Threshold-Based Admission Policies for Video Services

S.-H. Gary Chan
Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong

Fouad A. Tobagi
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
USA

Abstract— **In a video system, a central server stores all the files of interest to a large number of geographically distributed users. In order to lower the network transmission cost, the service provider may set up local servers to cache the video contents dynamically according to their local demands; thus forming a distributed video system. To lower the transmission cost further, the local servers can batch users together before presenting their video requests to the central server. In this case, if a video request to the central server is accepted, the provider enjoys the income collected by the batch while incurring a certain network transmission cost to the local servers. Due to the limited capacity at the central server, if video requests were accepted indiscriminately, the server may run out of channels too soon, which subsequently leads to the rejection of more profitable batches (and hence low profit). In order to better utilize the server channels, we consider admission policies which accept requests according to a threshold depending on the number of available channels at that time. We demonstrate that such admission policies can substantially increase the system profit while still meeting a certain user loss objective.**

Keywords— **Distributed video services, video-on-demand, admission policies, user loss and profit, request batching**

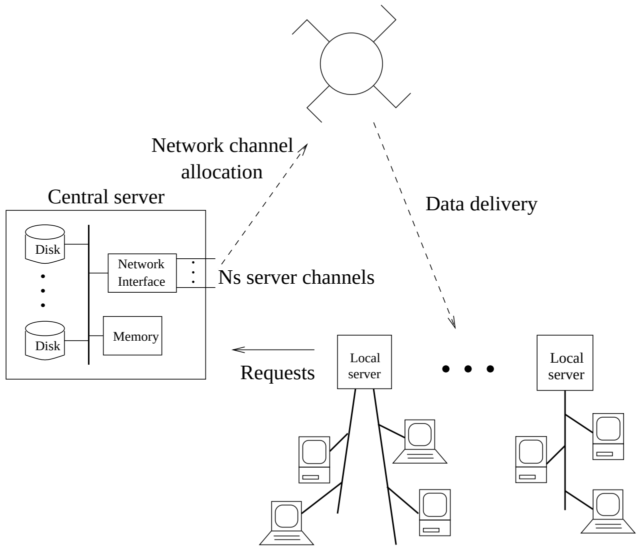

Fig. 1. A distributed video system.

I. Introduction

In order to provide video services such as movie-on-demand or distance-learning, a number of servers (collectively referred to as a central server) store all the files of interest to a large number of geographically distributed users. If the files were to be delivered directly to the users, the user capacity would be limited by the number of channels in the central server. Furthermore, since the users may not be co-located with the central server, the network transmission cost may be high. To overcome the problems and take advantage of the locality of access, local servers may be placed close to the user clusters. These servers obtain their files from the central server and cache them dynamically according to their local demands. (Such a distributed video system in fact has been discussed previously in [1], [2], [3], [4].)

We show in Fig. 1 a system for distributed video services, in which multiple head-end local servers stream the videos to their users through, for example, fiber or coaxial local loops. These servers may cache the more popular files in order to reduce the network bandwidth requirement. If users can tolerate some start-up delay (as in the application such as movie-on-demand), such bandwidth requirement can be further reduced by means of “batching” in which the local server groups user requests for the same video together before presenting a single request for the file to the central server. (Request batching for video applications is known as near video-on-demand. Interested readers are referred to [1], [5], [6] and references therein.) Since the number of users collected in a batch and the movie pay-per-view fee (PPV) may be different (due to, for example, users from different regions or priority classes), each batch to the central server presents different potential income to the provider. Since the central server has a limited number of channels, the main issue is then how to appropriately assign the available channels to file requests (or arrivals) presented to it in order to meet a certain performance objective, such as user loss rate or profit requirements. Upon the assigning of a server channel, a network transmission channel is then acquired to deliver the video file to the local server. The network channels may be acquired on-demand at some cost depending on the length of their usage. (This is the case for some satellite or ATM channels.) In this case, the challenge from the service provider's point of view is to balance the network usage cost with the revenue collected in serving the batches while achieving high service quality in terms of low user loss rate.

We focus on the central server. When its channels run out, the file requests would be blocked for service. We consider an on-demand central server in which a file request should be served immediately, and hence a blocked request (and by the same token all the users in the batch) can be

S.-H. Gary Chan was with the Department of Computer Science, University of California, Davis, when this work was done.

regarded as lost. The profit metric we use is the profit rate (i.e., the total collected revenue minus the total network transmission cost), which translates to the overall profit over a long period of time (e.g., over a year). Note that not all arrivals are profitable. Some batches have higher income than others, and the income of serving some batches may not even be able to cover their network channel transmission cost. However, in order to offer a reasonably low user loss rate, these "unprofitable" batches may still have to be served. If an indiscriminate admission policy were employed (i.e., all batches are indiscriminately accepted), the server channels would more often serve those unprofitable batches and run out very quickly. Once the channels run out, all the following requests have to be rejected, leading to high loss rate and low profit.

We therefore address the following question: given the limited number of channels in the central server, what should be the admission policy so as to achieve high profit while meeting a low requirement in the user loss rate (say around $5-10\%$)? To gain better control of the server channels, one policy is that we selectively reject requests earlier than the server actually runs out of channels. Obviously, we should not reject the requests randomly as some profitable batches would be inadvertently rejected as well. A better policy (from the profit point of view) is that we favor those more profitable batches by assigning a certain acceptance threshold depending on the number of channels currently in use. An arrival with a "characteristic" (such as the batch income) above the threshold is accepted; otherwise, it is rejected. Clearly, such a threshold-based admission policy achieve better profit by discriminating against low-income batches when the channels are about to run out. In this paper, we demonstrate how the threshold-based admission policies can turn a hardly profitable (or unprofitable) system into a highly profitable system. We show that low user loss rate and high profit may not be achieved at the same time (due to the fact that accepting more users may mean accepting more unprofitable batches).

This paper is organized as follows. After reviewing the related previous work in in Sect. II, we introduce the state-independent and state-dependent admission policies in Sect. III. In Sect. IV, we present some illustrative numerical examples on the performance of the admission policies: we first present the analysis of it with infinite server channels, and then present the simulation results and comparisons on the limited channels case. We conclude in Sect. V.

II. Prior Related Work

Chen et al. and Mundur et al. study admission policies for a media server in [7], [8], [9]. They consider two classes of users with different acceptance returns and rejection penalties, and investigate the respective number of channels each class should be assigned in order to maximize the reward rate. Instead of two distinct returns, we consider here arrivals with random returns and network transmission cost depending on the length of usage. We also study admission policies with admission threshold depending on the state of the system, which has not been studied previously.

Other previous work on admission policies for media servers mostly addresses disk access issues such as, given a certain disk bandwidth and access latency, what request-scheduling and data-storage/retrieval policies should be in order to support as many concurrent users as possible (see, for example, [10], [11]). Our work differs in that we consider *given* a certain number of server channels, how requests should be admitted so as to achieve high profit. While the previous work focuses on identifying the necessary system resources to serve a request (and the request is admitted once such resources are available), we consider here that even if there are idle channels, requests do not have to be admitted in order to achieve profit. The optimization we consider on the admission policies may resemble "continuous time Markov decision process," which deals with making optimal decision for a continuous time Markov process [12], [13]. However, since we are dealing with non-exponential service time with a cost structure depending on the service time of the request, and batch arrival with virtually infinite number of income possibilities (which leads to a large reward space in the system model), it is too complicated, if not impossible, to directly apply the theory here.

III. Admission Policies

Let N_s be the number of channels in the central server. Denote $s \in S = \{0, 1, 2, \ldots, N_s\}$ the state of the server at time t, where $s = i$ if i channels are being used. All requests arriving at $s = N_s$ are regarded as lost (i.e., cleared).

Let $Q_{th} \in S$ be the "alarm" state beyond which discriminate (i.e., threshold-based) admission policy is effected, i.e., all arrivals at system state $s < Q_{th}$ are accepted. Associated with each state $s \geq Q_{th}$ is an acceptance threshold α_s against which the characteristic of an arriving request for a file is compared. If the characteristic is less than α_s, the arrival is rejected; otherwise, it is served immediately. Clearly, one such threshold may be in terms of the income of the arrival in dollars (termed "revenue test"). The revenue test does not take into account the request's service time. For two arrivals with the same income, we may want to favor the arrival with shorter service time. Hence we can use a "rate test" in which we accept the batch if $i/T_h \geq \alpha_s$, where $\$i$ is the income carried by an arriving batch, and T_h is the file transmission time over the network (i.e., the service/holding time). From simulation, we find that the profit resulting from the two tests are very similar, and hence we only present results using revenue test in the remainder of our discussion.

In terms of the threshold value in each state, we can in general consider the following two admission policies:

• *State-independent policy:* The threshold is the same for all states $s \geq Q_{th}$, i.e., $\alpha_{Q_{th}} = \alpha_{Q_{th}+1} = \cdots = \alpha_{N_s} \equiv \alpha$, $\forall s \geq Q_{th}$. In other words, the admission criteria is independent of the system state for $s \geq Q_{th}$.

• *State-dependent policy:* In practice, we may want to accept more requests (i.e., a lower α_s) when the system is not loaded, so as to offer a lower loss rate and hence better service quality to users; however, when the system is loaded, more "stringent" admission policy can be employed to better utilize the channels so as to safeguard against frequent channel run-out. In the state-dependent admission policy, the threshold increases with the system state. In the absence of a proper theory for this increasing function, we consider an exponentially increasing threshold in $s \geq Q_{th}$ in the form of $\alpha_s = \beta^{(s-Q_{th})/(N_s-Q_{th})} \times 1$ dollar, where $\beta(\geq 1)$ is a dimensionless constant.

We consider that the cost in using a network channel is directly proportional to its holding time with the proportionality constant being C_{ch} (in \$/hr). Let R (in \$/hr) be the total revenue collected per hour over a long run, and E (in \$/hr) be the total expense (cost) of using network channels over a long run. We are interested in the profit θ in \$/hr given by $R-E$. Let p_L be the *user* loss rate. (Note that since a batch may contain many users, clearing the batch means rejecting multiple users; hence the batch loss rate may be different from the user loss rate). We would like to keep $p_L \leq p_L^*$.

We now derive the distribution of the batch income given the batch size (i.e., the number of user requests in the batch) and the PPV of the batch (all users in a batch pays the same PPV). Let B be the random variable of the batch size (i.i.d.) with $P(B=b) = r_b$ $(b \geq 1)$ and its expected value $E[B]$. Denote $G_B(z)$ its probability generating function, i.e., $G_B(z) \stackrel{\Delta}{=} \sum_b r_b z^b$. Let the random variable M (in dollars) be the PPV in the batches, with $P(M=m) = u_m$ and $G_M(z)$ its corresponding generating function.[1] Let W (in dollars) be the batch income. Given a batch of size b and PPV for each user in the batch being m dollars, the batch income is obviously given by $w = bm$. Let $P(W=w) = q_w$. By conditioning on the PPV M, the generating function for W is then given by[2]

$$G_W(z) = \sum_m u_m G_B(z^m). \tag{1}$$

We next derive the probability of a random user to be in a batch of income K dollars. Let such probability be given by $P(K=k) = t_k$. By conditioning on the batch size b, the generating function for K is

$$G_K(z) = \frac{\sum_b b r_b G_M(z^b)}{E[B]}. \tag{2}$$

IV. Illustrative Numerical Examples

In this section, we present some illustrative numerical results for the aforementioned admission policies. We show that substantial profit can be achieved by judiciously declining service for some of the unprofitable arrivals. We first analyze a case with infinite server channels which provides a first-cut insight into the trade-off between the user loss rate and profit. We have used simulation to study the case with finite channels (since the analysis for this case appears to be intractible), and the results will be covered next.

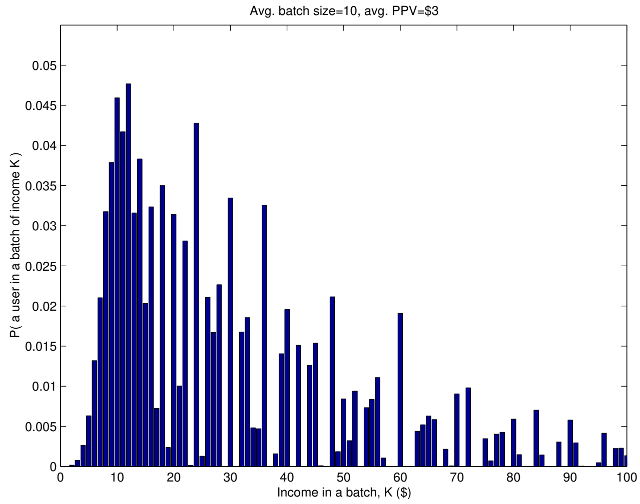

Fig. 2. The distribution of users in batches of income K (t_K vs. K).

We consider that requests arrive at the central server according to a Poisson process with rate λ batches/hr. We use $p_L^* \approx 6\%$, with C_{ch} =\$40/hr, $Q_{th} = 0$, and T_h being i.i.d. uniformly distributed between 30 minutes to an hour (giving $\overline{T}_h = 45$ minutes). The request rate under which the system is designed is $\lambda_0 = 200$ batches/hr. The batch-size distribution is a "truncated" (by 1) Poisson distribution of mean $\overline{n}$ (so that the probability with zero user in a batch is zero),[3] i.e., $r_b = \frac{\overline{n}^b}{b!(1-e^{-\overline{n}})} e^{-\overline{n}}$, and hence $E[B] = \overline{n}/(1-e^{-\overline{n}}) \approx \overline{n}$, for large $\overline{n}$. Consider also that the PPV in the batches is a geometric distribution with mean $\overline{g}$, and hence $u_m = (1-1/\overline{g})^{m-1}/\overline{g}$ $(m \geq 1)$. We will use $\overline{n} = 10$ and $\overline{g} = 3$ in the following.

We show in Fig. 2 the distribution of users in a batch with income K (from Eq. (2)). Note that while some batches are profitable (those with $K \geq C_{ch}\overline{T}_h = \$30$), many are not. In fact, since $E[B]\overline{g} \approx C_{ch}\overline{T}_h$ in our system, the system in the absence of admission policy hardly makes any profit at all. Many users are in batches with income around \$10 (the unprofitable batches). Therefore, in order to meet the loss rate requirement, the admission threshold is not expected to be higher than \$10.

A. The Case with Infinite Server Channels

We consider only state-independent admission policy in this section. Given a (constant) threshold of α dollars, clearly the user loss rate is given by $p_L = \sum_{k=1}^{\alpha-1} t_k$. The profit at a given arrival rate λ_0 is then given by,

$$\begin{aligned} \theta &= (1-p_L)\lambda_0 \left(E[W|p_L] - C_{ch}\overline{T}_h\right) \\ &= \lambda_0 \left(\sum_{w=\alpha}^{\infty} q_w w - (1-p_L) C_{ch}\overline{T}_h\right), \end{aligned} \tag{3}$$

[1] Though we consider integral values of the PPV, our results extend readily to the continuous case.

[2] We may condition on the batch size B to equivalently obtain $G_W(z) = \sum_b r_b G_M(z^b)$.

[3] Such distribution corresponds to the batch size distribution if the local requests arrive according to a Poisson process and the local servers batch the requests at regular intervals [5], [6].

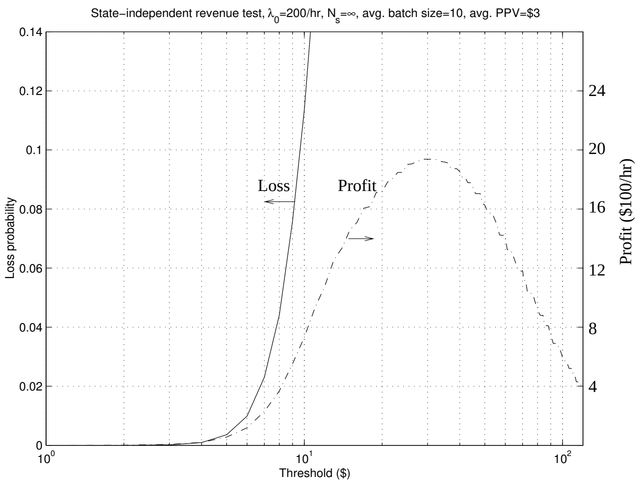

Fig. 3. Profit and loss rate vs. α for a video system.

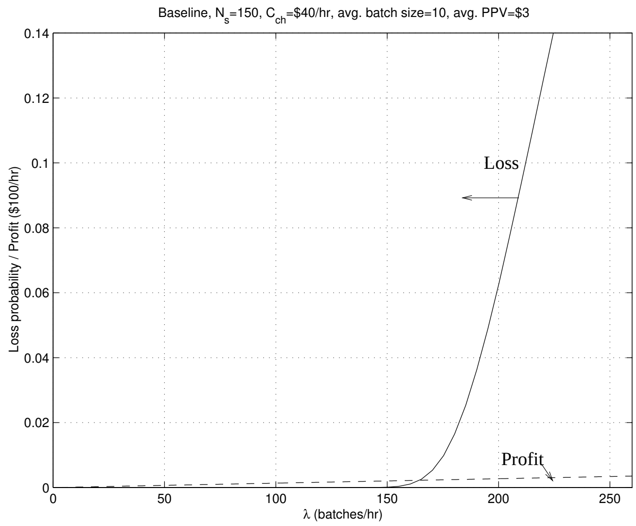

Fig. 4. Baseline system performance vs. λ.

where the first term in the bracket is the average revenue while the second term is the average channel cost.

We show in Fig. 3 the user loss rate p_L and the corresponding profit θ versus the threshold α. As α increases, p_L also increases, first quite slowly and then very sharply. The profit first increases (due to the acceptance of more profitable batches) and then decreases (due to rejection of too many batches, leading to under-utilization of channels and hence a decrease in θ). There is a maximum profit, but it can be hardly realized due to its corresponding high user loss rate. We see from the figure that given $p_L \leq 6\%$, α^* should be around \$8, achieving a high profit of more than \$800/hr. A lower threshold achieves lower loss rate and lower profit; nonetheless, the system can still be substantially profitable compared with a system without any admission policy at all (given by $\alpha = 0$). For example, with a threshold of \$7, a loss rate of as low as $\approx 2\%$ can be achieved while the system is still highly profitable (at about \$200/hr). The figure shows that the threshold can be chosen so as to achieve a balance in terms of loss rate and profit (which we have also observed in near-VOD systems [5]).

B. The Case with Limited Server Channels

In this section, we present the simulation results for the case with finite server channels. First we consider a baseline system without any admission policy, in which we accept all batches until all the channels run out. Given the number of channels N_s and the arrival rate λ batches/hr, the user loss rate is well-known to be $p_L = \frac{(\lambda \overline{T}_h)^{N_s}/N_s!}{\sum_{n=0}^{N_s} (\lambda \overline{T}_h)^n/n!}$, and the profit θ is hence given by $\theta = \lambda(1-p_L)\left(E[W] - C_{ch}\overline{T}_h\right)$. We show in Fig. 4 p_L and θ versus λ, with $N_s = 150$ channels (with the other parameters being the same as considered before). At the target arrival rate $\lambda_0 = 200$ batches/hr, even though $p_L \leq p_L^*$, the system is barely profitable ($\theta <$ \$0.4/hr). This shows that accepting requests indiscriminately may lead to low or even unprofitable services.

We next study the admission policies as given in Sect. III.

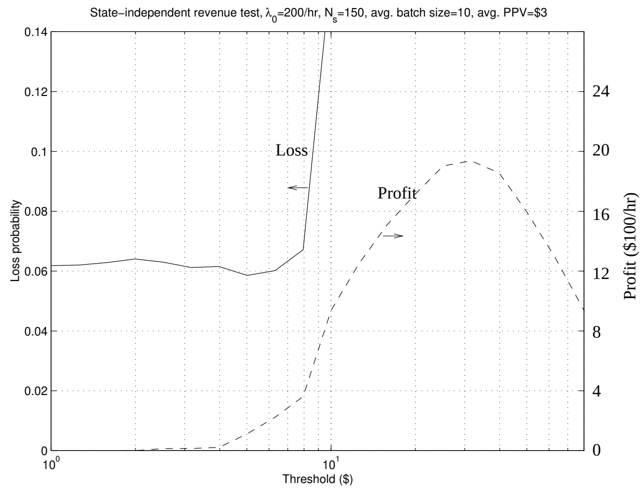

Fig. 5. p_L and θ vs. α with state-independent admission policy.

We first study the state-independent policy and show in Fig. 5 p_L and θ versus the threshold α given $\lambda_0 = 200$ batches/hr. As α increases, p_L first stays very flat until some point, beyond which the loss rate increases rapidly. This is expected by referring to Fig. 2, which shows that for $\alpha \leq\approx$ \$5, the fraction of users being rejected is quite low while the profit increases (as those highly unprofitable batches are discriminately rejected). Even up to $\alpha =$ \$8, p_L does not change much while θ has increased substantially from virtually zero to a significantly high value (of around \$400/hr).

Figure 5 indicates that for a target arrival rate of $\lambda_0 = 200$ batches/hr, we may choose a threshold of $\alpha^* =$ \$8 to achieve a high profit. We show in Fig. 6 the performance of such a system as the underlying arrival rate fluctuates around λ_0 given the threshold. Comparing with the baseline system (in Fig. 4), the system profit is indeed substantially increased. However, as λ drops below λ_0, p_L does not drop but remains at a rather constant level. This is because the threshold does not change according to the system state

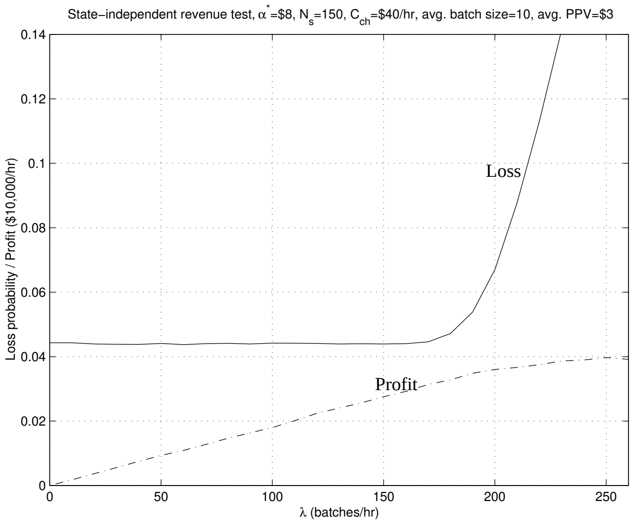

Fig. 6. p_L and θ vs. λ with state-independent admission policy.

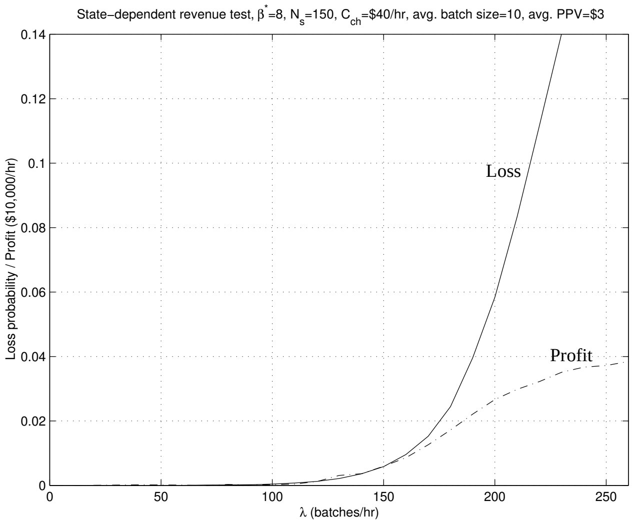

Fig. 7. p_L and θ vs. λ for the state-dependent admission policy.

and hence even when the system is not loaded, it still rejects unprofitable batches to achieve profit. Though the loss rate requirement is met, in reality we might want to offer a lower p_L when λ drops so as to offer a better service to the users. This is achieved by gradually lowering the threshold according to the system state by means of state-dependent admission policy.

With state-dependent threshold, we find that, given the same system parameters, the optimal value for β is 8. We show in Fig. 7 p_L and θ versus λ with such β. We see that indeed as λ fluctuates below λ_0, the loss rate decreases. At low λ, the state-dependent policy has a lower profit due to its indiscriminate nature in accepting requests to achieve a lower p_L.

V. Conclusion

In a distributed video system, a central server stores the files of interest to a large number of geographically distributed users, and local servers are set up to cache the contents dynamically depending on their local demands. To decrease the network transmission cost, the local servers may batch requests before presenting the file requests to the central server. An arrival at the central server, due to the difference in batch size and pay-per-view fee, has different batch income. Such income has to offset the network transmission cost to achieve profit. We have studied how server channels can be assigned to the incoming requests in a controlled manner by means of admission policies so as to achieve high system profit. The server is characterized by its state depending on the number of available channels. A certain threshold is set for each state so that the more profitable requests can be accepted. In the state-independent policy, the threshold is the same across the states, which leads to user loss even when the system is not loaded. On the other hand, in the state-dependent policy, the threshold is different for each state. This offers a good service quality in terms of user loss rate when the system is not loaded and high profit when the system is loaded. Such policy therefore can adapt to the fluctuating arrival rate to strike a balance between user loss rate and profit. We have demonstrated that applying the admission control policies can greatly increase system profit without sacrificing the user loss rate.

References

[1] V. O. K. Li and W. Liao, "Distributed multimedia systems," *Proceedings of the IEEE*, vol. 85, pp. 1063–1108, July 1997.

[2] A. D. Gelman, H. Kobrinski, L. S. Smoot, S. B. Weinstein, M. Fortier, and D. Lemay, "A store-and-forward architecture for video-on-demand service," *Canadian Journal of Electrical and Computer Engineering*, vol. 18, pp. 37–40, January 1993.

[3] F. A. Tobagi, "Distance learning with digital video," *IEEE Multimedia Magazine*, pp. 90–94, Spring 1995.

[4] S.-H. G. Chan and F. A. Tobagi, "Caching schemes for distributed video services," in *Proceedings of the 1999 IEEE International Conference on Communications (ICC'99)*, (Vancouver, Canada), pp. 994–1000, June 1999.

[5] S.-H. G. Chan and F. A. Tobagi, "On achieving profit in providing near video-on-demand services," in *Proceedings of the 1999 IEEE International Conference on Communications (ICC'99)*, (Vancouver, Canada), pp. 988–994, June 1999.

[6] S.-H. G. Chan, F. A. Tobagi, and T.-M. Ko, "Providing on-demand video services using request batching," in *Proceedings of the 1998 IEEE International Conference on Communications (ICC'98)*, (Atlanta, Georgia), pp. 1716–1722, June 1998.

[7] I.-R. Chen and T.-H. Hsi, "Performance analysis of admission control algorithms based on reward optimization for real-time multimedia servers," *Performance Evaluation*, vol. 33, pp. 89–112, July 1998.

[8] I.-R. Chen and C.-M. Chen, "Threshold-based admission policies for multimedia servers," *Computer Journal*, vol. 39, no. 9, pp. 757–66, 1996.

[9] P. Mundur, A. Sood, and R. Simon, "Threshold-based admission control for multi-class video-on-demand systems," in *Proceedings of the 1998 IEEE International Performance, Computing and Communications Conference*, (Tempe/Phoenix, AZ), pp. 154–60, February 1998.

[10] H. M. Vin, A. Goyal, and P. Goyal, "Algorithms for designing multimedia servers," *Computer Communications*, vol. 18, pp. 192–203, March 1995.

[11] F. Y.-S. Lin, "Optimal real-time admission control algorithms for the video-on-demand (VOD) service," *IEEE Transactions on Broadcasting*, vol. 44, pp. 402–8, December 1998.

[12] H. C. Tijms, *Stochastic Modelling and Analysis: A Computational Approach.* John Wiley & Sons, 1986.

[13] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming.* New York: Wiley, 1994.