An Adaptive Pricing Scheme for Content Delivery Systems

Srinivasan Jagannathan & Kevin C. Almeroth
Department of Computer Science
University of California
Santa Barbara, CA 93106-5110
{jsrini,almeroth}@cs.ucsb.edu

Abstract—**Content delivery has become an important enterprise on the Internet. There exists a huge demand for bandwidth-intensive, rich multimedia content like news, entertainment services and software. Two key issues associated with any content-delivery system are: revenue and user satisfaction. In this paper, we explore the domain of pricing for content delivery, and how it relates to revenue, and system utilization. A fundamental contribution of this paper is to consider a probabilistic user behavior model where users can refuse the offered service based on their capacity to pay and the price quoted. As a first step, we consider a system which delivers multimedia content on a First-Come-First-Served basis, and analyze pricing mechanisms which maximize the expectation of revenue. We argue that charging a constant price based on the customers' capacities to pay maximizes the expected revenue. Since the customers' capacities are highly varying and not known beforehand, we develop an adaptive pricing model and validate it using simulation. Our simulation results indicate that the adaptive pricing scheme generates nearly the same revenue as the theoretical expectation.**

I. Introduction

Content delivery is the new mantra in the Internet. Multimedia traffic is likely to grow to constitute a very large chunk of the overall Internet traffic. The sheer volume of data involved makes content delivery a lucrative business proposition. Much of the multimedia content distributed today is free, thus making it immensely popular. Associating a price with the content may significantly alter this trend. For instance, one may prefer to get news updates in plain text than say, in rich video format, if one has to pay a lot of money for the latter. In this way, pricing provides the content provider leverage to control the system utilization. A low price during off-peak hours may can attract customers while a high price can reduce system load during peak hours. Choosing the right price is thus of great importance for the success of the enterprise.

While our ultimate goal is to control system utilization using price, we would like to understand the causal effects of price. Therefore, in this paper, we investigate the problem of pricing at a content delivery server in the context of revenue generated. Under specific assumptions of user behavior and a system model, we analyze pricing mechanisms which maximize *expectation* of revenue. We use the term *expectation* in the statistical sense, because the revenue generated depends on a probabilistic user behavior model. To illustrate the probabilistic nature of user behavior, let us consider an example. Consider a teenager with $15 as pocket money at a video-game parlor. The latest release of a hit video-game is very attractive to him, but whether or not he chooses to play the game depends on the price associated with the game and the money he has with him. He may be very likely to play for $5, but not for $14. He may decide to wait for another month when the game is not so new and the price falls. There is a probability associated with his decision to play based on the price and his capacity to pay. We can see a direct correlation between the example described here and purchasing content on the Internet. In general, the probability that a customer buys the service is inversely proportional to the price and directly proportional to his or her capacity to pay. We try to capture this behavior in our work. Our work is based on a video-on-demand server, but it is sufficiently general to be applied to other forms of content.

Delivery of content depends on three factors—resource availability, customer capacity to pay and customer willingness to pay. One would like to preferentially serve customers who can and are willing to pay more for a service. To maximize revenue, one would have to make an *educated guess* about a customer's capacity and his willingness to pay. In this paper, we analyze pricing mechanisms under a *Pareto* distribution of customer capacity to pay. We also introduce a probabilistic model for user willingness to pay the quoted price. We argue that charging a constant price will maximize the expected revenue for any user willingness model in which user willingness decays with increasing price. We derive the constant price for the user willingness models we use in this paper. Since, the parameters of the Pareto distribution will not be known to the service provider, we develop an adaptive pricing model which estimates these parameters. We show, using simulations, that revenue using the adaptive pricing scheme matches closely with the expectation of revenue, given the probabilistic customer willingness to pay.

The rest of the paper is organized as follows. We describe our basic system model used in this paper in Section 2. We formulate the theoretical expectation of revenue in Section 3. In Section 4, we develop the adaptive pricing scheme. We validate it using simulations in Section 5. We conclude the paper in Section 6.

II. System Model

We consider a system where requests are satisfied if resources are available and the customer agrees to pay the quoted price. Resources are modeled as *logical channels*. Every request which is satisfied occupies a channel for some finite amount of time. For a video-on-demand server we can think of the channels as the number of movies that can be served simultaneously. In this paper we do not focus on how the channel is allocated or how an allocated channel is managed. These issues have been treated in detail in earlier work [1], [2], [3], [4]. We mainly focus on the interaction between the system and the customer before a channel is allocated. The sequence of actions resulting in content-delivery is depicted in Figure 1.

Economic theory has established that there are a large number of customers with a small income and a very small number of

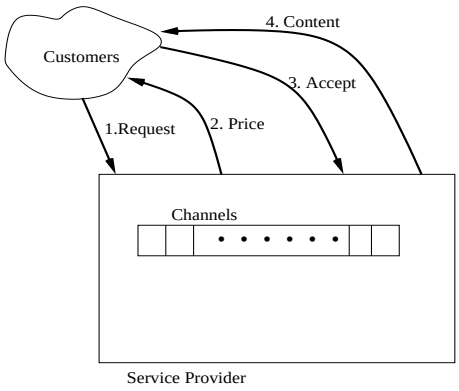

Fig. 1. System Model

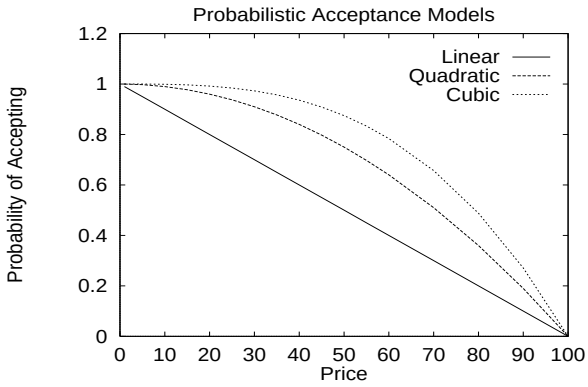

Fig. 3. User Willingness(Acceptance) Models

customers with very large income [5]. Any pricing scheme must be cognizant of this distribution. Currently, two probability distribution models – *Pareto* and *log-normal* are used to represent the distribution of incomes. In this paper, we use the Pareto distribution. Every customer has the capacity to pay based on a Pareto distribution with two parameters–shape α and scale b. All customers have capacities at least as large as b. The shape α determines how the capacities are distributed. The larger the value of α, the fewer the people with very large capacity to pay. The Pareto density function is defined as $f_{\wp}(x) = \frac{\alpha b^{\alpha}}{x^{\alpha+1}}$, for $x \geq b$. Figure 2 illustrates the Pareto density function for different values of shape α, and scale $b = 67$.

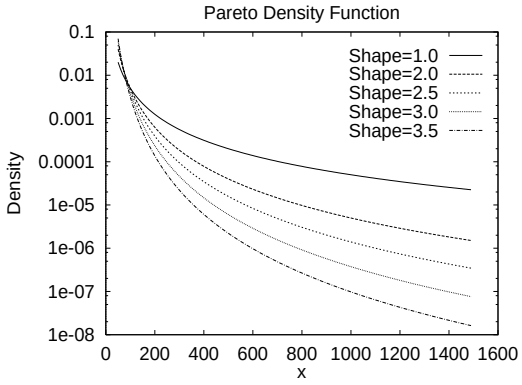

Fig. 2. Pareto Density Function

To adequately describe the willingness of customers to pay, we define a family of probability functions. Consider an arbitrary customer with capacity χ. We denote his decision to purchase the service, by the random variable Υ which can take two values–1 for accept and 0 for reject. As discussed in the example in the previous section, the probability that the customer accepts the price ψ, denoted by $P\{\Upsilon = 1 \mid \psi\}$ depends on his/her capacity χ, and the price ψ.

In this paper, we work with a simple model, where $P\{\Upsilon = 1 \mid \psi\}$ is defined as shown in Equation 1. By varying the parameter δ, we can make the willingness as *elastic* as desired. The higher the value of δ, the more willing are customers to spend money. We show three different willingness models for a customer having capacity 100, with δ values 1, 2, 3 respectively in Figure 3. As can be seen, the model with $\delta = 3$ makes the customer much more willing to spend money than in the case of the other two models. In this paper, we make a simplifying assumption that all customers will conform to one single model (as opposed to different customers obeying models with different values for δ).

$$P\{\Upsilon = 1 \mid \psi\} = \begin{cases} 1 - \left(\frac{\psi}{\chi}\right)^{\delta} & , \quad 0 \leq \psi \leq \chi \\ 0 & , \quad \psi > \chi \end{cases} \tag{1}$$

III. Expectations of Revenue and Acceptance

In this section, we discuss the dynamics of the user capacity model and user willingness model and how it affects revenue. Intuitively, if we do not know how much customers are capable or willing to pay, it makes sense to charge a constant amount of money from each customer. This is because we have no means of predicting which customer to charge a high price and which a lower price. By choosing a constant price we maximize the chances that they accept. We have proven this intuition correct using probability theory. We state the following theorems (without proof owing to reasons of space):

Theorem 1: If shape parameter $\alpha > 1$ (i.e., a finite mean for the Pareto distribution exists), willingness $P\{\Upsilon = 1 \mid \psi\}$ decreases monotonically with respect to ψ and tends to 0 as ψ approaches ∞, then the expectation of revenue, $E[\gamma]$ is maximum when ψ is a constant.

Theorem 2: For the user willingness defined in Equation 1, the expectation of the variable Υ given price ψ, $E[\Upsilon \mid \psi]$ is as follows:

$$E[\Upsilon \mid \psi] = \begin{cases} 1 - \frac{\alpha}{\alpha+\delta}\left(\frac{\psi}{b}\right)^{\delta} & , \quad 0 \leq \psi \leq b \\ \frac{\delta}{\alpha+\delta}\left(\frac{b}{\psi}\right)^{\alpha} & , \quad \psi > b \end{cases} \tag{2}$$

Theorem 3: For the user willingness defined in Equation 1, the expectation of revenue, $E[\gamma]$, is maximum when the price $\psi_{max} = \left[\frac{\alpha+\delta}{(\delta+1)\alpha}\right]^{\frac{1}{\delta}} b$. The expectation of Υ given price ψ_{max} is $\frac{\delta}{\delta+1}$.

According to Theorem 1, the content provider should charge a flat rate to maximize revenue for the system model used in this paper. Theorem 2 gives an estimate on the mean rate at which customers accept a quoted price ψ. Theorem 2 tells us that the mean rate of acceptance is at least $\frac{\delta}{\alpha+\delta}$ if the price is less than b, and at most $\frac{\delta}{\alpha+\delta}$ if the price is greater than b. Theorem 3 suggests what price should be charged to maximize revenue.

IV. Adaptive Pricing Scheme

The theory developed in the previous section tells us what price to charge to maximize the expected revenue. However, the price is dependent on the Pareto distribution parameters–shape α, and scale b and the willingness elasticity parameter δ. These parameters will not be known to the content provider. Moreover, they may not even be observable, because a customer will not disclose his capacity to pay. The only observable event, is the customer's acceptance of the quoted price (denoted by the binary variable Υ).

We have developed an adaptive pricing algorithm that learns from the rate at which customers accept a given price. We use the equations developed in Theorems 2 and 3 to relate the rate of acceptance and the price charged. There are three unknown variables, α, b and δ and only one observable event, Υ. Hence, we must assume some reasonable value for two of the unknowns to be able to predict the third. Since α is typically greater than 1 for income distributions[6], we assume one such value for α. We shall show later that mis-estimating α does not significantly alter our results. We now have to assume some reasonable value for one other unknown. Instead of assigning a single estimate we identify a set of possible values and then compute the other unknown. The parameter that we choose to identify a set for is the willingness elasticity parameter, δ. We choose the set Δ, consisting of feasible values of δ, such that it covers a wide range of elasticity of willingness. For instance, if we constrain δ to belong to the set $\{0.7, 1.0, 1.3, 1.6, 2.0, 2.4, 2.7, 3.0, 3.4, 3.8\}$, any actual value of willingness elasticity can be approximated to one of the elements in the set. Now, the problem of prediction is slightly more tractable.

Our algorithm adapts after observing a *round* of requests. Each round consists of 100 customer requests. A constant price ψ is charged in each round. This allows us to observe the rate of acceptance for price ψ. This observed rate is then equated to the formula for $E[\Upsilon \mid \psi]$ derived in Theorem 2. For each feasible value of elasticity δ, (i.e., elements of set Δ), we compute a possible value for b. We choose the appropriate equation to use from Equation 2 based on whether or not the observed rate of acceptance is greater or less than $\frac{\delta}{\alpha+\delta}$. Once we have a set of feasible values for b, we perform one more round of experiments. We choose an arbitrary price and compute the expected rate of acceptance for each of the feasible values of b. After this second round, we compute which feasible value of b most closely predicted the observed acceptance rate. We use that value for b, and the corresponding δ in Theorem 3 to get the price to be charged in the next round. The algorithm is presented in Figure IV.

1. Choose an arbitrary price ψ_0. and a value for α
2. Choose a set of elasticity values $\Delta = \{\delta_1, ..., \delta_n\}$
3. For the next 100 arrivals charge ψ_0.
4. Compute the observed acceptance rate p_0.
5. For each $\delta_i \in \Delta$, compute scale b_i using Theorem 2.
6. Choose another arbitrary price ψ_1.
7. For each $\delta_i \in \Delta$, compute expected acceptance rate for price ψ_1 using scale b_i and Theorem 2.
8. $k \leftarrow 1$
9. Repeat forever
10. For the next 100 requests, charge a price ψ_k
11. Compute the acceptance rate p_k.
12. For each $\delta_i \in \Delta$, compute scale b_i using Theorem 2.
13. Compare the acceptance rates predicted in round $k-1$ with the observed acceptance rate p_k.
14. Identify the δ_{opt} whose predicted acceptance rate most closely matches the observed acceptance rate p_k. Let b_{opt} be the scale computed in this round using δ_{opt}.
15. Set price ψ_{k+1} using δ_{opt}, b_{opt} and Theorem 3
16. For each $\delta_i \in \Delta$, compute expected acceptance rate for price ψ_{k+1} using scale b_i and Theorem 2.
17. $k \leftarrow k+1$.
18. End loop

Fig. 4. Our adaptive pricing algorithm

V. Validation

In this section, we validate the adaptive pricing scheme described in the previous section using simulations. We have implemented a simulator to model the content delivery system. The following is a list of parameters that we used for our analysis.

- **System Capacity:** This measures the number of simultaneous streams that can be served. We performed simulations with 500 logical channels.
- **Playout Duration:** The playout duration is the amount of time for which a logical channel is occupied for serving some request. For the results presented in this paper, we assume a duration chosen from a uniform distribution between 90 and 110 minutes.
- **Channel Allocation Policy:** We use a FCFS policy to allocate channels. Requests arriving when there are no free channels are rejected. There is no waiting queue.
- **Request Arrival Pattern:** A Poisson arrival process is simulated.
- **Customer Capacity:** This refers to the amount of money the customer can pay. In this paper, the capacities of individual customers are chosen from a Pareto distribution with scale 67 and shape 3.
- **Accounting Policy:** Each customer is charged a price suggested by the adaptive scheme. If the customer agrees to the price a channel is allocated.

The system capacity, the request arrival rate and the playout duration contribute to the system load and hence affect availability of resources. Our main interest though is the predictive capability of our algorithm, which is largely independent of these parameters[1]. Hence, we choose some constant values or ranges for these, and do not vary them in our simulations. The adaptive algorithm critically depends on the value of α that we assume and the set Δ we choose. We investigate the impact of assuming a wrong value for α and show that the revenue earned is not affected even for highly erroneous assumptions for α. We choose only one representative Pareto distribution of capacities for our simulation. This is because of two reasons. First, the actual shape of the distribution does not affect the results. It is the relative error between the assumed and actual values that will affect the simulation results. Second, the actual value of b is more related to the service being sold and its perceived value.

[1] A very high arrival rate would impact the maximum revenue that can be earned. However, we do not focus on this case in this paper.

Hence it is largely independent of our prediction algorithm.

Let λ be the arrival rate, n the number of channels and d, the mean playout duration. Then, using Theorem 3, the expected maximum revenue over time t is $min\{\ \frac{nt}{d},\ \frac{\delta\ \lambda\ t}{\delta+1}\ \} \times \left[\frac{\alpha+\delta}{(\delta+1)\alpha}\right]^{\frac{1}{\delta}} b$. We shall use this as our baseline for comparing our simulation results.

We present results to 1) validate Theorems 2 and 3, 2) show that a wrong assumption of α does not affect results and 3) show that the algorithm is able to generate nearly the predicted revenue for a wide range of willingness elasticity.

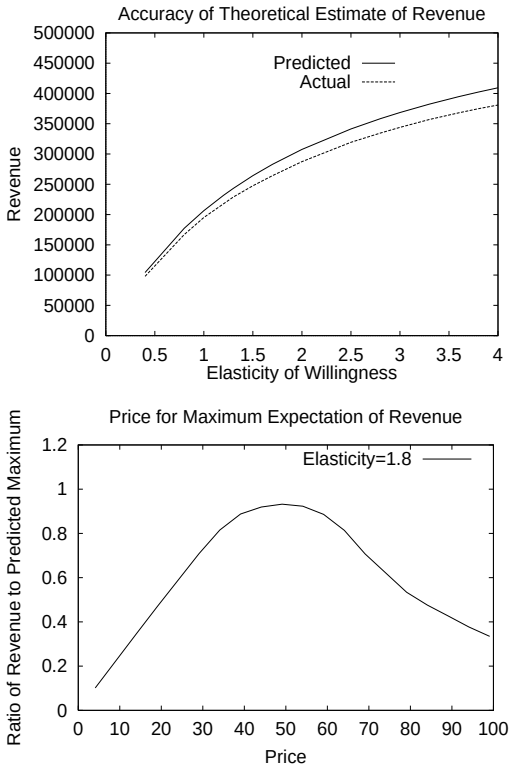

Fig. 5. Predicted and Actual Revenue

We ran numerous simulations to establish that Theorem 2 presents the expectation of the rate of acceptance. We show a plot of one such set of simulations with 500 channels, an arrival rate of 6 per minute, a mean play-out duration of 100 minutes and varying degrees of willingness elasticity. In this simulation, the content provider is assumed to have full knowledge of α, b and δ. The system charges the constant price suggested by Theorem 3. Note that the expectation of Υ presented in Theorem 3 is actually derived from Equation 2 in Theorem 2 by substituting the value for ψ_{max}. Figure 5 plots the revenue derived above as well as the actual revenue earned in the simulation. The duration of simulation was one day(simulated time). As can be seen, the actual revenue earned very closely matches the predicted revenue. The actual revenue earned is slightly less than the predicted revenue because some requests were rejected due to lack of resources when all channels were occupied. The predicted revenue assumes infinite resources. Figure 5 also illustrates one instance of a set of simulations to validate Theorem 3. Using Theorem 3, for δ=1.8, α=3 and b=67, the expectation of revenue is maximum when we charge a price of 49.1. We ran a set of simulations, charging a constant but different price in each of them. The ratio of revenue earned to the theoretically predicted maximum revenue is plotted with respect to the price charged in each simulation. As can be seen, the ratio is maximum and nearly 1.0 for price around 49.

An interesting consequence of Theorem 2, is its impact on system utilization. The system utilization given by $\rho = min\{1.0, \frac{\lambda E[\Upsilon=1|\psi]d}{n}\}$, is dependent on the price and the elasticity parameter δ. Figure 6 plots the predicted utilization and observed utilization, for the second set of simulations shown in Figure 5. The observed utilization is the fraction of simulated time for which the channels were occupied. As expected, the observed utilization closely matches the predicted utilization. Also of interest if the tradeoff between utilization and revenue. If user behavior were not probabilistic, one would expect increased revenue with increased utilization. This clearly is not the case with probabilistic user behavior.

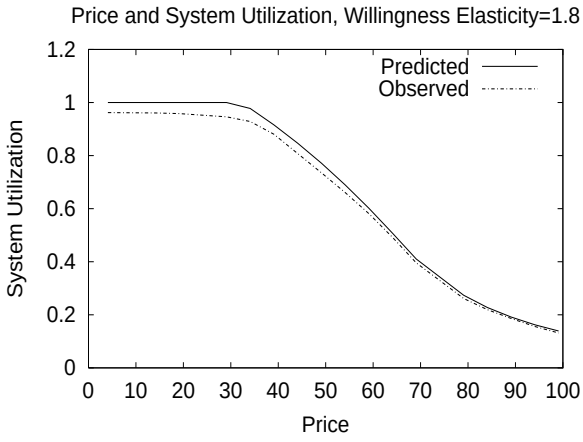

Fig. 6. Predicted and Observed Utilization

To validate the adaptive algorithm, we compare the revenue earned by our algorithm with the revenue earned by a prescient algorithm which knows the correct values for all three parameters α, b and δ and uses Theorem 3. We also compare it with the predicted maximum revenue. To show that a wrong assumption of α does not affect revenue, we ran simulations for different assumptions of α, (the actual value was 3). For each assumed value of α, we ran a range of simulations with different values for δ. The range of values of δ that we used was different from the set Δ which we use in the adaptive algorithm. For each such set of simulations, we found the average of the ratio of revenue earned to the predicted maximum. The arrival rate, system capacity and mean play-out duration were kept constant as in the earlier simulations. Figure 7 plots this average ratio for different assumed values of α. The average ratio is uniformly around 0.88 irrespective of the assumed value of α. The prescient algorithm has a ratio of 0.94. Note that the prescient algorithm does not need to make any assumptions as it knows all the values beforehand. This plot illustrates that 1) Wrong assumptions about α do not affect results and 2) the adaptive algorithm earns a revenue which is very close to the maximum possible.

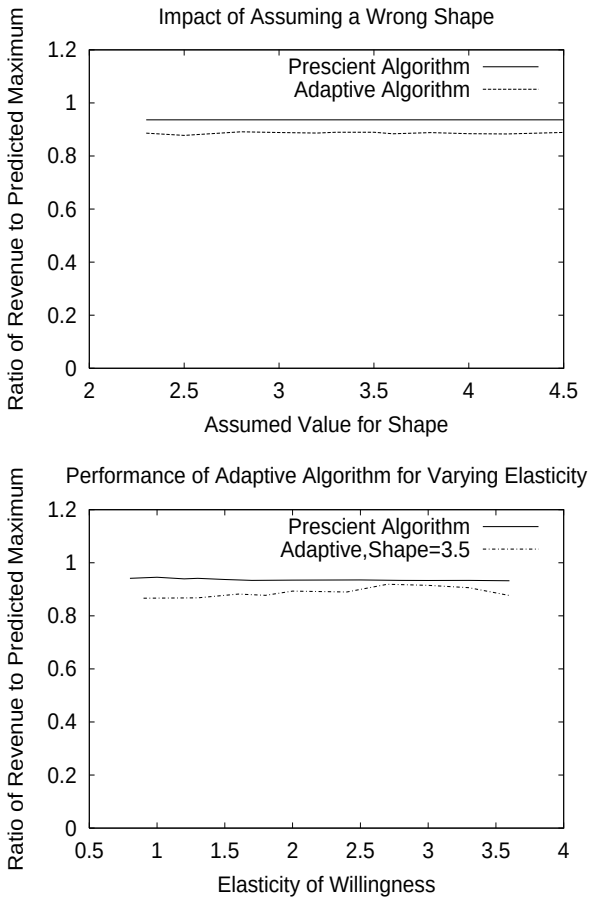

Fig. 7. Performance of Adaptive Algorithm

Figure 7, shows one representative plot (α was assumed to be 3.5 for this plot) from a set of simulations where we assumed some value for α and varied the willingness elasticity δ. Since, the maximum predicted revenue varies with δ, we again summarize the results using the ratio of revenue earned to predicted maximum. The corresponding ratio for the prescient algorithm is also shown. The adaptive algorithm clearly generates revenue very close to that generated by the prescient algorithm as well as to the predicted maximum.

Though we performed simulations to study the impact of the starting value ψ_0, we are unable to show any representative plots due to lack of space. We found that the starting value only influences convergence time.

VI. Conclusions and Future Work

In this paper, we focussed on pricing models for a content delivery system. We introduced the idea of probabilistic user behavior and analyzed its impact on revenue. We developed a theoretical framework for maximizing revenue and introduced an adaptive pricing scheme based on it. We validated the theoretical framework and also showed that, for the system model assumed, the adaptive pricing scheme generates revenue very close to the maximum possible.

Our future work is to focus on using price as an effective tool for content delivery management. An interesting application of Theorem 2, which relates price and acceptance rate, is to use price to control system utilization. For instance, price can be used to attract peak-hour customers to purchase the service during off-peak hours. Theorem 2 provides the tool for achieving this. An eventual goal of our work is to allow customer and provider to negotiate the price. Theorem 3 provides a baseline for controlling the negotiation. The system model used in this paper is simplistic and our work but a first step. The impact of batching, content popularity, and temporal changes in user behavior need to be studied in greater detail.

References

[1] K. Almeroth and M. Ammar, "The use of multicast delivery to provide a scalable and interactive video-on-demand service," *IEEE Journal on Selected Areas in Communications*, vol. 14, pp. 1110–1122, August 1996.

[2] K. Almeroth, A. Dan, D. Sitaram, and W. Tetzlaff, "Long term channel allocation strategies for video applications," in *IEEE Infocom*, (Kobe, JAPAN), April 1997.

[3] A. Dan, D. Sitaram, and P. Shahabuddin, "Scheduling policies for an on-demand video server with batching," in *ACM Multimedia*, (San Francisco, California, USA), October 1994.

[4] T. Little and D. Venkatesh, "Prospects for interactive video-on-demand," *IEEE Multimedia*, pp. 14–23, Fall 1994.

[5] B. C. Arnold, *Pareto Distributions*. Burtonsville, Maryland: International Co-operative Publishing House, 1983.

[6] H. Aoyama, S. W., Y. Nagahara, M. Okazaki, H. Takayasu, and M. Takayasu, "Pareto' s law for income of individuals and debt of bankrupt companies," *Fractals*, vol. 8, no. 3, pp. 293–300, 2000.