OPTIMAL BANDWIDTH ALLOCATION SCHEME WITH DELAY AWARENESS IN MULTIMEDIA TRANSMISSION

Mei-Ling Shyu[1], Shu-Ching Chen[2], Hongli Luo[1]

[1]Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124, USA

[2]Distributed Multimedia Information System Laboratory, School of Computer Science Florida International University, Miami, FL 33199, USA

ABSTRACT

Recently, efficient network resource management and quality-of-service (QoS) guarantee become more and more important for multimedia applications and services, especially when considering network delays. In this paper, an optimal bandwidth allocation scheme that achieves maximal utilization of the client buffer and minimal allocation of bandwidth for each client is introduced. The proposed scheme allocates bandwidth to multiple clients requesting services from a server by adjusting the transmission rates based on the client buffer occupancy, the playback requirements of the individual client and the network delays. Simulations for the single client and multiple client scenarios are conducted under different network congestion levels. The simulation results show that our approach performs better in comparison with the fixed rate allocation approach and the rate by playback requirement approach since it avoids underflows and overflows efficiently and provides QoS for more clients with the limited available bandwidth in the network.

1. INTRODUCTION

There is a growing demand to provide a wide range of multimedia services to the clients with guaranteed quality-of-service (QoS), for example, the Internet. Multimedia services include data, voice, video and graphics, each with its own traffic characteristics and distinct QoS requirements (e.g., packet loss rate, delay, jitter, etc.).

Lots of research has been done on resource allocation schemes in transmission. In [7], the resource reservation algorithms were studied. Though bandwidth reservation can provide QoS, it is time consuming and cannot optimally utilize the bandwidth. Adaptive allocation of network resources is more desirable for QoS provision. Several research works concentrated on adaptive rate control for Internet video transmission and statistics methods were used. In [3], the authors introduced an algorithm for the evaluation of the traffic parameters that statistically characterize the video stream when a given QoS is required. Bandwidth can be dynamically renegotiated according to changes in the bit stream level statistics [2]. In [8], their approach used a neural network method and concentrated on the video content and traffic statistics analysis. In their approach, the resource request method depends on the prediction of future traffic patterns using the content and traffic information of short video segments. In [1], the artificial intelligence techniques such as fuzzy-logic-based and artificial-neural-network-based techniques were used in traffic control mechanisms.

In our previous work [4][5], we proposed a closed-loop framework for multimedia transmission and investigated the single-server/single-client situation. In [6], the single-server/multiple-clients multimedia network transmission model assumes the network delays are negligible. In this paper, an optimal bandwidth allocation scheme that dynamically changes the transmission rates for multiple clients considering the relationships among the transmission rates, client buffer occupancies (i.e., total sizes of packets in the client buffers), playback rates, and network delays is introduced. The simulation results show that our approach better utilizes the client buffer and supports more clients under the constraint of the limited available bandwidth in both the single client and multiple client scenarios in comparison with the fixed rate approach and the rate by playback requirements approach.

The paper is organized as follows. In the next section, the optimal delay awareness bandwidth allocation scheme for both single client and multiple clients are presented. Simulation results for the various scenarios are given in Section 3. Conclusions are presented in Section 4.

2. OPTIMAL DELAY AWARENESS BANDWIDTH ALLOCATION SCHEME

Assume there is an end-to-end transmission from the server to a client. Normally, the delay in the Internet is time varying and hard to be precisely modeled. In this paper, we attempt to capture the delay from the server to client in terms of the percentage of packets sending at the server at a time interval while arriving at the client at a later time interval. Such information is provided as feedback to the server regularly for rate control so that the server adaptively adjusts its transmission rates according to the current network delay.

2.1. Optimal Rate Control Algorithm for a Single Client

We enumerate below all the variables that are necessary for our optimal delay awareness rate control algorithm. For simplicity and easy-to-understand purposes, we describe the algorithm using the single-server/single-client situation here. The optimal

bandwidth allocation for multiple clients is presented in the next subsection.

Let Q_r be the allocated buffer size for each client at the setup of the connection. At time interval k, let R_k be the total size of packets transmitted from the server, P_k be the total size of packets arriving at the client buffer, and L_k be the total size of packets used for playback. In addition, let Q_k and Q_{k+1} denote the client buffer occupancy at the beginning of time intervals k and $k+1$, respectively. Hence,

$$Q_{k+1} = Q_k + P_k - L_k . \quad (1)$$

Due to the network delay, P_k is not equal to R_k. The network delay considered here is the delay that a packet experienced before it arrives at the client, and is the total of transmission delay, queue delay and propagation delay. It is practical to assume that the packets arriving at the client buffer at time interval k comprise of packets transmitted from the server at the time intervals k-d, k-d-1, …, and k-d-i+1. Thus we have

$$P_k = b_{1,k} R_{k-d} + b_{2,k} R_{k-d-1} + ... + b_{i,k} R_{k-d-i+1} , \quad (2)$$

where the subscript k-d is to denote the closest time interval when the transmitted packet can arrive at the buffer, and k-d-i+1 denotes the farthest time interval when the transmitted packet can arrive at the buffer at time interval k. $b_{i,k}$ is used to denote the corresponding percentage of the packets transmitted at k-d-i+1 that arrive at the client buffer at time interval k. Since each packet transmitted from the server is appended a sequence number and a timestamp of the time it is transmitted, the value of $b_{i,k}$ can be calculated at the client by counting the total size of packets arriving at a certain time interval. Similar to $b_{i,k}$, the value of d can also be obtained at the client by checking the timestamp of the arriving packets.

The introduction of the parameters d and $b_{i,k}$ in our algorithm can effectively capture the changing network delays as follows. First, a larger d value indicates larger network delays since it takes longer for a packet to arrive at the client buffer. Second, a larger i value also means larger network delays since it takes longer for all the packets transmitted at the previous time intervals to arrive at the client buffer. Third, a larger $b_{i,k}$ value means a larger percentage of packets transmitted at time interval k-d-i+1 can arrive.

Optimal utilization of the network resource includes maximal utilization of client buffers and minimal allocation of bandwidth. We need to keep the transmission rate small, and at the same time keep the difference between the buffer occupancy and the allocated buffer size small. The optimization object is to find a suitable sequence of R_k that can minimize the following quadratic performance function.

$$J_k = (w_p Q_{k+d_0} - w_q Q_r)^2 + (w_r R_k)^2 , \quad (3)$$

where w_p, w_q, and w_r are the weighting coefficients. Because of the network delays, the change of transmission rate at time interval k (R_k) will result in a change of the client buffer occupancy at time interval $k+d_0$ (Q_{k+d0}). Therefore, d_0 is introduced in the performance index as a transmission control delay parameter.

Using time domain representation to combine Equations (1) and (2), we have the following equation

$$A(z^{-1}) Q_k = z^{-d_0} B(z^{-1}) R_k - L_k , \quad (4)$$

where $A(z^{-1}) = 1 - z^{-1}$,

$$B(z^{-1}) = b_{1,k} + b_{2,k} z^{-1} + ... + b_{m+1,k} z^{-m} . \quad (5)$$

z^{-1} is the delay operator (i.e., $z^{-1} Q_k = Q_{k-1}$). Here we assume that $b_{1,k} \neq 0$ and d_0 is determined during the computation. When 1 is divided by $A(z^{-1})$, the following quotient and remainder can be obtained.

$$1 = A(z^{-1}) F(z^{-1}) + z^{-d_0} G(z^{-1}) . \quad (6)$$

From Equation (4), we get

$$Q_{k+d_0} = B(z^{-1}) F(z^{-1}) R_k - F(z^{-1}) L_{k+d_0} + G(z^{-1}) Q_k . \quad (7)$$

Use Equation (7) to replace Q_{k+d_0} in the objective function in Equation (3), and differentiate J_k with respect to R_k, then the optimal transmission rate R_k can be obtained as follows.

$$(w_p^2 B(z^{-1}) F(z^{-1}) + \frac{1}{b_{1,k}} w_r^2) R_k = -w_p^2 G(z^{-1}) Q_k + w_p w_q Q_r + w_p^2 F(z^{-1}) L_{k+d_0} \quad (8)$$

The above equation is a recursive equation for R_k represented in terms of R_{k-1}, R_{k-2}, …, Q_k, Q_{k-1}, …, and Q_r. Therefore, the computation complexity needed to calculate the optimal transmission rate is low. The values of Q_k, Q_{k-1}, … are sent from the client to the server as feedback information at a certain time period, e.g., a round trip time (RTT), and then the optimal transmission rate is calculated at the server.

2.2. Optimal Bandwidth Allocation for Multiple Clients

Normally, there are a large number of active clients requesting services from a server, and the server needs to allocate its available bandwidth to provide QoS to a large population of clients. If the server allocates each client a fixed bandwidth or the bandwidth only according to the playback requirement during the connection, it will be a great waste of the network bandwidth since each client has its own traffic requirement. In order to provide services to the maximal number of clients with their QoS requirements and to achieve high utilization of the bandwidth resource, the bandwidth allocated to each client needs to be minimized.

Assume there are m active clients simultaneously requesting services from the server. Let $J_{j,k}$, $e_{j,k}$, $R_{j,k}$, $Q_{j,k+d_0}$ and $Q_{j,r}$ be the corresponding values as defined in the previous subsection for the j^{th} client. When there are multiple clients, the optimization function for resource allocation of the server becomes the following:

$$J = \sum_{j=1}^{m} J_{j,k} = \sum_{j=1}^{m} (e_{j,k}^2 + w_r^2 R_{j,k}^2) , \quad (9)$$

where $e_{j,k} = w_p Q_{j,k+d_0} - w_q Q_{j,r}$. (10)

Since the m clients are independent, if the $J_{j,k}$ value of the j^{th} client is minimal, the sum of the $J_{j,k}$ values for all m clients (i.e., the J value) is also the minimal. Hence, the optimal transmission rate for each client that achieves the optimization of the performance index in Equation (9) can be obtained.

Our bandwidth allocation mechanism is used with connection admission control (CAC). When the total optimal bandwidth requirement is larger than the available bandwidth, a bandwidth reallocation is required. To provide fairness among all clients, the bandwidth reallocated should be proportional to the actual requirement of each client under the constraint of the available bandwidth.

3. SIMULATION RESULTS

To study the performance of our approach, we compare our approach with other bandwidth allocation mechanisms such as the fixed rate approach and the rate by playback requirement approach under the single client and multiple client scenarios. For the fixed rate transmission approach, each client is allocated with a constant bandwidth. For the rate by playback requirement approach, the server allocates the bandwidth to the clients according to their playback requirements. In our simulations, it is assumed that the buffer at each client is 2 MB (Mbytes), the available bandwidth is 5 MBps (Mbytes/second), and the playback rates are generated randomly between [$0.1x10^5$, $0.9x10^5$] Bps. The simulations were run in 10,000 time intervals with the increment of 1 time interval.

3.1. Single Client Scenario

The scenario of one client requesting data from the server is chosen to illustrate how the transmission rate is adjusted in our approach. The playback rate requirement is between $0.1x10^5$ Bps and $0.3x10^5$ Bps. Figure 1 gives the comparison between our approach (the solid lines) and the rate by playback requirement approach (the dashed lines) by showing how the transmission rate is adjusted according to the playback rate and client buffer occupancy.

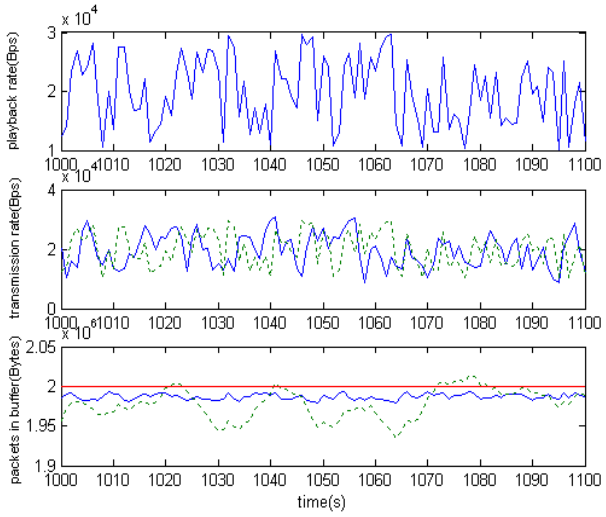

Figure 1. Comparion of our approach (in solid lines) and the rate by playback requirement approach (in dashed lines) in transmission rates and buffer occupancies.

When the client buffer occupancy is above the horizontal solid line (denoting the allocated maximal buffer size), overflow occurs. For example, in Figure 1, between 1020^{th} and 1025^{th} time intervals, overflows occur in the rate by playback requirement approach, while there is no overflow in our approach. It can also be easily seen from this figure that generally the buffer utilization of our approach is better than that of the rate by playback requirement approach. In addition, the changes of the transmission rates and the client buffer occupancies in our approach are less drastic than those in the rate by playback requirement approach. Hence, our approach is more robust in adaptive bandwidth allocation.

Our approach is also compared with the fixed rate approach that the bandwidth is allocated according to the peak playback requirements ($0.3x10^5$ Bps). The client buffer occupancies are continually increasing so that overflow occurs and it cannot be recovered.

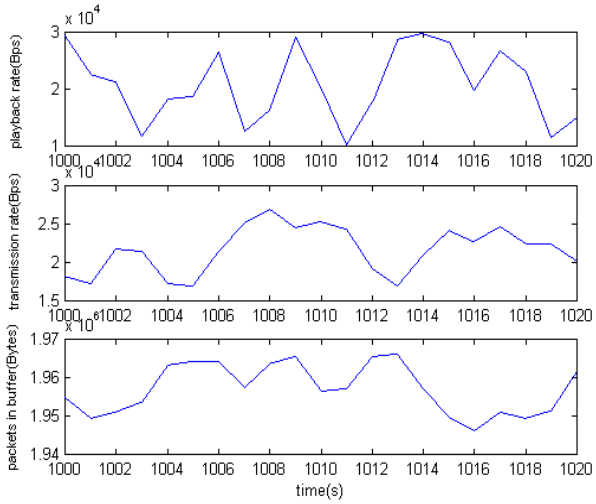

Figure 2. Transmission rate changes with the client buffer occupancies and the playback rates in a single client connection during time intervals [1000,1020] in our approach.

To better illustrate how the transmission rate is adjusted in our approach, the transmission rate changes during time intervals [1000, 1020] are presented in Figure 2. For example, at the 1007^{th} time interval, though the playback rate is low and buffer occupancy is medium, the transmission rate is high. This is because there is a peak playback requirement at 1014^{th} time interval. Considering there is a network delay, it takes a while for the transmitted packets to arrive at the client buffer. Similarly, as can be observed, the playback rate at 1008^{th} time interval is low but the transmission rate at that time interval is high since the playback rate at 1015^{th} is also high. Transmission rates at 1007^{th} and 1008^{th} time intervals should be high enough so that the packets arriving at the client buffer at 1014^{th} and 1015^{th} time intervals can satisfy the playback requirements in the next time interval. It means that the transmission rate should be adjusted beforehand. From the simulation result, it can be seen that our approach has the capability of predicting the necessary packets at the buffer so that the server can transmit beforehand enough packets to provide for the future use.

3.2. Multiple Client Scenario

We investigate the multiple client scenarios to compare the performance of our approach (approach A) with the other approaches. Three different scopes of playback requirements

are used to simulate three levels of network congestion. That is, ($0.1x10^5$ Bps ~ $0.3x10^5$ Bps), ($0.4x10^5$ Bps ~ $0.6x10^5$ Bps) and ($0.7x10^5$ Bps ~ $0.9x10^5$ Bps) are used to simulate the less, medium and severe congested network scenarios.

For the comparison, in the rate by playback requirement approach (approach B), when the total bandwidth requirement of the *m* clients is larger than the bandwidth at the server, we adjust the rates and distribute the bandwidth fairly to the clients according to the playback requirements. In the fixed rate approach (approach C), the fixed rate is obtained by equally allocating the bandwidth to all of the clients. The maximal number of clients that can be supported by our approach and the rate by playback requirement approach under the same bandwidth constraint is shown in Table 1.

Table 1. Maximal numbers of clients supported.

	Less Congested	Medium Congested	Severe Congested
Approach A	250	100	62
Approach B	183	87	57

As can be seen from Table 1, for the less congested network scenario, the maximal numbers of clients that can be supported are 250 and 183 in our approach (approach A) and the rate by playback requirement approach (approach B), respectively. Similarly, the maximal numbers are 100 and 87 for the medium congested scenario, and 62 and 57 for the severe congested network scenario. Hence, our approach can support more clients than the rate by playback requirement approach.

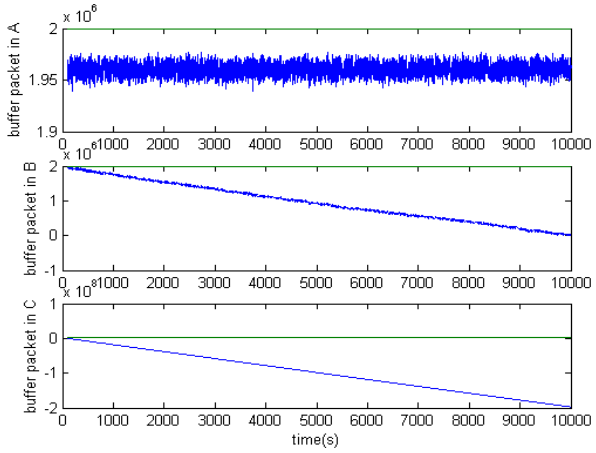

Figure 3. Client buffer occupancies of three approaches under the less congested network condition with 184 clients.

In Figure 3, the overflow/underflow situations for the three approaches with 184 clients in the less congested network scenario are presented. When the number of clients is 183, there are no underflow and overflow in the buffers in approach A, but there are overflows in approach B and the buffer occupancy tends to decrease. In Figure 3, when there are 184 clients requesting services from the server, under the bandwidth limitation, the buffer occupancy in approach B is decreasing to below zero (i.e., underflow occurs). Hence, the maximal number of clients the server can support is 183 in approach B. It is also obvious to observe that approach C behaves badly from this figure. Moreover, the client buffer occupancies in approach A are normally more than those in approach B, which means that our approach can better utilize the buffer capacity than approach B.

4. CONCLUSIONS

In this paper, an optimal delay awareness bandwidth allocation scheme that achieves the maximal utilization of client buffer and the minimal allocation of bandwidth is presented. In our approach, the transmission rate for each client is determined dynamically based on the playback requirement, buffer occupancy, and changing network delays. We studied both the single client and multiple client scenarios to illustrate how our approach performs. Simulations were run under different network congestion levels to show the efficiency of our approach in admitting more clients in multiple client scenarios. The optimized transmission rate can be adjusted to satisfy the constraint of the network bandwidth and to avoid overflows and underflows. Comparisons are made with the fixed rate approach and the rate by playback requirement approach. From the simulation results, it can be easily seen that overflows and underflows occur frequently in the fixed rate approach and they are unrecoverable. In addition, our approach is efficient in supporting QoS to more clients under the constraint of the bandwidth limitation in comparison with the other two approaches. In other words, our approach is more robust in avoiding overflows and underflows, and adapting to the changing network delays.

5. REFERENCES

[1] C. Douligeris and G. Develekos, "Neuro-Fuzzy Control in ATM Networks," *IEEE Commun. Mag.*, pp. 154-161, May 1997.

[2] M. R. Izquierdo and D. S. Reeves, "A Survey of Statistical Source Models for Variable Bit-rate Compressed Video," *Multimedia System*, vol. 7, no. 3, pp. 199-213, 1999.

[3] A. Lombardo, G. Schembra, and G. Morabito, "Traffic Specifications for the Transmission of Stored MPEG Video on the Internet," *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 5-16, 2001.

[4] M.-L. Shyu, S.-C. Chen, and H. Luo, "Optimal Resource Utilization in Multimedia Transmission," *IEEE International Conference on Multimedia and Expo (ICME)*, Waseda University, Tokyo, Japan, pp. 880-883, August 22-25, 2001.

[5] M.-L. Shyu, S.-C. Chen and H. Luo, "An Adaptive Optimal Multimedia Network Transmission Control Scheme," *Second IEEE Pacific-Rim Conference on Multimedia 2001 (PCM'2001)*, Beijing, China, pp. 1042-1047, October 24-26, 2001.

[6] M.-L. Shyu, S.-C. Chen, and H. Luo, "Self-Adjusted Network Transmission for Multimedia Data," *Third IEEE Conference on Information Technology: Coding and Computing (ITCC-2002)*, Las Vegas, Nevada, USA, pp. 128-133, April 8-10, 2002.

[7] P. P. White, "RSVP and Integrated Services in the Internet: A Tutorial," *IEEE Commun. Mag.*, vol. 35, no. 5, pp. 100-106, May 1997.

[8] M. Wu, R. A. Joyce, H. Wong, L. Guan, and S. Kung, "Dynamic Resource Allocation via Video Content and Short-Term Traffic Statistics," *IEEE Transaction on Multimedia*, vol. 3, no. 2, pp. 186-199, June, 2001.