****

ATM Forum Document Number: ATM_Forum/96-0517

****

Title: Buffer Requirements for TCP over ABR

****

Abstract:

In our previous study [2], it was shown that cell loss due to limited buffering may degrade throughput considerably. The key question is how much buffering is required to avoid cell loss. This contribution attempts to answer that question.

We show that the maximum buffers required at the switch is proportional to the maximum round trip time of any VC through the link. The number of round-trips depends upon the the switch algorithm used. With our ERICA (modified) switch algorithm, we found that the buffering required is independent of the number of TCP sources. These observations are valid even when VBR or two-way traffic is used. We substantiate our arguments with simulation results.

Our simulations are carried out with a modified version of the ERICA algorithm. We also give a brief description of the modifications to ERICA which include avoidance of unnecessary spikes, correct counting of bursty sources and enhanced fairness.

****

Source:

Raj Jain, Shiv Kalyanaraman, Rohit Goyal and Sonia Fahmy
The Ohio State University
Department of CIS

Raj Jain is now at Washington University in Saint Louis, jain@cse.wustl.edu http://www.cse.wustl.edu/~jain/

Saragur Srinidhi
NASA Lewis Research Center and Sterling Software
Scientific Systems Division
21000 Brookpark Road, MS 141-1
Cleveland,OH 44135
Phone: 216-433-8987, Fax: 216-433-8000

The presentation of this contribution at the ATM Forum is sponsored by NASA.

****

Date: April 1996, Anchorage, Alaska.

****

Distribution: ATM Forum Technical Working Group Members (Traffic Management)

****

is offered to the Forum as a basis for discussion and is not a binding proposal on the part of any of the contributing organizations. The statements are subject to change in form and content after further study. Specifically, the contributors reserve the right to add to, amend or modify the statements contained herein.

****

INTRODUCTION

Given the popularity of TCP/IP, it is important to verify that all source, switch, and destination rules specified for ABR perform as expected for TCP/IP traffic. In our last contribution [2] and a related paper [3], we studied the throughput and loss behavior of TCP over ABR with limited buffers. We observed a considerable drop in throughput even though the CLR was very small. Increasing buffers was found to improve TCP throughput. Maximum TCP throughput (with zero cell loss) was observed for cases with sufficiently large buffers. We also reported that the buffers should not be dimensioned based on the TBE parameter.

It was pointed out that most ABR switches would provide at least a round trip time worth of buffers, where the round trip time is measured for the longest VC passing through the switch. Hence, in this contribution, we attempt to quantify the buffer requirements for ABR to achieve the maximum TCP throughput with zero loss. We observe that the buffer requirement is proportional to the round trip time. The results depend upon the switch algorithm used. For our modified ERICA switch algorithm, we found it to be independent of the number of TCP sources. These observations are valid even when VBR or two-way traffic is used. Vanilla UBR, in comparison, requires buffers proportional to the sum of the receiver windows [4], which is proportional to the number of TCP sources.

We experiment with an infinite TCP source running on TCP over an ATM WAN. The TCP source always has a frame to send. However, due to TCP window constraint, the resulting traffic at the ATM layer may or may not be continuous. A description of our TCP code and source model is given in [2]. Our simulations are carried out with a modified version of the ERICA algorithm, described later in this contribution. The original ERICA algorithm was described in [1].

TCP OPTIONS:

We use a TCP maximum segment size (MSS) of 512 bytes. The MTU size used by IP is generally 9180 bytes and so there is no segmentation caused by IP. We implemented the window scaling option so that the throughput is not limited by path length. Without the window scaling option, the maximum window size is 2**16 bytes or 64 kB. We use a window of 16x64 kB or 1024 kB. The network consists of three links of 1000 km max each and therefore has a max one-way delay of 15 ms (or 291 kB at 155 Mbps). In our simulations, we have not used the "fast retransmit and recovery" algorithms. These algorithms exhibit different behavior for bursty losses which we plan to study separately. However, the zero-loss buffer requirement is valid for fast retransmit and recovery too.

THE N SOURCE + VBR CONFIGURATION

The N Source + VBR configuration has a single bottleneck link (LINK1) shared by the N ABR sources and possibly a VBR source. All links run at 155 Mbps and are of the same length. We experiment with the number of sources, the link lengths, and with/without the VBR background.

The VBR background is optional. When present, it is an ON-OFF source with a 100 ms ON time and 100 ms OFF time. The VBR starts at t = 2ms to avoid certain initialization problems. The maximum amplitude of the VBR source is 124.41 Mbps (80 of link rate). VBR is given priority at the link, i.e, if there is a VBR cell, it is scheduled for output on the link before any waiting ABR cells are scheduled.

All traffic is unidirectional. A large (infinite) file transfer application runs on top of TCP for the TCP sources. N may assume values 1, 2, 5, 10, 15 and the link lengths 1000, 500, 200, 50km.

SEVEN FACTS ABOUT TCP's CONGESTION CONTROL

---

1. TCP slow start successfully avoids congestion collapse

2. TCP can automatically fill any available capacity

3. TCP performs best when there is NO packet loss. Even a single packet loss can reduce throughput considerably

4. Slow start limits the packet loss, but loses considerable time. With TCP, you may not lose too many packets, but you lose time.

5. Bursty losses cause more throughput degradation than isolated losses.

6. Fast Retransmit/Recovery helps in isolated losses, but not in bursty losses {losses in a single window}.

7. Timer granularity is the key parameter in determining the time lost.

SEVEN OBSERVATIONS ABOUT TCP OVER ABR:

---

1. ABR performance depends heavily upon the switch algorithm. The following statements are based upon the modified ERICA switch algorithm.

2. Other key parameters are: round-trip time, number of sources, and feedback delay (from the bottleneck switch to the source and back).

3. There is no loss for TCP, if the switch has buffers equal to 4*RTT. This is true for any number of sources.

4. Observation 3 is true, even when there are CBR and VBR traffic in the background.

5. If there is no VBR in the background, then 3*RTT buffers are sufficient for no loss.

6. Under many circumstances 1*RTT buffers may do.

7. Drop policies improve throughput. But a proper drop policy is less critical than a proper switch algorithm.

The derivation of 4*RTT is based along the following arguments:

1. Initially the TCP load doubles every RTT.

2. The minimum number of RTTs required to reach rate-limited operation decreases as the log of the number of sources.

3. When the pipe just becomes full, the maximum queue is 1*RTT*Link Bandwidth

4. Queue Backlogs due to bursts smaller than RTT is 1*RTT*Link_Bandwidth

5. Bursty behavior of ACKs causes an additional 1*RTT*Link_Bandwidth queues

6. VBR contributes 1*RTT*VBR_bandwidth to the queue

7. Switch Schemes may contribute some more to the queue.

The sum of all these components is approximately 4*RTT.

Modified ERICA:

ERICA has been modified for the following:

1. To elminate many short spikes in ACR

2. To provide fast response even when the link is underutilized.

3. Correctly counts bursty sources

4. Allows multiclass scheduling in the presence of CBR, VBR, UBR, etc.

5. Achieves better fairness in many cases

SAMPLE SIMULATION RESULTS

We present only the maximum queue results for some of our simulations here. A larger set of simulation results and graphs will be presented at the ATM Forum and will be available at our www site after the Forum meeting.

In almost all the cases, we observe that the maximum queue is bounded by 3*RTT*Link Bandwidth. The bound is 3 * 30 ms * 368 ms = 33120 cells for 1000km(30ms) configurations, 16560 cells for 500km (15ms) configurations, 6624 cells for 200 km (6ms) configurations, 1656 cells for 50km (1.5ms) configurations.

Table 1: Effect of number of sources

Number of Sources	RTT(ms)	Feedback Delay (ms)	Max Q size(cells)

1	30	10	2 = 0*RTT
2	30	10	3056 = 0.37*RTT
5	30	10	10597 = 0.95*RTT
10	30	10	14460 = 1.31*RTT
15	30	10	16128 = 1.46*RTT

In Table 1, we notice that the maximum queue size grows with the number of sources. But, stablizes at 1.46*RTT.

We repeated simulations with different link lengths (All links in each case are of the same length.) The results are shown in Table 2.

Table 2: Effect of lower RTT and Feedback delay

Number of Sources	RTT(ms)	Feedback Delay (ms)	Max Q size(cells)
15	15	5	10910 = 2*RTT
15	6	2	6842 = 3*RTT
15	1.5	0.5	2108 = 4*RTT

From table 2, we find that the maximum queues may cross the estimate of 3*RTT*Link Bandwidth. This is because, the RTT values are lower and in such cases, the maximum queue depends upon the parameters used in the switch scheme.

Now we introduce VBR. The results are shown in Table 3.

Table 3: Effect of VBR

Number of Sources	RTT(ms)	Feedback Delay (ms)	Max Q size(cells)
15+VBR	30	10	22036 = 2*RTT

This experiment was with 15 sources + VBR. We observe larger queues queue to the introduction of VBR. The excess queue (5908 cells, compared to value in Table 1) is bounded by 1*RTT*VBR_Bandwidth = 1*30*368*0.8 = 8832 cells. Note, that VBR_Bandwidth is limited to 0.8 of the Link_Bandwidth.

All the results presented so far are for ERICA with only spike fix. We then used other enhancements. The results are shown in Table 4.

Table 4: Effect of Switch Scheme

Number of Sources	RTT(ms)	Feedback Delay (ms)	Max Q size(cells)
15 (Erica +spike fix)	30	10	16128
15 (Erica+ +spike fix +bursty count +fairness)	30	10	1045

Notice that the enhanced scheme reduces the queue considerably.

Thus, the enhanced scheme needs even fewer buffers than those mentioned earlier while providing the same or better throughput.

SUMMARY

We observe that TCP can run over ABR with zero-cell loss if sufficient buffers are provided. The buffer requirement is proportional to the round trip time and heavily depends upon the switch scheme used. In particular, for a modified version of ERICA, zero loss was achieved with 4*RTT buffers regardless of number of sources even in the presence of VBR background.

REFERENCES:

[1] R.Jain, S.Kalyanaraman, R. Viswanathan, R. Goyal,"A Sample Switch Algorithm," ATM Forum/95-0178R1, February 1995.

[2] R.Jain, S.Kalyanaraman, R. Goyal,S.Fahmy, F.Lu, S.M.Srinidhi, "TCP/IP over ABR(Was: TBE and TCP/IP Traffic)," ATM Forum/96-0177R1

[3] S.Kalyanaraman, R.Jain, S.Fahmy, R. Goyal, F.Lu, S.M.Srinidhi, "Performance of TCP/IP over ABR," submitted to Globecom'96.

[4] R.Jain, R. Goyal, S.Kalyanaraman, S.Fahmy, S.M.Srinidhi, "TCP/IP over ABR(Was: TBE and TCP/IP Traffic)," ATM Forum/96-0177R1

All our past ATM forum contributions, papers and presentations can be obtained on-line at http://www.cse.wustl.edu/~jain/