An Adaptive Advance Reservation Algorithm for QoS Sensitive Multimedia Networks

Sungwook Kim, and Pramod K. Varshney, Fellow, IEEE
skim@ecs.syr.edu and varshney@syr.edu
Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NY 13244, USA

ABSTRACT

Next generation communication networks will be required to support many multimedia services including voice, video and data. Since, bandwidth is an extremely valuable and scarce resource in multimedia networks, efficient bandwidth management is necessary in order to provide high quality service to users. Some applications need bandwidth reservations at an earlier time so that desired QoS requirement could be ensured in the future. In this paper, we propose a new bandwidth management algorithm for advance reservation. The main design principle underlying our algorithm is adaptive online control based on real time estimation. This online approach is dynamic and flexible that responds to current network conditions. Therefore, our scheme is able to resolve conflicting QoS requirements while ensuring efficient network performance. With a simulation study, we demonstrate the excellent performance of our scheme under widely diverse traffic load intensities.

Keywords: Multimedia networks, On-line decisions, Bandwidth reservation, Adaptive bandwidth allocation, Preemption, Quality of Service.

1. INTRODUCTION

In recent years, the requirements placed on communication networks have increased dramatically. The explosive growth of new communication services over the Internet and the rapid and widespread proliferation of multimedia data have necessitated the development of an efficient network management system. The management system is expected to provide diversified traffic services and enhance network performance simultaneously [1,2,3]. Multimedia is a keyword in the evolving information age of the 21st century. Usually heterogeneous multimedia data can be categorized into two classes according to the required Quality of Service (QoS): class I (real-time) traffic services and class II (non real-time) traffic services [4,5].

Sometimes, real-time applications need reservations at an earlier time so that desired QoS requirement could be guaranteed in the future. Our proposed scheme supports an advance reservation by an extension of the common bandwidth reservation concept. Bandwidth reservation is an effective technique for enhancing network QoS. To handle advance reservation requests at a future time, we split the available bandwidth into different partitions by means of a moving boundary, which means that the partition boundary can be changed adaptively.
The fundamental problem of management algorithms is uncertainty regarding the future. Control decisions for network management have to be made in real time, and without the knowledge of future traffic requests at the decision time. This uncertainty makes the problem more challenging. Therefore, online algorithms [6] are natural candidates for the design of efficient bandwidth control schemes in QoS sensitive multimedia networks.

Recently, Schelén and Pink [7] proposed the resource advance reservation (RAR) scheme for advance reservations. The main objectives of this scheme are similar to our objectives in this paper. However, this existing scheme has several shortcomings as described in Section 3. Compared to this scheme [7], our proposed online bandwidth management scheme attains better performance for QoS sensitive multimedia services. Our attempt in this paper is to approximate the optimal network performance by adaptive online controls. In addition, to balance the network performance among conflicting QoS parameters, our online decisions are mutually dependent on each other under widely different and diversified network traffic situations. The most important contributions of our proposed scheme are (i) ability to adaptively allocate bandwidth to maximize network performance, (ii) a well-balanced network performance among contradictory QoS requirements, (iii) low complexity of decision mechanisms to make it practical for real network operations, (iv) ability for advance reservation for multimedia networks.

This paper is organized as follows. Section 2 describes the proposed scheme in detail. In Section 3, performance evaluation results are presented along with comparisons with the scheme proposed in [7]. Finally, concluding remarks are given in Section 4.

2. PROPOSED ADVANCE RESERVATION SCHEME

In our proposed advance reservation scheme, call requests are classified into three different categories – immediate calls, advance reservation request calls and accepted advance reservation calls. An immediate call is serviced immediately. Advance reservation call requests will be serviced in the future if the call request is accepted. For each call request, we assign different call priorities. Advance reservation requests are usually less urgent than immediate call requests, but, if an advance reservation request is accepted, this reserved call has higher priority than immediate call requests. Therefore, we give higher priority to accepted advance reservation calls than immediate call requests. Advance reservation requests have lower priority than immediate call requests.

Our advance reservation algorithm splits the available bandwidth by the use of a moving boundary. Moving boundary [8] means that the partition boundary can be changed adaptively to avoid bandwidth inefficiency, which is a strategy that provides a compromise between complete sharing and a fixed partitioning policy. We partition the available bandwidth into three parts: current_partition is for new incoming immediate call requests, advance_partition is assigned to advance reservation call requests and saving_partition is for approaching accepted advance reservations. To find the adaptive amount of bandwidth for each partition, we dynamically adjust these partition amounts based on current network conditions. This control decision has to be made in real time, and without the knowledge of future traffic requests at the decision time.

Each type of call request has its own bandwidth partition. When its own partition is saturated, the partition boundary can be moved. By bandwidth borrowing from other partitions, we can adaptively move each partition boundary. Bandwidth borrowing decisions are made based on the predefined target CBP (\( P_{\text{CBP}} \)), priority of the requested call and the amount of accepted advance reservations. To ensure high bandwidth utilization while handling advance reservation, we propose the following control rules for moving the boundary by bandwidth borrowing. First, bandwidth can be borrowed from advance_partition to current_partition if there is available bandwidth. Second, bandwidth can be borrowed from current_partition to advance_partition, if the following conditions are satisfied. Required bandwidth is available in current_partition and the current CBP is lower than \( P_{\text{CBP}} \). If any one of the conditions is not satisfied, bandwidth borrowing is not allowed. Third, the reserved bandwidth in saving_partition can be borrowed temporarily for adaptive allocation of existing class II (non real-time) call services by bandwidth adaptation strategy [9,10].
Based on our bandwidth partitioning approach along with the moving boundary rules, admission decisions for immediate and advance reservation requests are mutually dependent on each other. When an immediate call request arrives, call admission is granted if sufficient bandwidth in the current_partition is available. If the current_partition is full, instead of rejecting the request, our scheme tries to borrow the required bandwidth based on moving boundary rules. If it fails to borrow the required bandwidth, the new immediate call request is rejected. When an advance reservation request arrives, call admission is granted if the available bandwidth in advance_partition is sufficient. If, during the requested future period, the advance_partition is full, our scheme tries to borrow the required bandwidth based on the moving boundary rules. If it fails to borrow the bandwidth, the new advance reservation request is rejected.

To allow for advance reservations, the future time axis is divided into intervals in terms of unit_time. In our scheme, we define a future point on the time axis as the saving_start time. We also define the time interval between the saving_start time and the actual starting time of the approaching advance reservation as the saving_period. During saving_period, our scheme actually starts making bandwidth preparation in the saving_partition for accepted advance reservations. However, under a heavy traffic load situation, there is a risk that the available bandwidth during the saving_period is not enough for the accepted advance reservations. If this is the case, it would be beneficial to preempt one or more existing lower priority calls to satisfy the higher priority calls.

The preemption probability can be controlled by means of saving_start time. It is clear that the preemption probability can be reduced by an increase in the saving_period. Therefore, there is a trade off between preemption probability, bandwidth utilization and immediate call blocking probability. For higher bandwidth utilization and low blocking and preemption probabilities, setting of the saving_start time is controlled adaptively based on current network conditions. Therefore, it is also an online computation problem controlled by real time measurements. For adaptive online adjustment, our algorithm provides a coordination paradigm by employing two different components - event driven and time driven components.

2.1 Time Driven Component for saving_start time Decision

In the time driven component for the saving_start time decision, we define a traffic window to decide this value. This window keeps the traffic history of call termination (W_{\text{return}}). By using this window, we can estimate the returning bandwidth, which is released bandwidth per unit_time Ret_B, which is obtained as

\[
Ret_B = \sum_{j \in W_{\text{return}}} (B_j \times N_j) / U
\]

where \(N_j, B_j\) are the number of returned calls and the corresponding bandwidth respectively. \(U\) is the length of the traffic window in terms of unit_time. Based on \(Ret_B\), we can adjust the saving_start time and the saving_period adaptively to ensure sufficient bandwidth for accepted advance reservation calls. Every unit_time, \(Ret_B\) value is dynamically adjusted in an on-line manner.

2.2 Event Driven Component for saving_start time Decision

The event driven component is triggered by a call preemption. When a call preemption occurs, if the current call preemption probability is larger than its target preemption probability (\(P_{\text{cpp}}\)), the saving_start time is made closer to the current time (the saving_period size is increased) by unit_time. If the following conditions are satisfied, the saving_start time is moved away from current time (the saving_period size is decreased) by unit_time. First, bandwidth is successfully allocated for advance reservation. Second, the current preemption probability is lower than \(P_{\text{cpp}}\). Third, the size of the saving_period is larger than the obtained size through the time driven component. Therefore, as network
conditions change, our proposed advance reservation algorithm dynamically adapts the size of \textit{saving\_period} reflecting the current traffic conditions.

Fig. 1 illustrates admitted immediate and accepted advanced reservation calls in a time-bandwidth diagram. Bars a, b, c, d and e are bandwidth allocations for accepted immediate calls. The total allocated bandwidth for immediate calls can be reduced to minimum allocation by the bandwidth adaptation technique [9,10]. The reducible bandwidth amount of existing calls is the size of the \textit{advance\_partition}. Bars f, g and h are accepted bandwidth allocation for advance reservations. Our scheme begins to set aside available bandwidth in \textit{saving\_partition} for e,f and g calls during \textit{saving\_period}. The main steps of our proposed advance reservation algorithm are given next.

- At the initial time, all the available bandwidth is in \textit{current\_partition}.
- When an immediate call request arrives, call admission is granted if bandwidth in \textit{current\_partition} is available.
- If the \textit{current\_partition} is full, our scheme tries to borrow required bandwidth from \textit{advance\_partition}. If it fails to borrow bandwidth, a new immediate call request is rejected.
- When an advance reservation request arrives, call admission is granted if bandwidth in \textit{advance\_partition} is available. If the \textit{advance\_partition} is full, our scheme also tries to borrow required bandwidth from \textit{current\_partition}.
- If the borrowing conditions based on moving boundary rules are satisfied, bandwidth borrowing is carried out. If it fails to borrow bandwidth, the advance reservation request is rejected.
- Bandwidth in \textit{saving\_partition} can be allocated temporarily only for existing calls.
- Using the traffic window, \( R_{ret} \) is dynamically estimated every \textit{unit\_time}. By using this value, we can decide the \textit{saving\_start} time and the \textit{saving\_period}.
- When call preemption occurs or bandwidth is successfully allocated for advance reservation, the \textit{saving\_start} time and \textit{saving\_period} size are adjusted by \textit{unit\_time}.
3. PERFORMANCE EVALUATION

To evaluate the performance of our proposed online scheme, we employ a simulation model. Based on this simulation model, we compare the performance of our scheme with an existing scheme [7]. The assumptions for our simulation study are as follows.

- The arrival process for new call requests is Poisson with rate $\lambda$ (calls/s), and the range of offered load was varied from 0 to 3.0.
- The total capacity of network bandwidth is 30Mbps.
- Network performance measures are plotted as a function of the offered load per second (calls/s). Based on this assumption, unit time in our simulation model is one second.
- In order to represent various multimedia data call requests, eight different traffic types are assumed based on connection duration, bandwidth requirement and required QoS. They are generated with equal probability.
- The durations of call connections and the start times of advance reservation requests, i.e., how far in advance from the current time, are exponentially distributed with different means for different multimedia data types.

Performance measures obtained through simulation are bandwidth utilization, call blocking probability (CBP) of new calls, call preemption and completion probability (CPP and CCP), and network revenue. These performance measures obtained on the basis of 10 simulation runs are plotted as a function of the offered load (call arrival rate) $\lambda$. Table 1 shows the multimedia traffic types and system parameters used in the simulation. With multiple classes of traffic, each has different traffic characteristics - its own requirements in terms of bandwidth, QoS guarantee and call connection time.

![Table 1. Multimedia traffic and system parameters used in the simulation experiment](image)

We compare the performance of our proposed online bandwidth management scheme with an existing scheme - the Resource Advance Reservation (RAR) scheme [7]. The RAR scheme provides an advance reservation mechanism in multimedia network operations. This scheme shares available bandwidth to support both immediate calls and advance reservations. When an immediate call is overlapped in time with advance reservation, the immediate call is preempted. The RAR scheme tries to maximize bandwidth utilization while providing advance reservation technique, however, there...
are several disadvantages. Under heavy traffic load situations, sharing bandwidth by immediate calls and advance reservations without partitioning can cause higher immediate call blocking. In addition, this scheme is not dynamically adaptable for various network traffic load intensities. With the RAR scheme, the bandwidth sharing techniques [10] can be applied to support advance reservations in multimedia networks. There are two well-known bandwidth sharing schemes; the complete bandwidth sharing scheme (shared scheme) and the fixed bandwidth partitioning scheme (fixed scheme). In the shared scheme, all traffic classes share the entire bandwidth. In the fixed scheme, bandwidth is divided into distinct parts corresponding to particular traffic classes. To confirm the effectiveness of our scheme, we also compare the performance with these two schemes.

The curves in (a) of Fig.2 show the performance comparison in terms of network revenue. As the call arrival rate increases, the performance of our scheme becomes much higher than other schemes. The curves in (b) and (c) of Fig.2 show the performance comparison of CBP and CCP for all types of traffic services. Our scheme has better performance than other schemes and quite adaptable from low to high network traffic intensities. This feature is highly desirable under widely different traffic situations. The curves in (d) and (e) of Fig.2 show the CBP and CCP of immediate call requests. The simulation results show similar trends in the curves in (b) and (c) of Fig.2. The curves in (f) of Fig.2 show the call preemption probability (CPP). From the simulation results we obtained, it is clear that our online scheme, in general, achieves superior network performance for varying traffic load intensities in multimedia networks while other existing schemes cannot offer such an attractive trade off.
4. SUMMARY

In order to provide advance reservation in QoS sensitive multimedia networks, an adaptive online bandwidth management scheme is proposed. An important design principle underlying our scheme is adaptive online control based on real time estimation. Our approach is able to resolve conflicting QoS performance criteria while ensuring efficient network performance. In addition, proposed online management has low complexity making it practical for multimedia networks. Performance evaluation results clearly indicate that our scheme maintains well-balanced network performance in widely different and diversified traffic load situations while other schemes can not offer such an attractive trade off. The main concept of our online scheme can also be applied beyond bandwidth management in multimedia networks to other resource management algorithms, e.g., CPU scheduling, disk and memory management, etc.

REFERENCES


