Admission Control and Capacity Management for Advance Reservations with Uncertain Service Duration

Yeali S. Sun[1], Yung-Cheng Tu[2], and Meng Chang Chen[3]

[1] Dept. of Information Management
National Taiwan University
Taipei, Taiwan
[2] Dept. of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
[3] Institute of Information Science
Academia Sinica
Taipei, Taiwan

Abstract. Different from Immediate Request (IR) service in packet-switched networks, admission control for Advance Reservation (AR) service is more complex - the decision points include not only the start time of the new connection, but also the instants that the new connection overlaps with connections already admitted in the system. Traditional approach on advance reservation considers only a fixed scheduled period. When overtime occurs (often quite approaching the end of the originally scheduled service period) depending on network load and resource usage, the service may easily be disrupted due to insufficient resources available. Examples include the broadcasting of sports events and business video-conference calls. In this paper, we study the problem of admission control and resource management for AR service with uncertain service duration. The objective is to maximize user satisfaction in terms of service continuity and guarantee of QoS while minimizing reservation cost and call blocking probability of the AR service. An innovative two-leg admission control and bandwidth management scheme is proposed. Service continuity, user utility and reservation cost functions are proposed here to evaluate user's satisfaction and the efficiency of resource allocation. Simulation results are presented.

1 Introduction

Many signaling and admission control designs of the quality of service (QoS) support in packet-switched networks such as RSVP [1] focus on requests that must be served immediately, commonly known as Immediate Request (IR) service. In today's Internet, there is demand for Advance Reservation (AR) service. For example, many important business conference meetings and calls are pre-planned and scheduled. By advance reservation service, users can know whether they can

E. Gregori et al. (Eds.): NETWORKING 2002, LNCS 2345, pp. 190–201, 2002.

get full QoS support of their communication needs over the Internet in advance. From the service provider's perspective, knowing the future needs ahead allows them to better manage the allocation and sharing of network resources between users, and to serve their customers in a more affirmative, predictable way.

In order to perform admission control and resource allocation, requests for advance reservation must specify three basic data: service start time, QoS requirement and duration of service. Recently a few works were proposed. They all assume these parameters are given and of fixed value when requests are submitted [2,3,4,5,6,7,8]. In reality, these information may not be known in prior, especially the service duration. Examples include the broadcasting of sports events and business videoconference calls. Typically, there is so-called scheduled duration, e.g., two hours for a broadcast sports event. But often there are overtimes. Traditional approach on advance reservation considers only a fixed scheduled period. When overtime occurs (often quite approaching the end of the originally scheduled service period) depending on network load and resource usage, the service may easily be disrupted due to insufficient resources available. Therefore, it becomes a challenge to the service provider to fulfill the needs of such types of requests assuring both the continuity of service and guarantee of QoS given the uncertain service duration at the time the request was scheduled while in line with its goal of maximum network resource utilization.

In this paper, we focus on AR request with longer lifetime such as Internet broadcast events and videoconferences. Here, we propose an innovative two-leg admission control and resource reservation scheme for AR requests with uncertain service duration over the Internet. The idea is to perform bandwidth reservation in *multiple* stages. Each stage has a fixed duration and specific level of quality of service to assure. Thus, service provider can efficiently manage network resources and allocate bandwidth necessary to guarantee service quality requirements of individual connections in each stage.

To further tackle uncertainty and to maximize network resource utilization, an update mechanism is used. A convex user utility function is defined to characterize the level of user satisfaction for those admitted AR connections with the combined bandwidth allocation and service continuity. A reservation cost is also defined to evaluate the efficiency of the overall network resource allocation in advance reservation service.

Other works related to advance reservations in the past include extensions to the existing protocols and signaling capabilities, e.g., extension of ST2 protocol [2,9] and RSVP [3]. In [5], the authors proposed a distributed reservation scheme and its possible implementation. In [10], the authors studied AR requests with uncertain duration. It does not address the service continuity problem. In [8], a measurement-based approach is proposed to estimate the bandwidth used for existing connections with fixed duration. In [6,7], they discussed the admission control for connections in progress that are preemptable or interruptible. Specifically, in [6], they studied the issue of resource sharing between AR and IR services. In [7], they gave a general description of the policy and pricing schemes for advance reservations. Most of these works assumed that service durations are fixed and available at the admission control time. In [6,11], they assumed the service times follow some distribution. An estimate or a safe upper bound of the service duration must be given at the request submission time. In this paper, we focus on the admission control and bandwidth allocation problem for AR requests with uncertain service duration.

The organization of this paper is as follows. In Section 2, we present the proposed Two-leg bandwidth reservation scheme. The definitions of service continuity and user utility are given. An update mechanism is also presented. In Section 3, admission control of the proposed scheme is described in detail. In Section 4, reservation cost is presented. In Section 5, simulation results are presented to show the benefits of the proposed scheme. Finally, we give a conclusion in Section 6.

2 The Two-Leg Resource Reservation Scheme

For advance reservation service, there are more than one decision points to check. They include all the time instants the duration of the new connection overlaps with the start time of any connections already admitted in the system. Figure 1 depicts the admission control decision points for AR connections with a fixed duration; the number of decision points is finite. This is, however, not true for the case of uncertain duration.

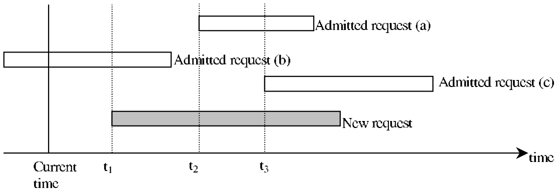

Fig. 1. There may have more than one decision points $\{t_1, t_2, t_3\}$ to consider in the admission control of a new advance reservation request.

In this section, we propose a new admission control scheme with two-leg bandwidth reservation to address the problem of uncertain service duration in AR service. To deal with uncertainty, estimation is used here which is based on the observation that the probability many Internet AR applications will last longer than a duration t is very small when t is sufficiently large (e.g., VCD films information from Blockbuster Homepage [12]). First, we assume for AR requests without specifying service duration, the actual lifetimes will follow certain distributions. Thus, requests can be classified into different categories; each has its own *characteristic lifetime distribution function.* In reality such functions can be obtained through proper data collection, sampling, analysis and characterization from the real world [6].

2.1 Age Function

Let $a_i(t)$ be the probability density function of the nominal duration of type i AR connections and s_i is the start time of the connection. We define the age function of connection i, $A_i(t)$, as the probability that connection will end after time $s_i + t$,

$$A_i(t) = Pr\{duration \geq t\} = \int_i^{\infty} a_i(x)dx, \tag{1}$$

2.2 User Utility

We characterize the level of user satisfaction for those admitted AR connections with the combined bandwidth allocation and service continuity. First, a convex function $s_i(T_i)$ is defined to describe the level of satisfaction in terms of service continuity for connection i which lifetime is T_i and D_i is the nominal duration of the corresponding event, i.e.,

$$s_i(T_i) = \begin{cases} e^{k(\frac{T_i}{D_i}-1)} & \text{if } T_i \leq Di \\ 1 & \text{if } T_i > Di \end{cases} \tag{2}$$

The constant k is used to reflect the weight of such effect. Figure 2 shows the values of (2) under different k's. The larger the k the more utility gain is stressed on the continuity of service especially towards the end of the event. For example, if a live broadcast of a basketball game was initially scheduled for three hours but due to overtimes, the event is in fact three and half hours. The nominal duration is three and half hours. The lifetime of the connection however depends on whether a service extension request is issued, say two hour and 45 minutes after the event. If accepted, T_i is equal to nominal duration; otherwise it is three hours.

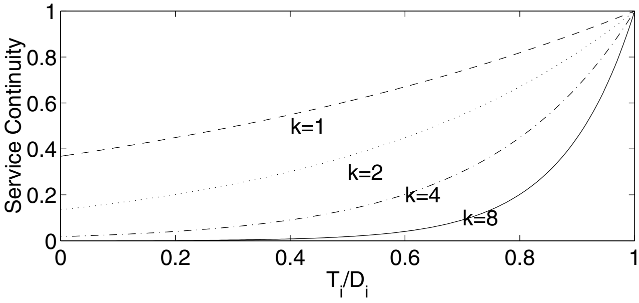

Fig. 2. The values of service continuity under different k's.

Now, we define the *user utility* for connection i as the combination of both service continuity and bandwidth allocation. It is an increasing convex function:

$$u_i(T_i) = \int_0^{T_i} \left(\frac{ds_i(t)}{dt} \times \frac{r_i(t)}{R_i} \right) dt \tag{3}$$

Note that $r_i(t)$ is the bandwidth allocation to connection i at time t and R_i is the requested bandwidth. This function contrasts bandwidth allocated vs. requested during its lifetime continuity of service. The *user utility* is the integral of satisfaction over the nominal service duration. Essentially, the utility value increases when service continues.

2.3 Two-Leg Bandwidth Allocation

Instead of reserving bandwidth indefinitely for connections as in the traditional way, we propose to perform a two-leg admission control and bandwidth reservation for an AR request. The scheme works as follows. Initially, when the first time an AR request is issued *Leg-One admission control* is performed. In this phase, admission control only considers resource allocation for an initial fixed period of time called the *full warranty period* during which the requested bandwidth is reserved for its use if admitted. To handle situations where events may last longer than the warranty periods, a second leg - *Leg-Two admission control* is performed in which an at least minimum amount of bandwidth is reserved at the same time for another fixed period of time called the *at least minimum warranty period.* Admitted AR connection may issue warranty period extension requests at any time afterwards. Additional admission control will be required.

We choose two-leg absolute service warranties than statistical guarantee as in [11]. We believe that this model of advance reservation service is more meaningful because users clearly know the requested QoS is assured during the period. There are several advantages of this model. First, it is easy to implement by service providers. Second, the model is simple enough for the average user to understand so that the users feel comfortable. Known expectations of service assurance reduce risks. Moreover, the administrative cost of tracking usage is low.

Full Service Warranty Period

In the full service warranty period, a connection is assured with full bandwidth allocation. The choice of a good warranty period is essential to the assurance of service continuity. It indeed depends on the nature of the application, i.e., the age distribution function. If a larger value is chosen, the system must reserve resources for a longer period of time. Although this can increase service provider's confidence on service quality delivery and minimize the likelihood of service violation, a major concern is that network resources may be underutilized. Adversely, if a smaller value is used the system can achieve better resource utilization and blocking probability performance. The tradeoff is that more frequent service disruption and lower user satisfaction.

At Least Minimum Bandwidth Reservation for After-Warranty Period

The full service warranty period only represents expected or average duration. The rational behind the design of after-warranty period is to avoid sudden service disruption for connection whose event time is longer than this period, e.g., overtime of sports broadcast events. With resource reservation for the at least minimum warranty period, if a service extension request is rejected, the connection at least has a minimum bandwidth available to continue the service although the quality may degrade. The second leg warranty period is denoted as $D_{i,amw}$. Let parameters $\beta_{i,fw}$ and $\beta_{i,amw}$ be the probability thresholds of the full warranty period $D_{i,fw}$ and at least minimum warranty period $D_{i,amw}$(see Fig. 3).

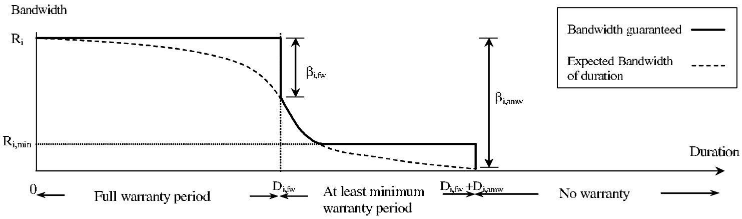

Fig. 3. The amount of bandwidth reservation at different warranty periods.

Compared to full bandwidth reservation, the tradeoff is link utilization. In fact, many Internet applications such as real-time audio/video streaming media are capable of adapting themselves to the network state and can tolerate certain degree of performance degradation. Hence, the bandwidth requirement of an AR request in the proposed service model is given in the form of $< R_i, R_{i,min} >$ where R_i is the bandwidth requirement of the service warranty period and $R_{i,min}$ is the minimum amount of bandwidth acceptable to the connection. The actual bandwidth reservation in the at least minimum warranty period for connection i would be in the range of $< R_i, R_{i,min} >$ (see Fig. 3). R_i can be the effective bandwidth [13,14] or the peak rate.

2.4 Revising Uncertainty with New Data

Uncertain resource allocation is complicated because of the form of ”probabilistic” in the duration of which the requested resources are needed as opposed to the fixed duration. Our work focuses on using new data to revise imperfect user-supplied initial knowledge of how long the connection will last. During the course of service, the service provider could periodically poll service user to update his/her knowledge of the connection lifetime or the service user can issue a status update to the service provider notifying whether an extension or early termination of the connection is needed.

3 Admission Control of Full Warranty Period and at Least Minimum Warranty Period

Let the total link capacity designated to the AR service denoted as C_{AR} ($C_{AR} < C_{link}$, C_{link} is the link capacity). Let $A'_i(t)$ is defined for each connection i:

$$A'_i(t) = \begin{cases} 0 & , t < 0 \\ 1 & , 0 \leq t \leq D_{i,fw} \text{(full warranty period)} \\ A_i(t) & , D_{i,fw} < t \leq D_{i,fw} + D_{i,amw} \text{(at least minimum warranty period)} \end{cases}$$

Consider the admission control of a new AR request. Let $< R_{new}, R_{new,min} >$ be the bandwidth requirements of the new connection; $D_{new,fw}$ and $D_{new,awm}$ are the full warranty period and at least minimum warranty service period of the new connection, respectively.

3.1 Leg-One Admission Control

Let P_w be the set of admission control decision points identified, i.e. $P_w = \{t_k, t_k - s_{new} \leq D_{new,fw}\}$, $W(t_k)$ is the set of connections that overlap with new connection at time t_k. The admission decision is based on the following equation:

$$\forall j, j \in W(t_k) \quad \sum_j max(R_{j,min}, (R_j \times A'_j(t_k - s_j)) + R_{new} \leq C_{AR} \tag{4}$$

3.2 Leg-Two Admission Control

Let P_{amw} be the set of admission control decision points identified, i.e. $P_{amw} = \{t_k, D_{new,fw} < t_k - s_{new} \leq D_{new,fw} + D_{new,amw}\}$. $W(t_k)$ is the set of connections that overlap with new connection at time t_k. The admission decision is based on the following equation:

$$\begin{aligned} \forall j, j \in W(t_k) \quad \textstyle\sum_j & max(R_{j,min}, (R_j \times A'_j(t_k - s_j))+ \\ & max(R_{new,min}, (R_{new} \times A'_{new}(t_k - s_{new})) \leq C_{AR} \end{aligned} \tag{5}$$

4 The Reservation Cost

We distinguish two costs for each admitted advance reservation request i: the reservation cost $c_{i,res}$ and actual cost $c_{i,act}$ defined as follows:

$$c_{i,res} = \int_0^{D_{i,amw}} max(R_i \times A'_i(t - s_i), R_{i,min}) dt \tag{6}$$

$$c_{i,act} = \int_0^{T_i} max(R_i \times A'_i(t - s_i), R_{i,min}) dt \tag{7}$$

Equation(6) is the integral of the total bandwidth reserved to connection i. This is the cost paid by the service provider. Equation(7) is the integral of the bandwidth actually used by connection i. The normalized reservation cost of the system for an interval τ is defined as follows:

$$c_{sys} = \frac{\sum_{i \in AR(\tau)} c_{i,res}}{\sum_{i \in AR(\tau)} c_{i,act}} \tag{8}$$

Its value is no smaller than 1. It will be used to evaluate the performance of the proposed scheme in the next section.

5 Performance Evaluation

In this section, we study the performance of the proposed two-leg advance bandwidth reservation scheme via simulation. The network configuration is shown in Fig. 4. The simulation period is 30 days (we take the daily average statistics from 30 days). For all sets of experiments, the requests are assumed of the type of videoconferences whose nominal service duration is a Pareto distribution with mean 120 minutes and shape=1.8. All requests have the same age distribution function and bandwidth requirements and $< R, R_{min} >=< 1.5Mbps, 256kbps >$

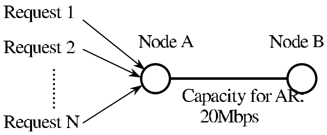

Fig. 4. Network Configuration of the simulation

The arrival process of advance reservation calls is assumed to be a Poisson process. Each call makes a connection reservation with start time in the next day. In each day, we divide 24 hours into "peak zones" (9am-12noon and 2-5pm) and "off-peak zones" (the other times of the day). The probabilities of reservations starting at peak zones or off-peak zones are assumed to be .7 and .3, respectively. Here, we assume the start time of an AR call must be at full or half o'clock (e.g., 9am, 9:30am, etc.). The start time distributions for calls in peak and off-peak zones are all uniform distribution.

5.1 Service Continuity, User Utility, and Reservation Cost

In this set of experiments, we compare user utility and reservation cost of the proposed Two-Leg bandwidth reservation with that of the traditional one-time reservation approach referred to as one-leg reservation. D_{fw} and D_{amw} are set to 80 minutes ($\beta_{fw} = 0.5$) and 116 minutes ($\beta_{amw} = 0.75$), respectively. Both these two interval values are used as the durations of the one-time reservation for the sake of comparison. In the Two-Leg bandwidth allocation scheme, service update is issued 60 minutes after connection starts. The arrival rate of the AR calls is 0.06 calls/minute. Table 1 shows the user utility. We can see that for connections whose nominal durations are greater than the full warranty period but less than the at least minimum warranty period, in terms of service continuity, it is one under the Two-Leg reservation with or without update. With update, the user utility is further improved. For those connections whose nominal durations are greater than the at least minimum warranty period, the Two-Leg reservation scheme outperforms one-time reservation scheme with D_{fw}. It is as expected that the Two-Leg reservation scheme is not as good as one-time reservation with duration D_{amw} because in the former, the bandwidth allocated after the full warranty period is a function of the bandwidth available at the time service extension request was issued. We know service extension requests often come as short notice and approach the end of the event. How to increase the acceptance probability of the service extension requests is one of the issues that we are currently looking into. In the aspect of reservation cost, the Two-Leg bandwidth allocation scheme performs very well, close to that of one-time with D_{fw}. This implies that the bandwidth reserved in the at least minimum warranty period is efficiently used.

Table 1. Comparisons of service continuity, user utility and reservation cost of the Two-Leg and traditional one-time bandwidth reservation schemes.

	$D_i \le D_{fw}$		$D_{fw} < D_i \le D_{amw}$		$D_{amw} < D_i$		c_{sys}(24-hours)
	$s(T_i)$	$u(T_i)$	$s(T_i)$	$u(T_i)$	$s(T_i)$	$u(T_i)$	
1-leg(D_{fw})	1.00	1.00	0.56	0.56	0.12	0.12	1.11
2-leg(No update)	1.00	1.00	1.00	0.64	0.35	0.17	1.16
2-leg(Update)	1.00	1.00	1.00	0.84	0.46	0.22	1.14
1-leg(D_{amw})	1.00	1.00	1.00	1.00	0.35	0.35	1.38

Figure 5 shows comparisons of call blocking probability under different arrival rates for the two schemes. As expected, because the extra bandwidth reservation for at least minimum warranty period in the Two-Leg reservation scheme with or without update, the call blocking probabilities are higher than that of one-time reservation with duration D_{fw} but lower than that of the one-time reservation with D_{amw}. Figure 6 shows the reservation cost. One can see that the reservation costs for the Two-Leg reservation scheme with or without update in peak zones, are close to those in the off-peak zones. Adversely, the reservation costs for the one-leg reservations are much higher than those of two-leg approach. Moreover, in the Two-Leg reservation scheme, the reservation cost when with update is lower that when without update. This is because that the bandwidth reserved after update is much likely utilized, thus lowering the reservation cost.

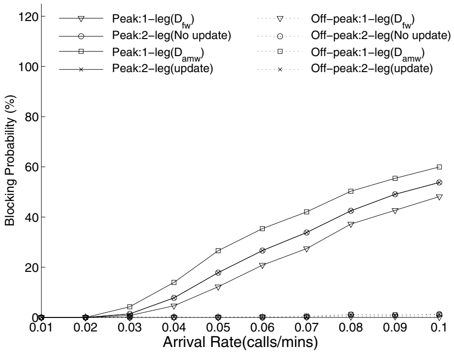

Fig. 5. Comparisons of call blocking probability.

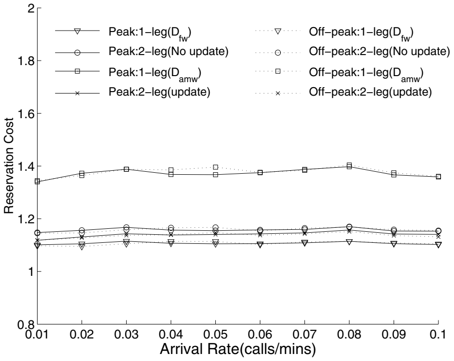

Fig. 6. Comparisons of reservation cost.

5.2 Service Continuity, User Utility, and Reservation Cost

The parameters β_{fw} (D_{fw}) plays an important role in the proposed Two-Leg bandwidth reservation scheme. In this set of experiment, we study the effect of different choices of the full warranty period on user utility and reservation cost. Here, the β_{amw} is fixed and set to 0.75. In Fig. 7, even with update in the Two-Leg reservation scheme, the improvement is limited. This is again because if the update is issued late during the connection lifetime, the blocking probability is likely high. Figure 8 shows the reservation cost under different full warranty periods.

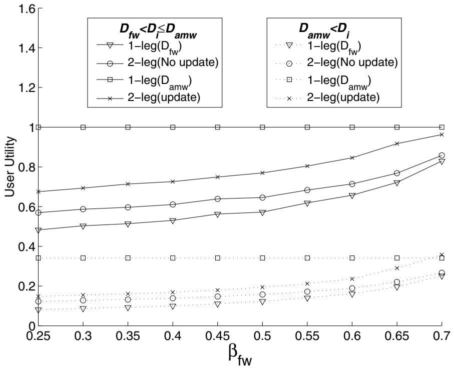

Fig. 7. Comparisons of user utility for different full warranty periods.

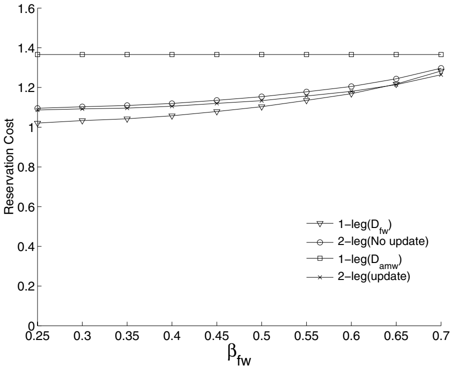

Fig. 8. Comparisons of reservation cost for different full warranty periods.

5.3 Update of Service Duration

In this set of experiments, we study the effect of at different times the service extension request is submitted in the Two-Leg reservation scheme in improving service continuity or user utility for those connections lasting longer than the initial full warranty period. In Fig. 9, we can see that both service continuity and utility do not change much when updates are submitted during the first 60% of the nominal service duration. After that, both increase. Figure 10 shows that reservation cost, has not much changes for different update submission times.

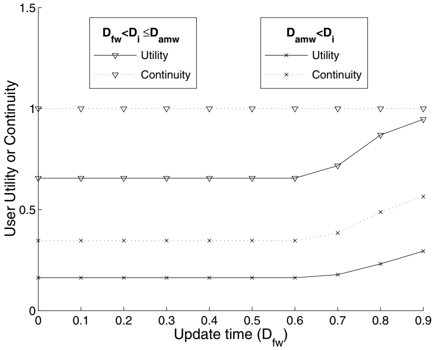

Fig. 9. Comparisons of service continuity and utility for service duration update of different submission times.

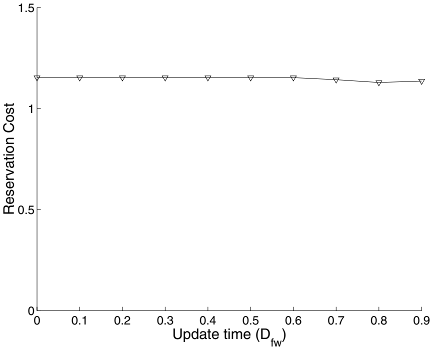

Fig. 10. The reservation cost for service duration update of different submission times.

6 Conclusion

It is difficult to do efficient resource management for advance reservations with uncertain service duration. In this paper we have presented a Two-Leg bandwidth reservation and admission control scheme. The idea is to perform bandwidth reservation in multiple stages. Each stage has a fixed duration and specific level of quality of service to assure. Thus, service provider can efficiently manage and allocate bandwidth needed to guarantee service quality to the connections at individual stages. Under the scheme, bandwidth reserved to an admitted AR request with uncertain duration includes a full bandwidth reservation for initial the service warranty period (Leg-One) and at least minimum bandwidth as well reserved for the after-warranty period (Leg-Two). The focus of the proposed scheme is not only to address the admission control issue at the initial call setup time but also the continuity of the service when events like overtimes occur.

An update mechanism is used to allow service user to update the network, especially a service duration extension is requested. If an update request cannot be satisfied, instead of reject a duration that best matches user's requirement is selected. In the worst case, Leg-Two bandwidth reservation assures at least minimum amount of bandwidth available to a connection to carry on the service.

The proposed scheme aims to provide service users a more predictive, affirmative service guarantees than gradually degrading service.

Finally, simulations are performed to evaluate the proposed schemes. Results show that the proposed scheme makes a good use of the bandwidth and outperform traditional one-time reservation in service continuity and user utility. The reservation cost is minimum, close to one-time reservation with fixed duration equal to the full warranty period even with additional bandwidth reserved for the at least minimum warranty period.

References

1. R. Braden, L. Zhang, "Resource ReSerVation protocol (RSVP) - Version 1: Functional specification, RFC 2205, 1997
2. W. Reinhardt, "Advance reservation of network resources for multimedia applications," in Proceedings of 2nd Intl. Workshop on Advanced Teleservices and High-Speed Communication Architectures (IWACA'94), Heidelberg, Germany, Sep. 1994.
3. A. Schill, F. Breiter, and S. Kahn, "Design and evaluation of an advance resource reservation protocol on top of RSVP," in Proceedings of IFIP Broadband'98, Stuttgart, Germany, April 1998.
4. F. Breiter, S. Kuhn, E. Robles and A. Schill, "The Usage of Advance Reservation Mechanisms in Distributed Multimedia Applicationsm," in Computer Networks and ISDN Systems, 30(16-18), Sept. 1998, pp. 1627-1635.
5. D. Ferrari, A. Gupta and G. Ventre, "Distributed Advance Reservation of Real-Time Connections," in NOSSDAV'95, Durhum, USA, Springer LNCS 1018, April 1995.
6. A. G. Greenberg, R. Srikant, and W. Whitt, "Resource Sharing for Book-Ahead and Instantaneous-Quest Calls," in IEEE/ACM Transactions on Networking, 1(7), April 1999.
7. M. Karsten, N. Berier, L. Wolf and R. Steinmetz, "A Policy-Based Service Specification for Resource Reservation in Advance," in Proc. of ICCC'99, Tokyo, Japan, 1999.
8. M. Degermark, T. Kohler, S. Pink and O. Schelen, "Advance Reservations for Predictive Service," in Proceedings of NOSSDAV' 95, Durham, NC, April 1995.
9. L. Delgrossi and L. Berger (Eds.), "Internet Stream protocol version 2 (ST2), protocol specification - version ST2+," in RFC 1819, Internet Engineering Task Force, August 1995.
10. R. A. Guerin and A. Orda, "Networks With Advance Reservations: The Routing Perspective," in Proceedings of INFOCOM 2000.
11. D. Wischik and A. G. Greenberg, "Admission Control for Booking Ahead Shared Resources," in Proceedings of INFOCOM'98 San Francisco, CA, April 1998.
12. The Blockbusters web page. http://www.blockbuster.com/
13. G. de Veciana, G. Kesidis and J. Walrand, "Resource Management in Wide-Area ATM Networks Using Effective Bandwidths", IEEE JSAC, Vol.13, No. 6, August 1995.
14. R. Guerin, H. Ahmadi and M. Naghshineh, " Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks", IEEE JSAC, Vol. 9, No. 7, September 1991.