The Dynamics of Price, Revenue and System Utilization

Srinivasan Jagannathan and Kevin C. Almeroth

Department of Computer Science, University of California, Santa Barbara, CA 93106-5110
{jsrini,almeroth}@cs.ucsb.edu

Abstract. Content delivery is a growing enterprise in the Internet. Critical to the management of content delivery systems is understanding customer behavior, its impact on the consumption of system resources and how together, these affect revenue. We believe that price dictates overall customer behavior and that understanding the relationship between price and customer behavior is the key to effective system management. To this end, we study the impact of price on revenue and system utilization. Our service model is based on customers being able to refuse content based on their capacity to pay and their *willingness* to pay the price quoted. We quantify the effects of such customer behavior on revenue and system utilization. We argue that charging a constant price based on the customers' capacities to pay maximizes the expected revenue and formulate how to achieve this maximum. However, since customer behavior and characteristics are highly varying and not known *a priori*, we develop an adaptive pricing model which tracks user behavior as well as the arrival process to maximize revenue. We validate it using simulation. Our simulation results indicate that the adaptive pricing scheme generates nearly the same revenue as the theoretical expectation under very dynamic workloads.

1 Introduction

The volume of multimedia traffic has been steadily growing over the past few years. It is very likely that in the future, it will constitute the largest percentage of data transmitted in the Internet. The sheer volume of data involved makes content delivery a lucrative business proposition, so much so, that specialized overlay networks called *Content Delivery Networks (CDNs)* are being deployed to meet the demand. Some of the most important issues for managing content delivery networks include: (1) capacity planning, (2) controlling system utilization, and (3) maximizing revenue. Capacity planning is essential to meet future demand. System utilization is a measure of the popularity of the enterprise as well as an indicator of the short term availability of resources. Control over system utilization allows the content provider to provision resources in an efficient manner. At the same time, revenue maximization is the central goal of any business enterprise. Therefore, an important task is to develop mechanisms that control system utilization as well as maximize revenue. In this paper we investigate pricing and price adaptation as mechanisms to achieve these goals.

Pricing has long been used to control the arrival process and service time in telephony traffic. High peak-hour rates and low off-peak hour rates regulate the frequency and duration of long distance calls. However, one cannot directly apply these principles to content delivery networks. Some arrival patterns may not be easily influenced.

For instance, demand for movies and other entertainment software is likely to surge after work hours, e.g., during evenings. Similarly, demand can be expected to be higher during weekends. It is difficult to quantify the impact of discounts during off-peak hours on these trends. Interestingly, peak hours for content delivery may occur during off-peak hours of telephony. Moreover, it may not be possible to control service time as in telephony. For instance, it would be very difficult to significantly edit the size of a movie on a video-on-demand server. And it is more natural to charge a price per-movie rather than per-minute (or hour) of the movie. In this paper, we study pricing strategies for dynamically changing workloads for a First-Come-First-Served content provider serving content with well defined play-out duration.

A fundamental contribution of this paper is to recognize the probabilistic nature of user behavior and quantify its impact on the system. Let us consider an example to illustrate the probabilistic nature of customer behavior. 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. This probability is typically modeled using a utility function. In this paper, we choose a different but analogous model to represent this behavior. We can see a direct correlation between the example described here and purchasing content in the Internet. In general, the probability that a customer buys the service decreases with increasing price and increases with his or her capacity to pay. We try to capture this behavior in our work. 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. Our work is based on a video-on-demand server, but it is sufficiently general to be applied to other forms of content. Moreover, even though our work focuses on an Internet setting as a whole, it is equally applicable to content delivery on specialized broadband networks.

Delivery of content depends on three factors—resource availability, customer capacity to pay and customer willingness to pay. In this paper, we analyze pricing mechanisms for a system with limited resources, a *Pareto* distribution of customer capacity to pay and 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. We also derive a relationship between system utilization and price. Using this relationship, one can control system utilization in predictable ways by varying the price. Since the parameters of the customer capacity distribution will not be known to the service provider, we develop an adaptive pricing model which estimates these parameters. Our algorithm adapts to the user behavior as well as to dynamic workloads. 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.

Pricing mechanisms have been suggested earlier for managing resource allocation and congestion control [1, 2]. Researchers have also suggested differentiated pricing schemes based on service and quality [3, 4]. These differentiated schemes propose creation of service classes with different quality guarantees. In this paper, we assume one single service class for content delivery. Goldberg et al. [5] suggest flat rates for pay-per-view movies. But they argue that finding the optimal price to maximize profit is difficult. Our adaptive algorithm is a mechanism to overcome this problem. Basu and Little[6] consider the impact of price on the request arrival process for video-on-demand systems. They formulate pricing strategies assuming that a good user behavior model can be found using marketing analyses. In our work we take a different approach. We develop an analytical model that describes general user behavior and use an adaptive algorithm to learn the actual values of the parameters governing the user behavior.

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 and the relationships between system utilization and price 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.

2 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 [7–11]. We mainly focus on the interaction between the system and the customer before a channel is allocated.

Economic theory has established that there are a large number of customers with a small income and a very small number of customers with a very large income [12]. It is reasonable to assume that customers' capacities to spend will follow a similar behavior. Currently, two probability distribution models – *Pareto* and *log-normal* are used to represent the distribution of incomes[12, 13]. In this paper, we use the Pareto distribution to represent the capacity to spend. 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 a 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 1 illustrates the Pareto density function for different values of shape α, and scale $b = 67$[1]. Let us consider an illustrative example to understand the Pareto distribution of capacities. Consider a video-on-demand server. We can expect all customers to have a capacity to pay at least some money for the movie. We call the largest such amount that can be paid by all the customers as the scale of

[1] For, b=67 and $\alpha = 3$, the mean of the Pareto distribution is 100.

the distribution of their capacities and denote it as b. We would expect most of the customers to be able to pay only about this amount. There will be very few customers who can pay a lot more than the scale. This information is captured by the shape of the distribution, which we denote as α. The greater the value of α, the fewer the customers who can pay a lot more than b. For systems like video-on-demand servers, we would expect the shape to be very large.

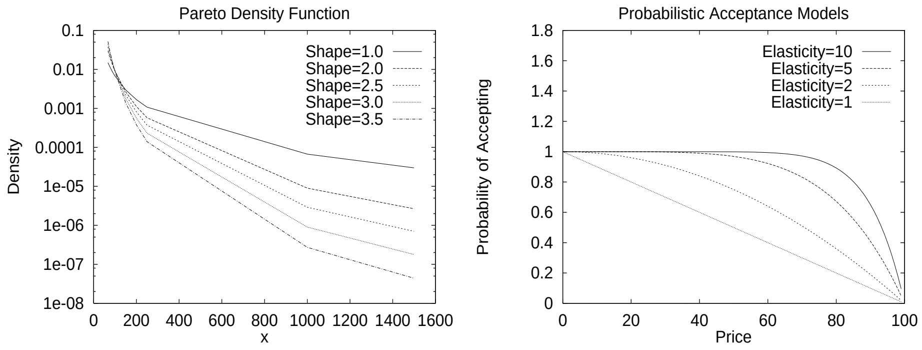


Fig. 1. Pareto Density(left) and Probabilistic User Willingness(right)

Even though customers *can* spend, they may not be *willing* to do so. 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/her 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.

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

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 2, 3 and 4 respectively in Figure 1. As can be seen, the model with $\delta = 4$ makes the customer much more willing to spend money than in the case of the other two models. In fact, as δ increases to around 4.0 or greater, the willingness begins to resemble a "step-function". To make our model easier to analyze, 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). As we shall show later, this does not affect the adaptive pricing algorithm that we develop in Section 4.

We chose this model of customer behavior over a utility function model because it is difficult to quantify the "value" a customer associates with the content. Choosing a probabilistic model on the other hand makes the analytical framework easier while at the same time capturing typical customer behavior. It can be shown that these models are equivalent to models where customers choose service-providers randomly and purchase service if the utility function is non-zero. The "value" a customer associates with the service in these models is a random variable with a density function which is closely related to the willingness model defined in Equation 1.

3 Price, Revenue and System Utilization

In this section, we discuss how customer capacities and their willingness to pay affects revenue and system utilization. We then derive expressions for the optimal price which maximizes the expectation of revenue under different workload conditions.

Even though the overall distribution of capacities is assumed to be Pareto, there is no way to know the capacity or willingness of any given customer. Hence, to maximize revenue, it makes more sense to charge every customer a constant price. By choosing a constant price we maximize the chances that they accept. We have proven this intuition correct using probability theory. The proof hinges on the fact that the expectation(mean) of a function(in this case revenue) is maximized when the probability that the function takes the maximum value is 1, i.e., the function is a constant. The actual value of the price for which expectation of revenue is maximized depends on the user willingness model. We state the following theorems without proof (owing to reasons of space):

Theorem 1. *If the shape parameter $\alpha > 1$ (i.e., a finite mean for the Pareto distribution exists), and 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 maximized 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 maximized 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.

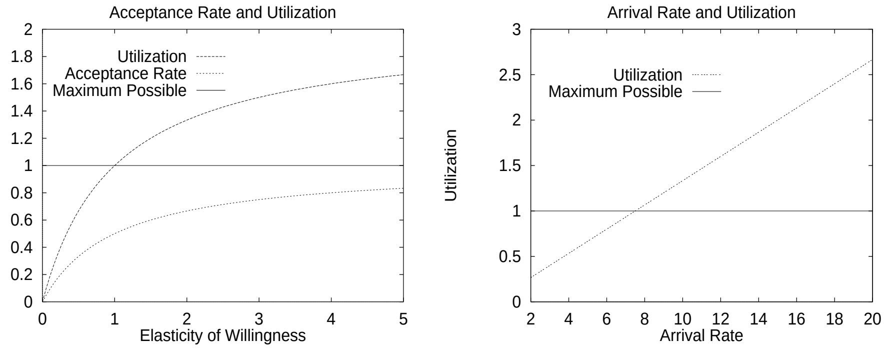


Fig. 2. Price, Arrival Rate, Acceptance Rate and System Utilization

Consider a system with n channels. If requests arrive at a rate λ, and the mean duration to service the request (e.g., the play-out time for a movie) is d, then the system utilization, ρ, when we charge price ψ is given by:

$$\rho = \frac{\lambda E[\Upsilon \mid \psi] d}{n} \tag{3}$$

According to Theorem 3, when the content provider charges a price $\left[\frac{\alpha+\delta}{(\delta+1)\alpha}\right]^{\frac{1}{\delta}} b$, the rate at which customers accept the price is $\frac{\delta}{\delta+1}$. Hence, the system utilization ρ_{max} which yields maximum expectation of revenue is given by $\frac{\lambda\delta d}{(\delta+1)n}$. Figure 3 shows how the acceptance rate, $E[\Upsilon \mid \psi_{max}]$, and ρ_{max} vary with the elasticity of willingness, δ, for a system with 500 channels, a mean play-out duration of 100min and an arrival rate of 10 requests/min. Also shown is how ρ_{max} varies with the arrival rate for the same system when the elasticity, δ, is 2.0. Note that system utilization cannot be greater than 1. This implies that resource constraints may prevent achieving the maximum expectation of revenue when either the arrival rate is high, or willingness to pay is very high. Therefore in an environment with large fluctuations in arrival rate or one in which customer willingness is high, the price charged must adapt itself to achieve maximum revenue. Note that in a system with n channels, and mean play-out duration d, on average, not more that $\frac{nt}{d}$ requests can be served in a time duration t. Also note that according to Theorem 3, the maximum expectation of revenue is achieved over time t, when $\frac{\lambda\delta t}{\delta+1}$ requests are served. Thus, when $\frac{\lambda\delta t}{\delta+1} > \frac{nt}{d}$, resources are insufficient to achieve maximum expectation of revenue. It is intuitive to increase the price at such a juncture to such an extent that the acceptance rate declines to equilibrium, i.e., only as many customers accept the price, as can be accommodated

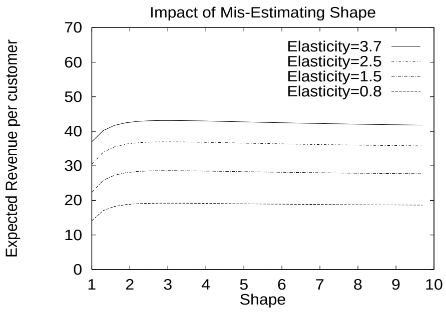


Fig. 3. Impacting of Mis-estimating Shape

by the system. We have proved using calculus that this does indeed achieve maximum revenue. We state the following theorem without proof:

Theorem 4. *Let customer capacities be Pareto distributed with shape α, $\alpha > 1$, and scale b. Let their willingness to pay be as defined in Equation 1. Consider a system with n channels serving content with mean play-out duration d. Let λ be the request arrival rate. The expectation of revenue for the system is maximum when the content provider charges a price ψ_{MAX} defined as follows:*

$$\psi_{MAX} = \begin{cases} \left[\frac{\alpha+\delta}{(\delta+1)\alpha}\right]^{\frac{1}{\delta}} b & , \quad \frac{\delta}{\delta+1} \leq \frac{n}{d\lambda} \\ \left[\left(\frac{\alpha+\delta}{\alpha}\right)\left(1 - \frac{n}{d\lambda}\right)\right]^{\frac{1}{\delta}} b & , \quad \frac{\delta}{\delta+1} > \frac{n}{d\lambda} \geq \frac{\delta}{(\alpha+\delta)} \\ \left[\frac{\lambda d\delta}{n(\alpha+\delta)}\right]^{\frac{1}{\alpha}} b & , \quad \frac{\delta}{\delta+1} > \frac{\delta}{(\alpha+\delta)} > \frac{n}{d\lambda} \end{cases} \tag{4}$$

Theorem 4 gives the content provider a mechanism to maximize revenue based on arrival rate and observed willingness elasticity δ. We delve into this in greater detail in the next section.

4 Adaptive Pricing Scheme

The theory developed in the previous section tells us how the user acceptance rate varies with price and what price to charge to maximize the expected revenue. However, the optimal 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 events are the customer's acceptance of the quoted price (denoted by the binary variable Υ), and the request arrival rate λ.

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 two observable events, Υ and λ. Of these two observables, λ cannot always be used. Hence, we must assume some reasonable value for two of the unknowns to be able to predict the third:

1. We assume an arbitrary value for α. This is because the expected revenue is not very sensitive to α even for moderately large values of α. And we expect α to be large for the customer capacities. Figure 3 shows the impact of assuming a wrong value of shape. The figure plots the expected revenue for different assumed values of shape. The actual value of shape used for calculations was 3.0, and the scale 67.0. As is seen, even a highly erroneous estimate of α does not significantly impact revenue. This is because of the heavy-tailed nature of the distribution. We shall also show later using simulations that mis-estimating α does not significantly alter our results.
2. 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.

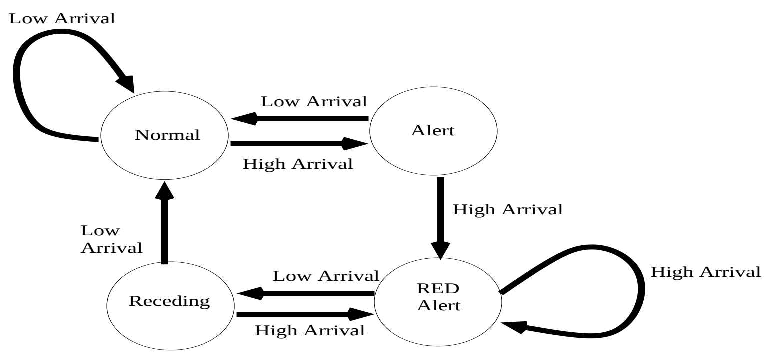


Fig. 4. State Diagram to Monitor Arrival Rate

Our algorithm adapts after observing a *round* of requests. Each round consists of 100 customer requests. Note that we chose 100 arbitrarily. We want to observe acceptance rate over a reasonable number of customers. 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. At the same time, the arrival rate is also monitored. The time elapsed for one round to complete gives an estimate of the current arrival rate. 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

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 and the arrival rate λ_0.
5. $\forall \delta_i \in \Delta$, initialize state S_i to "Normal" or "Alert" based on δ_i and λ_0.
6. $\forall \delta_i \in \Delta$, compute scale b_i using Theorem 2.
7. Choose another arbitrary price ψ_1.
8. $\forall \delta_i \in \Delta$, compute expected acceptance rate
for price ψ_1 using scale b_i and Theorem 2.
9. $k \leftarrow 1$
10. Repeat forever
11. For the next 100 requests, charge a price ψ_k
12. Compute the acceptance rate p_k and arrival rate λ_k.
13. $\forall \delta_i \in \Delta$ update state S_i based on λ_k.
14. For each $\delta_i \in \Delta$, compute scale b_i using Theorem 2.
15. Compare the acceptance rates predicted in round $k - 1$
with the observed acceptance rate p_k.
16. 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}.
17. Set price ψ_{k+1} using δ_{opt}, b_{opt} and Theorem 4
18. For each $\delta_i \in \Delta$, compute expected acceptance rate
for price ψ_{k+1} using scale b_i and Theorem 2.
19. $k \leftarrow k + 1$.
20. End loop

Fig. 5. Our adaptive pricing algorithm

acceptance rate. We use that value for b, and the corresponding δ in Theorem 4 to get the price to be charged in the next round. Since the price to be charged depends on the arrival rate, we maintain a state based on the state diagram shown in Figure 4. If $\frac{\lambda\delta}{\delta+1} \leq \frac{n}{d}$, then we term the arrival rate as "low". Otherwise, we call it "high". Since a high arrival rate in one single round of 100 requests can be a false alarm, we change state to "Red-Alert" only when the arrival rate is high enough in two successive rounds. Similarly, when in a "Red-Alert" state, there is a transition to "Normal" only after two successive rounds of low arrival rate. Since state transitions are dependent on elasticity value δ, for each $\delta_i \in \Delta$, we maintain a separate state S_i. Thus, depending on the current value of δ being used, we charge a price based on that δ and the corresponding b and state S. For states "Normal", "Alert" and "Receding", we charge $\left[\frac{\alpha+\delta}{(\delta+1)\alpha}\right]^{\frac{1}{\delta}} b$ while for the state "Red-Alert", we charge either $\left[\left(\frac{\alpha+\delta}{\alpha}\right)\left(1 - \frac{n}{d\lambda}\right)\right]^{\frac{1}{\delta}} b$ or $\left[\frac{\lambda d\delta}{n(\alpha+\delta)}\right]^{\frac{1}{\alpha}} b$ based on the value of $\frac{nd}{\lambda}$ as shown in Theorem 4. The algorithm is presented in Figure 5.

5 Simulations

In this section, we validate the theory and algorithm developed in the previous sections using simulations. We have implemented a simulator to model the content delivery system. Our simulations can be divided into two broad categories. One set of simulations validates the theoretical model developed in Section 3. We show that Theorem 2 accurately predicts the acceptance rate and that the price suggested by Theorem 3 maximizes the revenue. We show that the predicted system utilization closely matches the observed system utilization. We also validate Theorem 4. The second category of simulations deals with the adaptive pricing algorithm. The adaptive algorithm assumes some value for the shape of the customer capacity distribution. We show that the revenue earned is fairly insensitive to this assumed value. We compare the revenue earned using our algorithm with that earned by a prescient algorithm that knows all the parameters of the customer distribution. We show that the adaptive algorithm earns revenue very close to the predicted maximum expectation for a wide range of customer willingness elasticity and highly varying workloads. We also show simulation results that indicate that the adaptive algorithm is robust in dealing with situations not modeled by the theoretical framework developed in Section 3.

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. We chose a fixed number of channels because the system capacity typically does not change very often.
- **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. This closely models the typical length of movies.
- **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:** Dynamically varying arrival processes are simulated. We describe these in greater detail later in this section.
- **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 3.0. This distribution has a mean value of 100. We chose only one representative Pareto distribution of capacities 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. Hence it is largely independent of our prediction algorithm.
- **Elasticity of Willingness:** The parameter δ represents the willingness of the customers to purchase content at a certain price. We used $\delta \in [0.5, 5.0]$. When δ is small, the customer is extremely unwilling to purchase the content. For δ values around 5.0, the probability that the customer will purchase the content is nearly 1.0 (if the price is less than his capacity to pay).

The workload models we used in our simulations are shown in Figure 5. These models are adapted from the work on arrival-rate based scheduling by Almeroth et al. [9]. The workloads are modeled based on a 24 hour period beginning from 8.00am of one day and running to 8.00am of the next. "Prime time" periods see a surge in demand. We have used a steady baseline workload, with no surges in demand, and three non-steady workloads. The arrival rates during prime time for the non-steady workloads was around five times greater than the normal rate, based on statistics reported by Little and Venkatesh [11]. We simulated both gradual as well as sudden increases in arrival rate. We also used a workload with hourly spikes in arrival rate during primetime. This type of workload is based on the belief that the workload for some systems may be synchronized with an external event like wall-clock time.

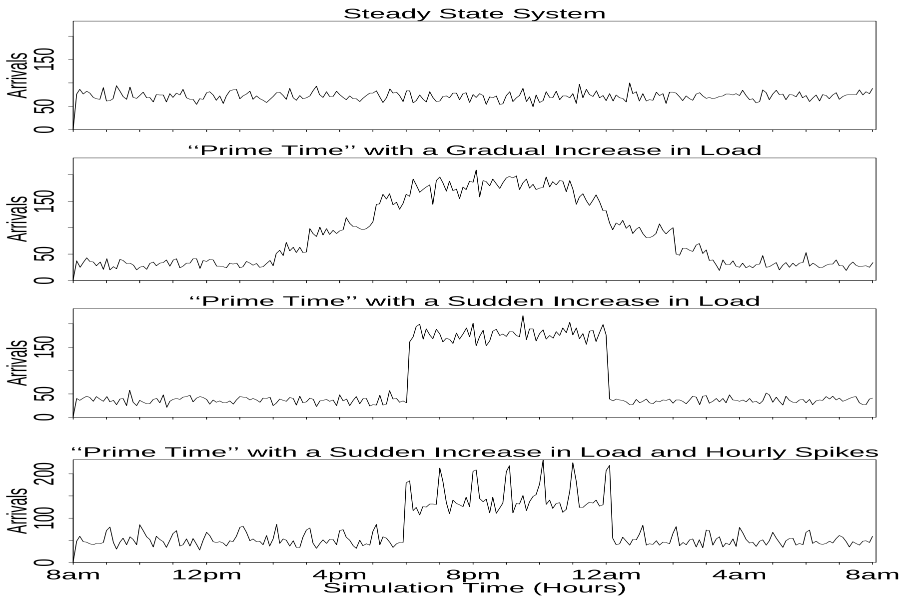


Fig. 6. Workloads

5.1 Validating the Analytical Framework

To validate Theorem 2, we ran different sets of simulations, charging a constant price from each customer and varying the price across simulations. We chose arrival rates which would not overload the system. Figure 7 shows the observed acceptance rate and the predicted acceptance rate for one such set of simulations. The arrival rate for this set of simulations was 6 per minute and the elasticity of willingness, δ, was 2. As is seen, the predicted acceptance rate matches the observed acceptance rate. The other graph shown in Figure 7 plots the predicted revenue and actual revenue earned in those simulations. The revenue is plotted with respect to the price charged. This graph indicates that the predictions are very close to the observed values. The maximum revenue is earned when a price close to 50 is charged. This corroborates Theorem 3, according to which, the expectation of revenue is maximum when the price is $\left[\frac{3+2}{(2+1)\times 3}\right]^{\frac{1}{2}} \times 67 = 49.93$. Also note that the acceptance rate for price 49.93 is around 0.66, which is very close to $\frac{2.0}{2.0+1}$ as predicted by Theorem 3.

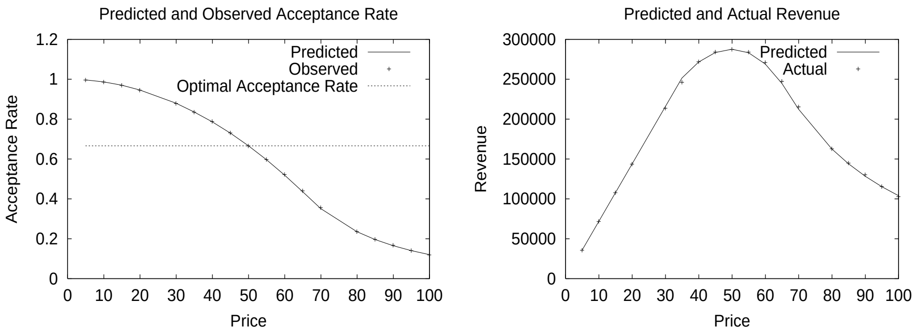
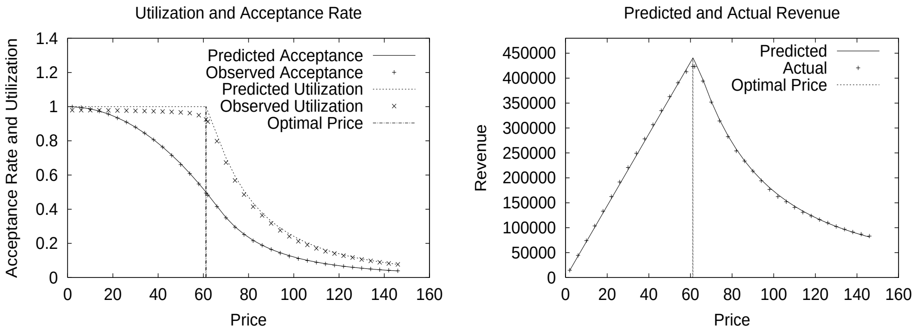


Fig. 7. Verification of Theorem 2 and Theorem 3

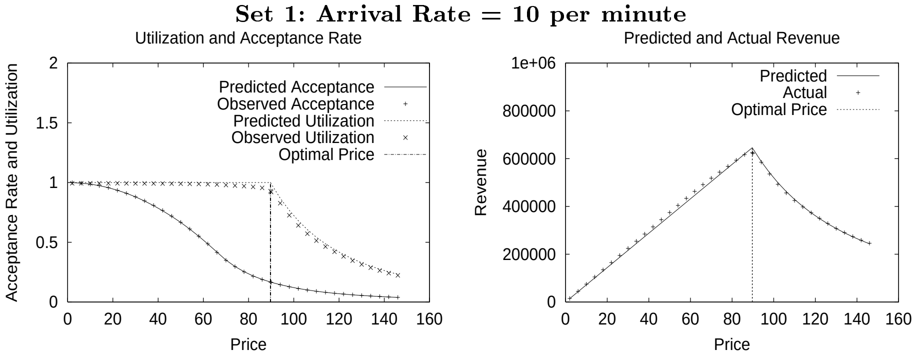
To confirm our intuition that during peak-hours, we must charge as much money as will result in highest predicted system utilization, we ran different sets of simulations with very high arrival rates. In each set of simulations, we varied the price across simulations. The price charged in each particular simulation was kept constant. We show the system utilization, acceptance rate and revenue earned in two such sets of simulations in Figure 8. The arrival rate in these sets of simulations were 10 per minute and 30 per minute respectively. Henceforth we shall refer to these sets as *Set 1* and *Set 2* respectively. The elasticity δ for both the sets was 2. Note that according to Theorem 3, a price of 49.93 will yield maximum revenue and result in an acceptance rate of 0.667. However, this means that on average we have 6.67 requests (for Set 1) and 20 requests (for Set 2) being serviced per minute. But the system cannot sustain more than $\frac{500}{100}$ requests per minute. Thus we cannot achieve the maximum predicted by Theorem 3 due to lack of resources. Theorem 4 was derived for such situations. Theorem 4 predicts that highest expectation of revenue is ahcieved at price 61.16 for Set 1 and at price 89.70 for Set 2. We observe that this prediction is indeed true for both Set 1 as well as Set 2. Furthermore, the acceptance rate predicted by our theorems are also accurate. The observed system utilization also closely matches the predictions. We found Theorem 4 to be accurate in all the sets of simulations we ran.

5.2 Validating the Adaptive Pricing Algorithm

Having validated our theoretical framework, we performed simulations to validate our adaptive algorithm. We ran simulations to: 1) evaluate how much revenue is generated using the adaptive algorithm, 2) study how the assumed value of shape impacts revenue, 3) study the impact of the starting price and, 4) evaluate the robustness of the algorithm. To evaluate the performance of the adaptive algorithm, we compared the revenue generated by it to the predicted expectation as well as to the revenue generated by a prescient algorithm. The presicent algorithm has knowledge of all the



Set 1: Arrival Rate = 10 per minute



Set 2: Arrival Rate = 30 per minute

Fig. 8. Verification of Theorem 4

parameters in the simulation except the exact knowledge of the capacity and willingness of each individual customer. The revenue generated by the prescient algorithm represents a physical upper-bound on the expected revenue. Since the predicted revenue will vary with willingness elasticity δ and the workload, we use the ratio of actual revenue and predicted revenue as a metric. We shall call this ratio as the *Revenue-ratio*. The greater this ratio, the better the algorithm. Note that this ratio can be greater than 1.0 because we are predicting the maximum expectation of revenue. We also use the predicted and observed system utilization as a measure of the performance of the algorithm. The closer the predicted and observed utilization, the better the algorithm. The system utilization achieved by the prescient algorithm gives us an idea of the physical bounds even when we have complete knowledge.

We ran a wide range of simulations varying the assumed value of shape as well as the elasticity of willingness δ. Figure 9 depicts the the observed range of Revenue-ratio for different assumed values of shape. For each assumed value of shape, we obtained this range by running simulations with different values of willingness elasticity δ. The values of δ used in the simulations was different from the set Δ used by the adaptive algorithm. Also shown is the mean revenue-ratio of the prescient algorithm for those values of δ. The starting price, ψ_0, was 20 and ψ_1 was 40, for all these simulations. As can be seen, the adaptive algorithm performs very well for each kind of workload irrespective of the assumed value of shape. Our simulations also indicate

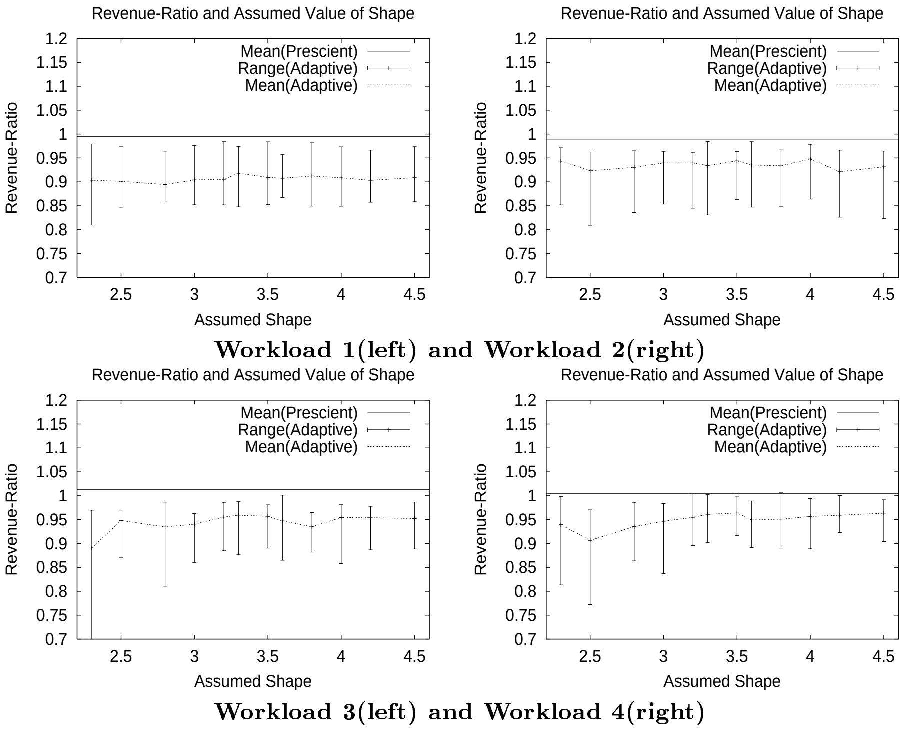


Fig. 9. Revenue-Ratio for Different Workloads

that using the adaptive algorithm results in system utilization close to that achieved by the prescient algorithm. Figure 10 shows the system utilization ahcieved in one set of simulations, with elasticity parameter $\delta = 1.6$. The predicted utilization and the observed utilization for the prescient and adaptive algorithms are plotted against the assumed value of α. Note that the presicient algorithm already has full knowledge of the parameters. Hence it is independent of the assumed value for α. As is seen in the figure, the observed system utilization when using the adaptive algorithm closely matches the system utilization for the prescient algorithm.

To study the impact of starting price on the revenue generated by the algorithm, we ran a number of simulations with different starting values. We varied ψ_0 in these simulations and always chose ψ_1 to be twice the starting price, i.e., $2 \times \psi_0$. Our simulations indicate that the performance of our algorithm is independent of the starting price. This behavior is expected because, the algorithm computes the scale for a wide range of feasible values of δ. Thus, it is able to adapt to the correct scenario no matter what the starting price. We present a representative plot of the results for each kind of workload in Figure 11. Shape, α, was assumed to be 2.0 (by the adaptive algorithm) for this set of simulations. The elasticity δ was varied. As before, we present the range of revenue-ratios we observed, when we varied δ for a given starting price ψ_0. Due to reasons of space we are unable to present a representative plot of the impact

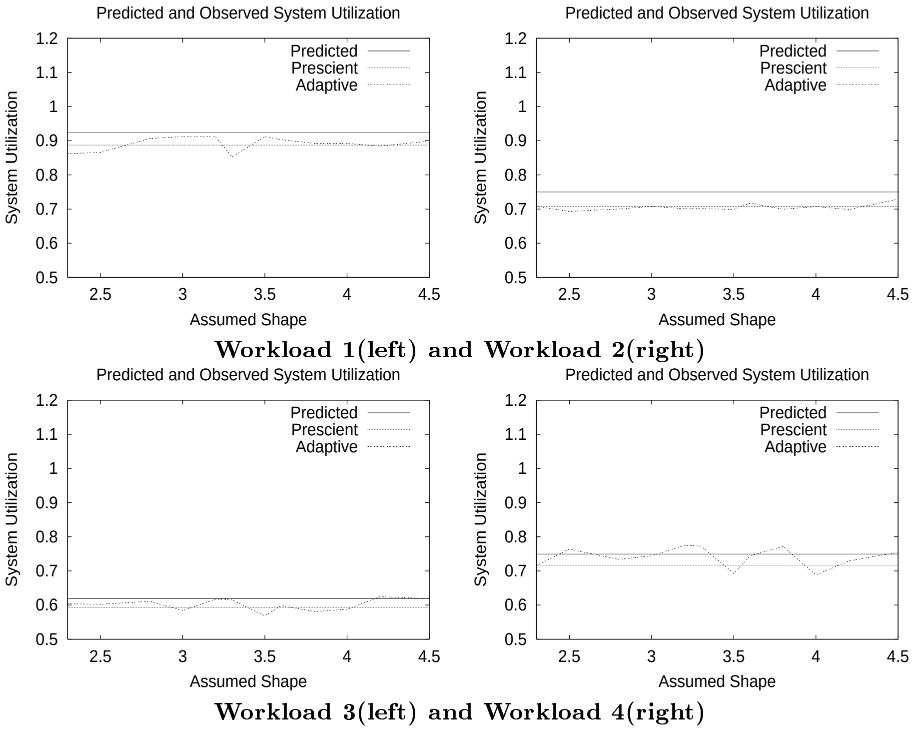
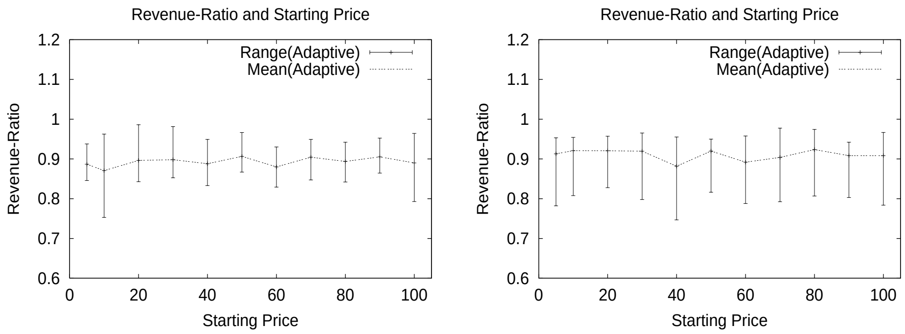


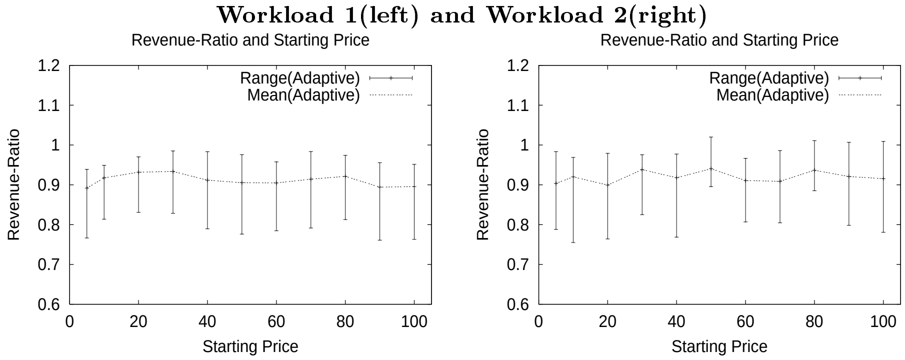
Fig. 10. System Utilization for Different Workloads

of starting price on system utilization. Our simulations indicate that, like revenue, system utilization too is relatively independent of the starting price.

In this paper, we made an assumption that all the customers conform to a single elasticity parameter δ. However, in reality, this parameter may vary from person to person. Our adaptive algorithm will be able to adapt to these variations. It will choose the closest δ which approxiamtes the overall customer behavior as a whole. To confirm that this is indeed true, we ran simulations in which the parameter δ for a customer was chosen from a uniform distribution. Note that for such customer behavior we do not know the maximum expectation of revenue. Therefore it is difficult to evaluate the performance of the algorithm. However, if the workload does not exceed the avaialble resources, according to Theorem 1, the maximum expectation of revenue is achieved when we charge a constant price. For such workloads, we can exhastively search for the optimal price using simulations. We chose a steady workload (arrival rate of 6 per minute) such that resources would be available even when δ for all customers is the maximum possible. We ran simulations to exhaustively search for the price which generates maximum revenue. We then ran simulations with our adaptive algorithm for the same customer behavior. We found that the algorithm performed remarkably well. The ratio of the revenue generated by the algorithm to the maximum revenue found using simulations is shown in Figure 12. We present the range of ratios generated when we used different assumed values for the shape, α, in the adaptive algorithm.



Workload 1(left) and Workload 2(right)



Workload 3(left) and Workload 4(right)

Fig. 11. Impact of Starting Price on Different Workloads

Though the performance in the graph does not "prove" that the algorithm is robust, it does indicate that the algorithm adapts fairly well to customer behavior.

6 Conclusions and Future Work

In this paper, we focused on pricing strategies that combine revenue goals with resource management. Specifically, we studied the impact of price on the revenue and utilization of FCFS content-delivery systems. We developed an analytical framework for maximizing revenue and quantified the relationship between price and system utilization. This relationship can be used to manage system utilization by controlling the rate at which customers purchase the content. Since the parameters governing customer behavior may not be known, we have developed an adaptive pricing scheme that tracks observable customer behavior to suggest the price. We performed simulations to validate our analytical framework and to evaluate the performance of the adaptive pricing scheme under dynamic workloads. Our simulations indicate that our algorithm is robust and generates revenue close to the maximum theoretical expectation. Therefore, with an effective dynamic pricing scheme, we now have a powerful mechanism to affect system utilization and revenue. Both of these characteristics are critical to successfully managing a content delivery service.

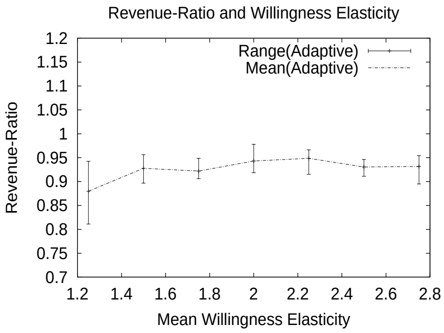


Fig. 12. Impact of Variable Willingness Elasticity

Our future work is to focus on using price to manage systems which use sophisticated scheduling mechanisms like batching. Also, the effects of content popularity and temporal changes in customer behavior need to be studied in greater detail. Content popularity has a significant impact on its *perceived value* and hence the revenue. An eventual goal of our work is to leverage content popularity and maximize revenue by allowing customer and provider to negotiate the price for the content.

References

1. D. F. Ferguson, C. Nikolaou, and Y. Y., "An economy for flow control in computer networks," in *Conference on Computer Communications*, (Ottawa, Canada), April 1989.
2. A. Ganesh, K. Laevens, and S. R., "Congestion pricing and user adaptation," in *IEEE Infocom*, (Anchorage, Alaska), April 2001.
3. J. Altman, B. Rupp, and P. Varaiya, "Internet user reactions to usage-based pricing," in *Proceedings of the 2nd Internet Economics Workshop*, (Berlin), May 1999.
4. J. Altman and K. Chu, "A proposal for flexible plan that is attractive to users and internet service providers," in *IEEE Infocom*, (Anchorage, Alaska), April 2001.
5. A. Goldberg, J. Hartline, and A. Wright, "Competitive auctions and digital goods," Tech. Rep. STAR-TR-99-01, InterTrust Technologies Corporation, November 2000.
6. P. Basu and T. Little, "Pricing considerations in video-on-demand systems," in *ACM Multimedia Conference*, (Los Angeles, California), November 2000.
7. S. Jagannathan, K. C. Almeroth, and A. Acharya, "Topology sensitive congestion control for real-time multicast," in *Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, (Chapel Hill, North Carolina, USA), June 2000.
8. S. Jagannathan and K. C. Almeroth, "Using tree topology for multicast congestion control," in *International Conference on Parallel Processing*, (Valencia, Spain), September 2001.
9. K. Almeroth, A. Dan, D. Sitaram, and W. Tetzlaff, "Long term channel allocation strategies for video applications," in *IEEE Infocom*, (Kobe, JAPAN), April 1997.
10. 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.
11. T. Little and D. Venkatesh, "Prospects for interactive video-on-demand," *IEEE Multimedia*, pp. 14–23, Fall 1994.
12. B. C. Arnold, *Pareto Distributions*. Burtonsville, Maryland: International Co-operative Publishing House, 1983.
13. E. L. Crow and K. Shimuzu, *Lognormal distrbutions: theory and application*. New York: Dekker, 1988.