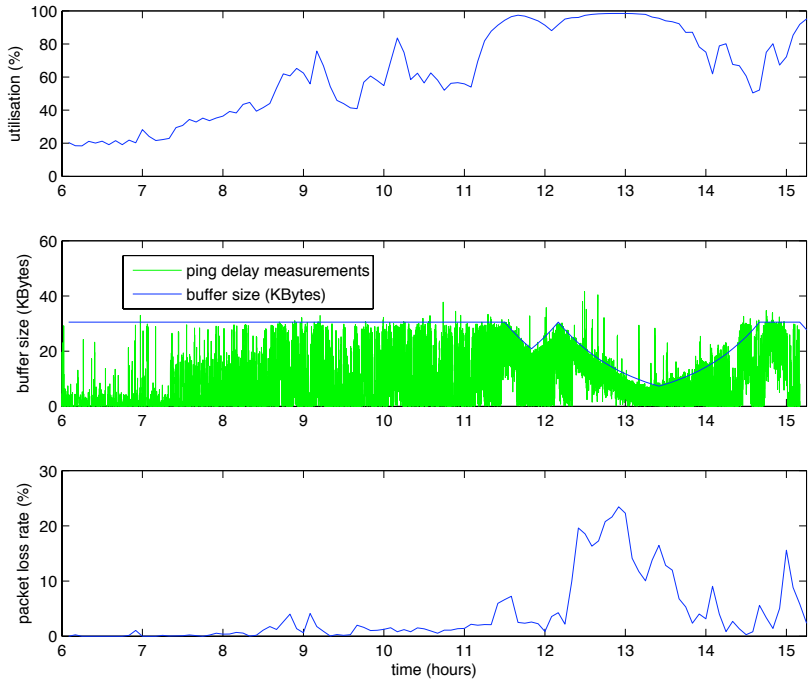

Figure 5: Evolution of the link utilization (top), buffer size and ping response time (middle), and loss rate (bottom) from 06:00 to 15.30 on Wed, 27 September 2006.

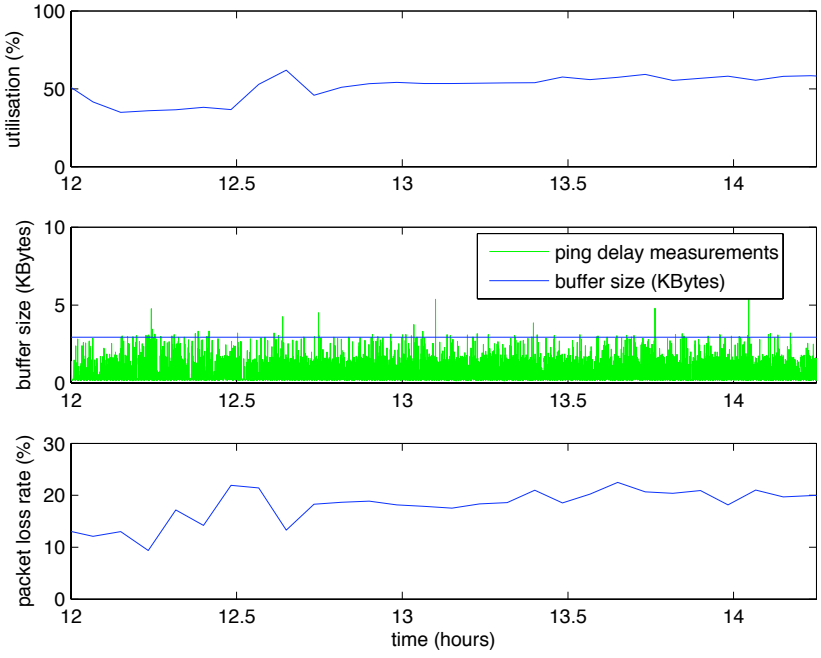

Figure 6: Small buffer (3KB or two 1500 byte packets). Evolution of the utilization (top), buffer size and ping response time (middle), and loss rate (bottom) from 12:00 to 14:25 on Fri, 29 September 2006.