A Control-theoretic Rate-based Control of Buffer-Occupancy Feedback for Real-time Multimedia Communications

Chia-Hui Wang\textsuperscript{1}, Jan-Ming Ho\textsuperscript{4}, Ray-I Chang\textsuperscript{4}, Shun-Chin Hsu\textsuperscript{4}

\textsuperscript{1}Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C. \{hoho, william\}@iis.sinica.edu.tw
\textsuperscript{4}Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. \{d5526006, schsu\}@csie.ntu.edu.tw

ABSTRACT

Rate-based flow control mechanisms have been widely deployed in real-time multimedia applications over best-effort networks due to its short propagation delay. In this paper, based on the feedback control of buffer occupancy (BO), a new rate-based flow control mechanism is proposed. Feedback control ideas have been used successfully in computer systems to control the uncertainty of complex and unpredictable workloads. Our proposed mechanism applies this idea to consider not only the behavior of TCP-friendly traffic but also the BO of the client to predict and prevent serious degradation of playback QoS. Then, an effective and efficient error control mechanism is furnished by tightly coupling rate control with packet retransmission. Such a fine-grain adaptation method can help clients without a large buffer (such as set-top box, PDA and cellular phone) to achieve a high playback QoS. We have deployed the proposed mechanism into a true VOD service testbed. Experiments show that the proposed mechanism outperforms not only the mechanism without BO feedback but also a BO feedback mechanism without the proportional and derivative (PD) control function.

Keywords: feedback control, buffer-occupancy, control theory, error control, video on demand, PD control

1. INTRODUCTION

Modern network environments such as Internet are open and unpredictable. The bandwidth of the public network backbone is shared by all kinds of applications to achieve statistical multiplexing gain. It introduces considerable uncertainty in workload and resource requirements. While delivering a packet through the network, the delay and jitter is unpredictable even if the packet is not lost (e.g. as shown in Figure 1). However, different from conventional text/image network applications, multimedia applications require end-to-end quality-of-service (QoS) with jitter-free playback of audio and video. A good real-time rate-based flow control mechanism is one of the most important parts in such time-critical applications.

Rate-based flow control prevails over window-based flow control for real-time multimedia applications because it provides end-to-end deterministic and statistical performance guarantees over the packet-switching network [1], it prevails over the window-base flow control for real-time multimedia applications in the Internet. Recent literatures [2-5] proposed the rate adjustment schemes considering TCP friendliness for the congestion condition. Their rate adjustments are basically based on the feedback from the information of packet loss and delay indicated in the periodic RTCP packet of the RTP protocol suite[12]. And their TCP-friendly AIMD (Additive Increase/Multiplicative Decrease) adaptive rate adjustment discipline consider packet loss ratio, utilization and fairness to perform the actual rate change dynamically.

However, while considering the ubiquitous connections of the Internet, many clients such as set-top box, PDA and cellular phone does not preserve much system resource. They may not be able to handle the elastic traffic introduced from the aforementioned rate adjustment schemes. As the buffer size is limited, the client has to stop decoding and displaying while playback running out of packets in the buffer. Likewise, the incoming packet will be discarded if it is flooding the buffer. Therefore, the underflow and overflow of client buffer introduced by the delay and jitter in transporting packets also reflect respectively displaying delay and jitter while playback decoding. They may jeopardize seriously the playback QoS.

In the study of buffer-occupancy feedback control, Sun et.al. [8] proposed an adaptive flow control based upon client’s queue length for the MPEG-1 stream. It considers data pre-loading, queue-length adaptive rate control and retransmission schemes related to the dynamics of running buffer occupancy. The packet sending rate \( s(t) \) is derived directly from a formula based on measured packet delay \( d(t) \), previous sending rate \( s(t-1) \), average decoding rate \( \mu(t) \), packet loss ratio \( p(t) \) and target queue length \( b_n \) where \( t \) is the rate control period. To smooth the impact of rate change to the network, the actual rate change was limited to less than nominally 1/3 of the previous sending rate in their simulation. They also assume the resultant packet receiving rate \( \lambda(t) \) would be the same as the playback rate \( \mu(t) \) to keep the queue length running nearly to \( b_n \).
Basically, our proposed rate control scheme also performs rate adaptation and retransmission based on the feedback of client’s buffer-occupancy. But we explore further for a fine-grain feedback control function and economical retransmission mechanism to provide reliable real-time multimedia communication service. In fact, the control theoretic feedback control idea has been used successfully in computer systems with complex and unpredictable workloads. They demonstrated the effectiveness of feedback control mechanisms in achieving predictable system performance without precise knowledge of worst-case load patterns. We extend the control theoretic feedback control idea on the unpredictable network performance for real-time multimedia communications. It takes much more challenge than the similar theoretical feedback control of local buffer-occupancy deployed in smoothing algorithm for the MPEG stream encoder [6]. Furthermore, the proposed rate adaptation must meet the following four objectives:

- minimize fluctuations of buffer-occupancy
- maximize the possibility for timely retransmission in error control.
- avoid sending unnecessary feedback control messages due to the communication overhead
- considering the TCP-friendly behavior

In Section 2, we introduce the feedback control mechanism and the linear proportional and derivative (PD) control function. Experimental results of our proposed method are shown in section 3. In section 4, we analyze the proposed scheme by mathematical proof. Concluding remarks and future works are given in the last section.

2. CONTROL THEORETIC RATE-BASED FEEDBACK CONTROL OF BUFFER-OCCUPANCY

In this section, firstly the applied control theory will be simply introduced and we describe the proposed buffer-occupancy rate-based control mechanism in detail. Secondly, we will present the applied PD control function. Finally, we continue to the retransmission mechanism based on our rate adaptation scheme.

2.1 Buffer-occupancy rate-based feedback control

In control theory, the feedback control system as shown in Figure 2(a) can be described by a device to be controlled, an actuator and a sensor. Timely, the system monitors and compares the error between the current value and the target value of the sensor. If necessary, the controller will perform a feedback function to the system based on this error. An adaptive algorithm is introduced to support applications in dynamic uncertain environments.

The feedback information is needed to provide the network dynamics and then to predict the workloads and to support a good rate regulator as shown in Figure 2(b). The running buffer-occupancy of the client results from not only the elastic traffic introduced by the rate adaptation but also the decoding rate while the client playback the received packets. Therefore, the proposed rate-based feedback control mechanism must be adapted the variations of server sending rate, client receiving rate and client playback rate (as shown in Figure 1) to keep the playback QoS intact at its best.

Note that in multimedia communications, each client has only a limited buffer. If the buffer is flooded, then the incoming packets are discarded. On the other hand, if the buffer is underflow, the display is then temporarily in freeze. In this paper, we regard the rate regulator as a controller, and use buffer's occupancy to predict its possible overflow/underflow. In this section, the target values of a sensor are thresholds of buffer-occupancy. In this paper, we use the low threshold \( b_l \) and the high threshold \( b_h \) of buffer-occupancy to guard against buffer overflow and underflow. Let the target value of buffer-occupancy be \( b_m \) where \( b_m \) is defined as the middle point of \( b_l \) and \( b_h \). Based on the feedback control theory, at any time \( t \), our algorithm controls the buffer-occupancy \( b(t) \) to be running close to \( b_m \). Therefore, the controller can prevent buffer overflow/underflow to achieve satisfactory playback quality.

To save communication overhead of timely feedback control, client can send the feedback control to server only when the buffer-occupancy does not run within \([b_l, b_h]\) interval. Statistically, the rate of feedback control can be deducted by

\[
\frac{b_l - b}{b_{max}} = \frac{b}{b_{max}}
\]

\( b_{max} \) is maximum size of the client buffer. It can sustain the delay jitter up to \( b_{max} \) divided by the network bandwidth.

Two terms are raised here as the criteria to judge the performance to achieve the first one of four objectives aforementioned in section 2 to minimize the fluctuations of buffer-occupancy. The first term is defined as “threshold miss” (T.M.) while the receiving packet did not arrive in the position within \([b_l, b_h]\). The second one is defined as “serious miss” (S.M.) while the client buffer was in the condition of overflow or underflow. We will utilize these terms in the experiments discussed in the section 3.

2.2 Rate control function

Figure 3. Illustrated curves of rate functions.
For the feedback controller, many possible designs are available. We focus on a linear proportional and derivative (PD) control [10] for its theoretical stability and practical values. Because the only proportional control of rate adaptation may result in rate oscillations theoretically, it won’t reach to stable state to maintain the playback quality.

Note that, as shown in Figure 3, the running position of buffer-occupancy depends on not only its current value $b(t)$ but also the rate-difference $r(t)$ between packet receiving rate $\lambda(t)$ and packet playback rate $\mu(t)$ and the server sending rate $s(t)$ will also effect the $\lambda(t)$. The network dynamic impact to the difference between $s(t)$ and $\lambda(t)$ will be discussed in detail later in the section 4. We can assume no difference between them and define $r(t)$ as follow:

$$r(t) = \frac{db(t)}{dt} = \lambda(t) - \mu(t).$$

Before we figure out the formula of the PD control function, we should consider some criteria the PD control function should meet in practice as follow:

- The idea curves for rate difference $r(t)$ and buffer-occupancy $b(t)$ are all illustrated in Figure 4 while the initial conditions both of $b(0)$ and $r(0)$ are zero. The mathematical results conducted from the continuous case of the proposed PD control function are mentioned.

### 2.3 Retransmission mechanism for error control

Considering the deployment of the error control mechanism for packet loss in the reliable multimedia communications without violating the real-time constraints. The proposed retransmission mechanism based on the buffer-occupancy feedback control is simply illustrated in Figure 5. Because the decoder will consume the buffer content in a FIFO manner, the retransmission packet should be sent out from server to client as soon as possible before client can not playback the retransmitted packet. The higher the buffer-occupancy running higher more valid retransmission will be received at client. Valid retransmission is defined as a retransmission packet can be decoded in time while server retransmit the packet upon client’s retransmission request. The running buffer-occupancy can help to indicate if the client is too late to send the retransmission request to prevent the redundant traffic in the network. Multiple retransmissions are applicable in our proposed scheme.

Besides, more than one consecutive packets are lost in network can be considered as an indication of congestion. At the mean time, the multiplicative decrease (MD) part of TCP-friendly AIMD rate adaptation can be applied to help our scheme to achieve TCP-friendly behavior.

### 3. EXPERIMENTS AND PERFORMANCE RESULTS

In this paper, the testbed is based on a previous work [8] of true VOD (Video-On-Demand) services in our ADSL trial. As shown in Figure 6, our clients are running on a different network segment with the server. A router in which a Dummynet simulation tool [11] is running then connects these two network segments. By applying this tool, we can approximate the real network model to simulate different effects of finite routing queue, bandwidth limitation, communication delay and packet.
drop rate. In the following, detailed experiments are shown to evaluate the performance of the proposed scheme.

3.1 Experimental background

In the test bed, there are three set-top boxes of clients are running. Two of them are the BORC (Buffer-Occupancy Rate Control) clients using our proposed feedback-control mechanism. The last one is a non-BORC client. The server performs the discrete PD control function while receiving the buffer-occupancy feedback from BORC clients. While receiving the feedback of loss and delay from non-BORC client, the server performs the rate control function of TCP-friendly rate adaptation mechanism proposed in [5]. It picks up the minimal one among the values derived from the rate-change-smoothing function, the exponential-rate-limit function and the TCP-bandwidth-share function [5] to increase the rate while no congestion.

Much like the period of RTCP [12], the non-BORC client sends out the feedback control while it receives every 500 packets or more. It usually takes 2 to 3 seconds to send out the control message. One of BORC clients and the non-BORC client added the function of multiplicative decreasing for the TCP-friendliness while congestion condition.

BORC clients check for their buffer-occupancies after receiving every 100 packets. Considering the overhead of the feedback control mechanism including communication and computation overheads at the client, one of the BORC clients would send out feedback control to the server only when the buffer runs either above the high threshold \( b_h \) or below the low watermark \( b_l \). The other BORC client would send out the feedback control while receiving every 100 packets.

Note that the system dynamic may be changed at any time. For fair comparisons, we need to make sure the system dynamics applied in different tests are the same. In our experiments, we first assume that the same content of media streams are requested to serve at the same time for the clients. The media content applied is MPEG-1 file of the motion picture “Star War, Episode I”. We take 20 minutes length of the movie with size 206577K bytes in the experiments. The packet size segmented for delivery is 1K byte. The arrival packet will be accommodated in a buffer unit at client. In the testing period, we recorded the variations of the sending rate and the running buffer-occupancy according the different deployed rate control functions. The maximum sending rate of the VOD server is 183024 bytes/s (it is measured from the 1.5 Mbps bandwidth of the ADSL downstream that simulated by the Dummiynet). All the BORC client and non-BORC clients are set-top box machines with the same limited buffer size of \( b \) and \( b \) equals to a very small size of 200 buffer units (i.e. 200K bytes). The low watermark \( b_l \) is \( b/4 \) and the high watermark \( b_h \) is \( 3b/4 \) in our experiments. The middle point of \( b_l \) and \( b_h \) is \( b_m \), and \( b_m \) is also aimed to the middle of buffer. Besides, the distances from overflow and underflow position on the buffer to \( b_l \) and \( b_h \) are the same. In statistics, we assume the chance of overflow and underflow for these thresholds will be the same in the long term without specific rate control.

3.2 Performance results and evaluation

To measure the performance of our proposed control theoretic buffer-occupancy rate adaptation mechanism, we have already defined two terms \( S.M. \) and \( T.M. \). in section 2.1. The experimental result as illustrated in Figure 7 shows that the \( T.M. \) and \( S.M. \) ratios of a BORC and the non-BORC clients experiencing the delays of 0ms, 100ms, 500ms and 1 second. The results indicate that all the \( T.M. \) ratios of BORC client are less than 2.3% and especially much less than the ratios in non-BORC client. It indicate also that our proposed rate control scheme with buffer-occupancy feedback outperform [4] in controlling buffer occupancy to within the designated size constraints.

Besides, the playback quality index \( S.M. \) of BORC approaches to the ratio of non-BORC while the delay is longer. After studying Figure 8 and Table 1, buffering underflow dominates in our performance study. We also observed that most the buffer underflow occurs at the beginning of streaming. Unfortunately, the feedback control mechanism is not yet stable for the preloading phase. Non-BORC client is also unstable at that time. However, non-BORC client suffers from much more overflow than does BORC clients do as shown in the experiments. Therefore, the BORC client outperforms the non-BORC client under all the measured criterions proposed in the experiments.

<table>
<thead>
<tr>
<th>Delay (ms)</th>
<th>0ms</th>
<th>100ms</th>
<th>500ms</th>
<th>1000ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-BORC</td>
<td>4.2</td>
<td>1147</td>
<td>614</td>
<td>3947</td>
</tr>
<tr>
<td>BORC</td>
<td>2800</td>
<td>2572</td>
<td>2668</td>
<td>2539</td>
</tr>
</tbody>
</table>

Table 1. Packet counts for overflow and underflow.

Moreover, as we studied the trace of buffer-occupancy and rate changes illustrated in the Figure 9-10, we may conclude that buffer-occupancy is controllable even under the long network delay with less communication overhead of feedback control message than the other without buffer-occupancy feedback control. The experimental results indicate that our predictive rate control mechanism of buffer-occupancy feedback control
maintains low T.M. ratio to prevent the degradation of playback QoS for resultant serious miss ratio (i.e. S.M.).

Figure 9. Trace of buffer-occupancy for 0 to 10ms network delay.

Figure 10. Trace of buffer-occupancy for 500ms to 1 sec network delay.

4. ANALYSIS OF IMPACT OF SYSTEM DYNAMICS TO BUFFER-OCCUPANCY

In a network system, the sending rate \( s(t) \) can be defined as
\[
s(t) = \lambda(t) + q(t)
\]
where \( q(t) \) represents the jitters or the packet loss. By applying the same idea, we can calculate the adaptation of sending rate \( s(t+1) = s(t) + \Delta r(t) + \Delta q(t) \) where \( \Delta r(t) \) is the picked up control function. For example, \( \Delta q(t) = -(s(t) - \lambda(t))^\varepsilon \) where \( 0 < \varepsilon < 1 \) is a small constant.

Buffering the data of media stream may absorb the variation of playback rate, network bandwidth and end-to-end delay jitter. The sending rate \( s(t) \) is ideally calculated by the summation of \( s(t-1) \) and two delta functions obtained (i.e. \( \Delta r(t-1) \) and \( \Delta q(t-1) \)) from the proposed control functions. But in the experiments for the control theoretic buffer-occupancy feedback control mechanism, the \( \Delta q(t) \) is considered to be zero for the assumption of the network noise \( q(t) \) remains invariant most of the time in the formula \( v(t) = \lambda(t) + q(t) \). In the following, we will show the reason why the \( \Delta q(t) \) can be discarded in the proposed feedback rate control algorithm in the experiments.

\[
\begin{align*}
\therefore r(t) &= \lambda(t) - \Delta t - \mu(t - \Delta t) - \mu(t - \Delta t) \\
\therefore r(t + \Delta t) &= s(t) - q(t) - \mu(t) \\
\therefore \Delta r &= r(t + \Delta t) - r(t) \\
&= s(t) - s(t - \Delta t) - (q(t) - q(t - \Delta t)) - (\mu(t) - \mu(t - \Delta t)) \\
&= \Delta s - \Delta q - \Delta \mu \\
\Delta b &= \Delta s \Delta r = (\Delta s - \Delta q - \Delta \mu) \Delta t
\end{align*}
\]

Therefore, the control function for \( \Delta q(t) \) can be approximately associated into a single \( \Delta r(t) \) control function for the coming instance of buffer-occupancy \( b(t + \Delta t) \) will partially reflect the accumulated network noise \( \Delta s \Delta r(t) \) by the difference \( \Delta b(t) \).

5. CONCLUSIONS AND FUTURE WORKS

Our experimental results show that the proposed PD control function efficiently control the occupancy of client buffer of small size. It is shown to perform well even for networks with large network delay. Using the scheme of predictive prevention from overflow and underflow of client buffer, a significant improvement of playback QoS can be achieved. It indicates that buffer occupancy should be taken into account in TCP-friendly rate control functions if the client buffer is small. For example, this buffer occupancy information can be placed in the optional field of RTCP packets. We will study further on the implementation of this extension to transport video over wireless network using RTP/CRTP (Compressed RTP [13], revised version of RTP/RTCP for multimedia transportation over low-speed serial links like wireless network).

REFERENCES