

A New Pricing Model for Competitive Telecommunications Services Using Congestion Discounts

N. Keon and G. Anandalingam*

Department of Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104-6315

July 2000

Revised: June 2001

* This research was partially funded by a grant from the National Science Foundation NCR-9612781 and forms part of the PhD dissertation of the first author. We thank Roch Guerin, Nelson Dorny, and Yannis Korilis for constructive criticisms on earlier drafts of this paper. We also thank the referees of this journal for insightful comments that improved the quality of the paper. We remain responsible for any remaining errors.

Abstract

In this paper, we present a new model for using prices as a way to shift traffic from congested peak periods to non-peak periods in telecommunications networks, and hence balance the load and also ensure that almost no one is turned away (or “blocked”) from being provided service. We use the offer of congestion discounts to customers who have the choice of accepting these rebates and returning during a subsequent non-peak period, or who can reject the offer and obtain services right away. We model the problem as a mathematical program in which the network provider tries to reduce cost by minimizing total discounts offered but at the same time ensuring that almost all (i.e. 99%) of those requesting services are served. We apply this model to various scenarios and show that, except during the situations of extreme persistence of high traffic volume, the scheme would lead to zero blocking and an increase in revenue over the non-discounting case.

1 Introduction

In this paper, we propose and analyze a pricing mechanism that could be used in telecommunications networks for connection-oriented services with guaranteed quality of service (QoS). Even with the expansion of high speed networks, new services such as video-on-demand, graphics and real-time audio and video, have emerged to consume the available bandwidth of existing networks during peak periods. In the future, it is expected that public and private networks with large bandwidths will be available to consumers with guaranteed QoS. Methods for allocating bandwidth among diverse users have become an important research topic. Using economic incentives to control users' behavior, such as with pricing schemes, appears to be an effective approach to produce fair and efficient use of resources. In addition, pricing is an effective means of controlling the flow into the network, and thus managing congestion.

A network with guaranteed QoS must use a call admission policy to ensure sufficient resources are available to each connection. This results in the rejection of some connection requests. For these types of networks, the proportion of blocked connection requests is an important measure of network performance. In a systems sense, effective flow and congestion control throughout the network could minimize connection blocking. In this paper, we present an adaptive price discounting scheme that could be used as an efficient form of flow control. The basis of the discounting scheme is the allocation of connections across several time periods based on individual users' valuations of the service, and the provision of a choice to users who willingly accept discounts (or rebates) for postponing service in place of immediate service. We implement the scheme for a single service, and examine how the discount offered can be adapted to demand fluctuations, and changes in the flow of connection requests to the network.

We are developing a pricing policy to cope with fluctuating demand over a relatively short period such as a few hours. In connection-oriented networks, with guaranteed QoS, only a fixed number of users for any service can be accommodated simultaneously, each with his or her own connection. Fluctuations in demand for connection-oriented services could therefore become a critical problem unless vast capacity is installed. Having large capacity could result in it being grossly under-utilized in most periods. A method to provide users an incentive to distribute demand evenly in the aggregate is therefore desirable. Time of day price schedules may be adequate in certain markets for certain services. However, even in such cases demand forecasting errors may require a real-time control mechanism to avoid blocking a high percentage of connection requests in a busy period. We present a pricing mechanism to cope with demand

fluctuations that cannot be easily predicted. We will implement the scheme under uncertain information, requiring no advance knowledge of the demand curve or users' preferences.

1.1 Our Model

When demand exceeds capacity at a particular price (i.e. when the network is congested), the service provider is faced with one of two choices: Either "block" the new user, i.e. do not allow the user access to the network, who will likely go to another service provider, or else provide an incentive for the user to return at a time when the network is not congested. In much of the telecommunications literature, access control and blocking is used as a mechanism for flow and congestion control. In our model, we use discounts (or rebates) as incentive prices to shift user demand to another period and hence also provide congestion control. The main focus in this paper is to model this process, and to derive optimal discount rates. Obtaining discount rates is complex because there is uncertainty about the level of demand and also as to what proportion of the users will accept the discounts to shift demand.

In our paper, we assume to be working in a regime where the prevailing price for the service during any period of time is a parameter fixed outside the demand regulation problem, at least in the short-term. This is consistent with the view that the user who is "blocked" can always obtain service at the prevailing price from another service provider. The inability of any particular service provider to change actual prices can be found in situations of perfect competition, or in a situation where there is a monopolist who is price regulated.

A perfect competition model is appropriate where there are many competing service providers, each offering an identical service, with no barriers to entry, and users have the ability to change from one service provider to another. In this case, no deviation from the market-clearing price is possible. Even in cases where there are small numbers of competitors but no significant barriers to entry, a competitive price as described above may exist, based solely on the threat of new competitors entering the market.

At the other extreme of the competitive setting, a monopolist can observe effects of prices on aggregate demand since a monopolist controls the entire market. Nonetheless, monopolists often must sell at a price set outside their control, if regulators deem the monopolist's uncontrolled behavior to be harmful to the public. This was the case in long-distance telephone service in the United States prior to deregulation.

We expect to see an environment between these two extremes in future markets for consumer telecommunications services. Such an environment could be described as monopolistic competition, if there are many firms, or an oligopolistic market, where there are only a few large

competitors. In the former case, the good or service is differentiated across competitors, resulting in some brand loyalty and allowing some marginal differentiation in prices among competitors for similar services. Nonetheless, such competitors do not have complete freedom to set prices, which are often set according to marketing considerations, and a broad differentiation in prices among providers is not likely. Strategic behavior by other competitors, which may occur within monopolistic competition but is more likely in the case of an oligopoly, makes large deviations from the prices of competitors very unlikely, especially if a service provider is a market follower in a setting with few firms.

1.2 Related Literature

This paper deals with the issue of anticipating and avoiding peak traffic in telecommunications networks. Peak-load pricing has been extensively studied in both the economics and electricity pricing literature (See for example [22][27][28]). Our paper is related to this literature but is part of an emerging literature specifically concerned with communications networks.

Much of the work on pricing for packet-switched networks offering best-effort service has focused on so-called incentive compatible pricing. [14] [15]. It has also been shown through simulations that priority pricing improved network performance when there was either single or multiple service classes [2]. It has also been shown by offering a number of routes, with a corresponding set of *relative* discount rates, that a network can elicit users to select routes for data traffic according to the desired operating point of the network provider [10]. The optimal discount rates discussed in [10] can be found using an adaptive rule on-line, and are consistent with congestion pricing. Finally, in [4], the authors show that marking individual packets at congested resources allows the network to estimate the shadow prices at individual resources in a network, according to models presented in [9]. Pricing has also been offered as a means of flow control for available bit rate service in ATM [3].

A dynamic pricing mechanism was proposed in [16]. This adaptive pricing scheme assumes no knowledge of the demand function on the part of the network or the individual users. The scheme does not always converge, due to errors in users' expectations and errors in price estimates, but exponential smoothing of prices and demand estimates across periods ensures convergence for the M/M/1 queue. Dynamic priority pricing has also been studied extensively by Gupta, et al. [6][7]. They have used an innovative approach based on dynamic programming to compare dynamic pricing with fixed prices. Given the computational intractability of this model, they used simulations to perform their assignments.

Other authors have considered the pricing problem in the context of networks offering Quality of Service (QoS) guarantees. The pricing decision for a single link point-to-point integrated services network was formulated as a constrained optimal control problem and a three-stage solution procedure was developed to calculate a price schedule in [25]. A negotiation based framework for allocating network resources, using effective bandwidth as a base for pricing was proposed in [8]. In another approach to addressing the QoS issue, some authors have proposed offering network resources such as bandwidth and buffer space directly to users as part of a bidding process in [12], and subject to announced prices in [21]. In such schemes, users could achieve a desired QoS by directly purchasing access to either reserved or shared resources in the network.

While much of the pricing literature assumes users will divulge their valuations of service in a bidding process, it seems more realistic to assume that network service providers will serve the demand at a single price faced by all users for the same service. However, there may be limited ability to set prices, as market forces dictate prices in a competitive setting. In this paper, we propose a simple pricing scheme that could be used only when network congestion seems imminent. Users are offered “discounts” (or rebates) to postpone their demand for service to a less congested period. Discounts can be adjusted, under varying demand, to control the flow of connection requests to the network.

1.3 Organization of the Paper

The paper is organized as follows: In section 2, we explain the price discount model for shifting demand from high demand periods to low demand periods using discounts offered to users. We include a model for the response of users to the discount offered by the service provider, which will enable us to estimate the proportion of users who actually accept the price discount. In section 3, we present the “service model”, i.e. the queuing model at the switch which represents the decision to either serve or not serve the connection request. We provide the necessary definitions of blocking probability and the maximum arrival rate tolerable for any blocking specification, using the queuing model. This also enables us to set capacity limits for the system. We derive the optimal discounts in section 4. We present some examples, which demonstrate the effectiveness of the scheme in controlling the flow of requests to the service provider in section 5. Finally section 6 contains concluding remarks, and the appendices include the proofs as well as detailed simulation results.

2 The Price Discounting Model

2.1 Overview

We consider a case where the price for a service is determined outside the problem and fixed. The service provider can only serve a certain number of connections at one time and would prefer to shift some demand from the higher demand periods to lower demand periods in order to limit the number of customers who are refused service, i.e. blocked. In order to shift demand, the service provider offers price discounts (i.e. rebates) to the users if they will postpone the fulfillment of the service by one period. Some users will accept the price discounts and obtain their service in the next period. Some users will reject the offered discount and insist on being served right away.

Clearly the service provider would prefer that *all* “excess” customers shift their demand to non-congested periods so that *none* are blocked. However, just in case this does not happen, the service provider must choose a reasonable number, called the “blocking probability” for the proportion of requests for service that may be blocked. The provider wishes to satisfy this limit on blocking in every period. In the next section, we will examine the “service model” where we describe the arrival process, queuing model, and the service process in more detail. In this section, we will describe the discounting model in more detail and provide an expression for the proportion of users who accept the price discount or rebate.

2.2 Shifting Demand Between Periods

In each period, there is a maximum feasible rate at which requests arrive (see (17) in section 3.2), for which the probability of blocking requests is below a limit prescribed by the service provider. The demand shifting we wish to accomplish is illustrated in Figure 1.

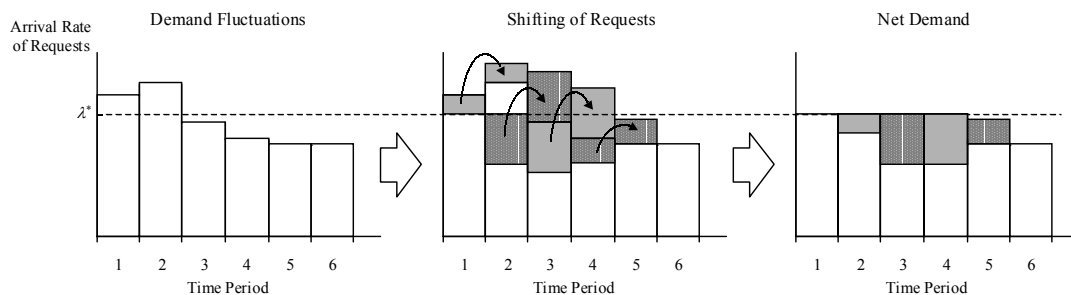


Figure 1. Demand shifting across periods.

In Figure 1, we illustrate a case where users are asked to delay service over one period. During periods 1 and 2, the rate of requests exceeds the feasible threshold of arrivals. Requests delayed from periods 1, 2, 3 and 4 are served during periods 2, 3, 4 and 5 respectively. The

delayed requests arriving from periods 1 and 2, necessitate further delayed requests from other users in periods 2 through 4. Figure 1 is a conceptual illustration and is not intended to be an accurate portrayal of the underlying queuing process.

We propose a strategy of offering price discounts or rebates to users to shift some demand. The discounts are offered as an incentive to users to delay consumption of the service. Only users who are sufficiently compensated by the discounts for their inconvenience will delay their consumption. When demand exceeds the maximum feasible level, the service provider sells the service to individual users according to the following sequence:

- An individual user requests service at the price, which is known publicly and fixed.
- A right to immediate service is sold to the user. The sale is binding for both the user and the provider.
- When congestion is imminent, a discount is offered to the individual user privately.
- The user chooses whether to relinquish the right to immediate service in period k , in exchange for a right to service at any time in period $k + 1$, at the discounted price.

In our model, we have assumed that the users will be delayed at most one period. Clearly if there is a very high arrival rate at any period, then delaying the excess arrivals by only one period will work only if the arrival rate in the next few periods is not too great. There are two implicit assumptions about the one-period delay. First, the competitive prices (see section 1.1) will take into account traffic flow and will be large enough to prevent the persistence of excess demand. Second, the definition of “period” will depend on whether or not there is peak traffic. Clearly, we are trying to move peak traffic to non-peak periods. The length of a period is an implementation issue of our price discounting strategy, and we will choose the length of a period based on how long the “peak” lasts.

The structure of the transaction is outlined in Figure 2 below:

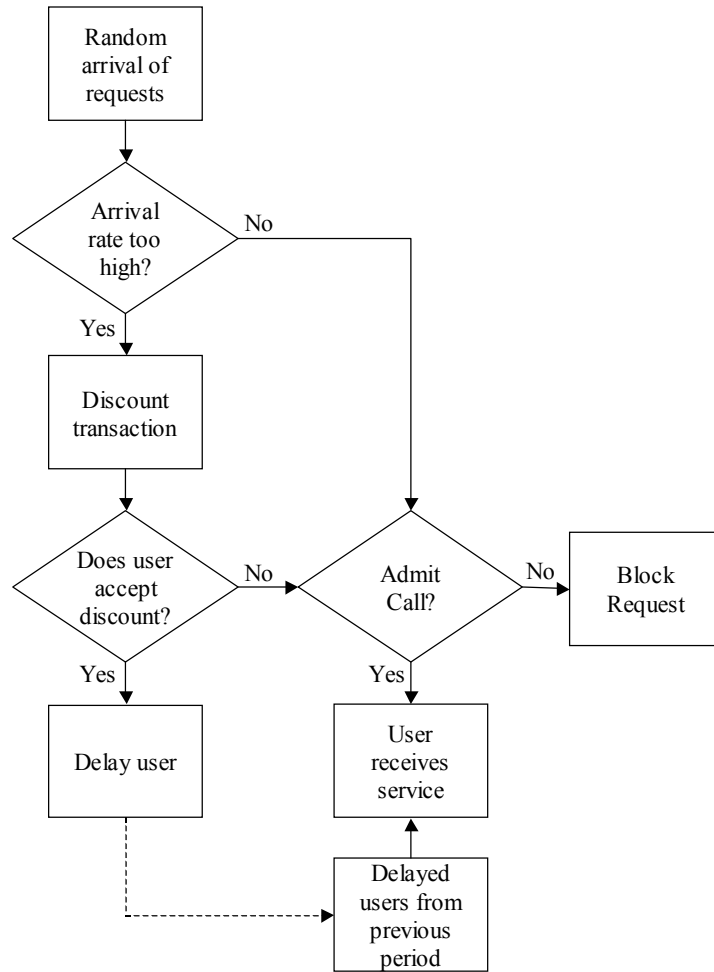


Figure 2. Congestion avoidance transaction using discounts

The discounts are offered in advance of call admission so that the likelihood of blocking is restricted. The goal of the system is to have a sufficient proportion of users accept the discount offered so that the sum total of current arrival who reject the discount and the returning users who have previously accepted discounts is limited to a level where service can be provided with acceptably high probability, e.g. 99% likely service will be provided, even in a stochastic setting where the possibility of blocking always exists.

2.3 Individual User Optimization Behavior

The discount, in return for delayed use of the service, is offered to every user requesting service in a period. Some users will accept the discount and postpone their requests, and others will refuse the discount and use the service immediately. If too many users request immediate service and the provider has to block some of the users, they have the recourse to go to alternative

service providers. Clearly, users who contract for service at a particular price may not take kindly to being blocked, and may choose legal recourse. We assume that this does not happen.

We wish to investigate the proportion of users who will choose to delay consumption of the service. In fact, the amount of the price discount has to be carefully chosen so that blocking is kept below a prescribed level at a minimum cost.

We make the following assumptions on the behavior of individual users:

- Users are unable to observe the overall level of demand, and there is no collusion among users, i.e. an individual user is uncertain whether a discount will be offered.
- Users will arrive based on the *total price* charged for service for that period alone; they do not see the price less the expected discount when they arrive. Note that this is similar to the papers by Mendelson and co-authors who have modeled arrival rate for computer and/or communications as a function of price [17][18][16]. The total price has two parts: competitive market price plus the opportunity cost of being blocked.
- Whether or not there is delay of service is entirely under the control of the user and depends on their willingness-to-pay (*WTP*). Thus, when they arrive they need not be concerned about the possibility of a delay.
- The service provider is temporally risk neutral, and treats all revenues the same.
- The inconvenience due to delay is identical for all users. Users delayed in period k are free to schedule the return for any time in the period $k + 1$.

At the prevailing price, the individual user solves a simple optimization problem:

$$\underset{u \in \{0,1\}}{\text{Max}} u(WTP - p) \quad (1)$$

$$u = \begin{cases} 1 & \text{if the user requests service} \\ 0 & \text{if the user does not request service} \end{cases} \quad (2)$$

where,

WTP = the individual user's willingness to pay for service

p = the total price = sum of competitive price and opportunity cost of being blocked

The optimal solution for the individual at the first stage is clearly:

$$u^* = \begin{cases} 1 & \text{if } WTP \geq p \\ 0 & \text{if } WTP < p \end{cases} \quad (3)$$

All users who have requested service and agreed to pay the announced price have acquired a right to the service, with a positive value equal to the individual consumer's surplus.

$$V = WTP - p \quad (4)$$

where,

V = value of individual right to service

The discount offered is a bundle, comprised of a fee and a right to the service after one period, offered in exchange for the user to relinquish his or her purchased right to service in the current time period. The discount and future right is weighed against the value of the right already purchased. In order to decide whether or not to accept the discount, the user has to solve the following optimization problem:

$$\underset{u_d \in \{0,1\}}{\text{Max}} (1 - u_d)V + u_d(\beta V + d) \quad (5)$$

$$u_d = \begin{cases} 1 & \text{if the user accepts the discount and the right to future service} \\ 0 & \text{if the user exercises the right to service immediately} \end{cases} \quad (6)$$

where,

d = the price discount (i.e. rebate) offered

β = discount factor reflecting the individual's time preference for use of the service

Substituting for V using (4), we get the individual user's optimal solution to be:

$$u_d^* = \begin{cases} 1 & \text{if } WTP \leq p + \frac{d}{1 - \beta} \\ 0 & \text{if } WTP > p + \frac{d}{1 - \beta} \end{cases} \quad (7)$$

2.4 Aggregate User Behavior: Proportion of Discounts Accepted

We now define a cumulative distribution function for willingness-to-pay (WTP). The probability that an individual has a WTP less than the price, p , is given by:

$$F_{WTP}(p) = P(WTP \leq p) \quad (8)$$

The function $F_{WTP}(p)$, must satisfy the simple properties of any distribution function, namely:

- $F_{WTP}(p) \rightarrow 1$ as $p \rightarrow \infty$
- $F_{WTP}(p) \geq 0$, for all p
- $F_{WTP}(p)$ is monotonically increasing.

The service provider treats each user identically and is interested in the probability that a user will accept the discount for postponing service, after purchasing the service at the public price.

Theorem: Let the distribution function of a user's willingness-to-pay, WTP , be given by $F_{WTP}(p)$, and the time value of consumption between periods is given by β , $0 < \beta < 1$, i.e. the value of consumption from one period to the next decreases from V to βV . Of the users who contract for service at price p , the proportion which will accept a discount, d , and a delay of service by one period is given by:

$$P(A) = \frac{F_{WTP}\left(p + \frac{d}{1-\beta}\right) - F_{WTP}(p)}{1 - F_{WTP}(p)} \quad (9)$$

Proof: See Appendix A.1.

Lemma: For a uniform distribution of users' WTP , on the interval $[a, b]$, $a \leq p \leq b$, provided $b'd \leq 1$, where $b' = 1/((1-\beta)(b-p))$, the probability of an individual user drawn at random, accepting a discount d to delay service by one period is proportional to the discount offered:

$$P(A) = b'd \quad (10)$$

Proof: See Appendix A.1.

3 The Service Model

3.1 Modeling the Connection Service

Even if the demand, λ_k , and the distribution function for WTP are known exactly and used to set price discounts, the actual number of arrivals and proportion of users who accept the discounts are stochastic and may be different from the expected values, which we use as the planning variables. Thus, some of the users who do not accept the discounts may have to be blocked. In order to state the optimization model for the service provider, we need to derive the constraint for the specification of the blocking probability. The blocking probability depends on the service model that we design for the service provider. In this section, we model a generic telecommunications service, provided to a group of users through a single switch. The service offered is the use of a connection with guaranteed quality of service (QoS). The service provider interacts with the users through the switch. Both pricing decisions and whether or not to admit the

user into the network is done at the switch. Once the user is admitted, the QoS is guaranteed. Given available bandwidth, the number of connection that can be served at any period of time with guaranteed QoS is limited, and given by c , as in Figure 3.

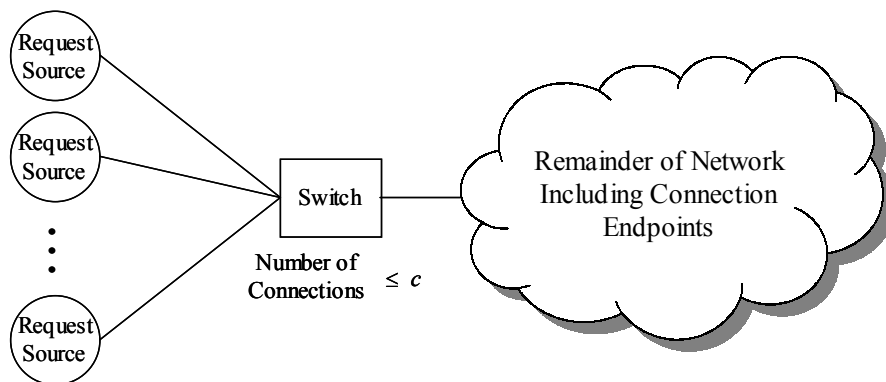


Figure 3. Service model.

We assume no scarcity of bandwidth between the switch and the users, or between the switch and the endpoints of the requested connections, i.e. the problem is a bottleneck at the local switch. Note that either the switch or the “connection endpoints” in Figure 3 would typically be referred to as servers. However, in the queuing model below, the “servers” are the c connections allowed through the switch and not the connection endpoints, from which users may be retrieving data. We have chosen this language to avoid confusion. This model could describe an ISP or perhaps a wireless voice or data service, where the bottleneck is the number of channels that can be supported in a particular cell, given the available spectrum.

We assume arrivals of user requests for service occur with exponential inter-arrival times. Requests, which arrive when the system is full, are denied access and lost to the system. Guaranteed QoS is offered by assigning a fixed amount of bandwidth to each connection. Without loss of generality, we assume that each connection requires the same amount of bandwidth. Thus, the maximum number of connections is a fixed integer, c , given by dividing the total amount of available bandwidth by the bandwidth required per connection. We assume no particular distribution on the holding time for the individual connections and that each connection is independent. This service can be modeled as an $M/GI/c/c$ queue. For a description of the properties of this queuing model, see [23]. For this system, the distribution governing the number of users in the system is the truncated Poisson distribution:

$$P[N = n] = \frac{\rho^n / n!}{\sum_{i=0}^c \rho^i / i!}, 0 \leq n \leq c \quad (11)$$

$$\rho = \lambda T \quad (12)$$

where,

N = number of ongoing connections

ρ = traffic intensity

λ = arrival rate of connection requests

T = expected holding time of a connection

The expected number of ongoing connections, $E[N]$, and the blocking probability, $B(\rho, c)$, are also known:

$$E[N] = \lambda E[T](1 - B(\rho, c)) \quad (13)$$

$$B(\rho, c) = \frac{\rho^c / c!}{\sum_{i=0}^c \rho^i / i!} \quad (14)$$

The *Erlang Loss Formula*, (14), gives the probability of blocking a request for the $M/GI/c/c$ queue.

3.2 Maximum Acceptable Arrival Rate, λ^*

Given the service model described above and its relaxation, we can now estimate the maximum acceptable arrival rate which depends on the service provider specified limit on the acceptable blocking probability, P_b :

$$B(\rho, c) \leq P_b \quad (15)$$

The probability of blocking, $B(\lambda, E[T], c)$ (14), is an increasing function of the arrival rate, λ , through the traffic intensity, $\rho = \lambda T$. We can calculate a maximum acceptable arrival rate λ^* , in order to satisfy (15), by setting the right hand side of the Erlang Loss Formula, (14) equal to P_b .

$$\rho^* = \rho \text{ such that } \frac{\rho^c / c!}{\sum_{i=0}^c \rho^i / i!} = P_b \quad (16)$$

The maximum feasible arrival rate, λ^* , is simply the maximum traffic intensity, ρ^* , divided by the average holding time, T :

$$\lambda^* = \frac{\rho^*}{T} \quad (17)$$

4 The Optimal Price Discount

In this section, we present the optimization model that is solved by the service provider in order to determine optimal discounts. We will first summarize the important conclusions of the analysis in the previous sections. In section 3, we discussed the case where the service provider treats call-blocking probability in a given period as a constraint (15) which determines the maximum feasible rate of requests, λ^* , (17). In section 2, we derived an expression for the proportion of users accepting a discount to defer service by one period; we showed that this proportion was related to the price discount through the constant b' in the case of uniform distribution of user valuations, (10).

We now require two further assumptions beyond those in sections 2 and 3:

- The holding times of the connections are relatively short compared to the scale of the time periods which exhibit peak and off-peak demand (the provider chooses the time periods for the model when implementing the proposed discount pricing scheme), e.g. if connections exhibit an average holding time of 5 minutes the peak period may be roughly 1 hour and the corresponding delay of service will be 1 hour.
- Delayed users from period k arrive during period $k + 1$, with exponential inter-arrival times. Thus, the delayed arrivals in addition to the underlying demand for the period are the sum of two Poisson processes and are in aggregate a Poisson process.

The net arrival rate of connection requests during a period is the arrival rate under the market price, less some proportion of users who accept the discount, plus delayed arrivals of users who had accepted a previous discount offer:

$$\lambda'_k = \lambda_k(1 - b'_k d_k) + \lambda_k^d \quad (18)$$

where,

λ_k^d = arrival rate of requests in period k , of users who previously accepted discounts

λ_k = arrival rate of *new* requests for immediate service

λ'_k = net arrival rate of *requests for immediate service*, after discounts offered

The service provider observes a Poisson process of arrivals, with a rate given by (18). The expected arrival rate of delayed requests in any period is a function of the discount offered in the previous period, (19). Delayed users cannot be delayed again. Only the first time arrivals, λ_k , are offered the discount, d_k .

Note that the arrival rate of delayed requests is determined by the discount offered in the previous period, d_{k-1} , and the resulting acceptance probability of users, given in (10). Multiplying

the acceptance probability by the arrivals of new requests in the previous period, λ_{k-1} , we obtain the expected arrivals of delayed requests:

$$\lambda_k^d = \lambda_{k-1} b'_{k-1} d_{k-1} \quad (19)$$

The objective function of the service provider is to minimize the total discounts paid to users. Minimizing the discounts paid to users reflects the view that prices and demand are outside the direct control of the service provider, who's primary goal is therefore to offer satisfactory service at all times by regulating blocking in all periods. Note that *expected revenue before the discounts are offered* is determined by the price (p), the arrival rate (λ), the expected holding time of a connection (T) and the proportion of requests actually blocked (B), i.e. expected revenue per unit of time before discounts = $(1-B)\lambda p T$. While λ , p and T are constants, the proportion actually blocked, B , is a complex nonlinear function that depends on a number of factors including λ , capacity c , etc., and once the discounting scheme is in place, the discount offered d , also determines B . To keep the analytics simple and the potential implementation realistic, we will focus only on reducing the cost of providing discounts. User acceptance of discounts is done period by period, and we formulate the model in a multi-period setting. The service provider minimizes the total expected costs of the discounts offered subject to the maximum feasible arrival rate. Therefore, in a given period, the optimization problem with regards to selecting the discounts is:

$$(P\text{-discount}) \quad \text{Min} \sum_{k=1}^N \lambda_k T_k d_k b'_k d_k \quad (20)$$

subject to,

$$b'_k d_k \leq 1 \quad 1 \leq k \leq N \quad (21)$$

$$\lambda_k (1 - b'_k d_k) + \lambda_k^d \leq \lambda^* \quad 1 \leq k \leq N \quad (22)$$

$$0 \leq d_k \leq p \quad 1 \leq k \leq N \quad (23)$$

The objective, (20), is to minimize the expected value of the discounts paid to users, as this quantity represents lost revenue. Note that discounts are offered to users before the system blocks any users. The first constraint, (21), restricts us to limit the expected acceptance rate of the offered discount to less than or equal to 100%. Obviously, one cannot delay more than 100% of new connection requests. The second constraint, (22), restricts the net arrival rate in each period, λ_k , to less than or equal to the maximum acceptable rate, λ^* -in each period. Finally, the discount is restricted to a non-negative range bounded by the price, (23); one cannot offer a discount more than the price itself.

Note that the for the overall optimal discount problem, (P-discount), the discount for any period k , d_k , is a function of the discounts offered in previous periods, d_{k-1}, d_{k-2}, \dots . However, we prove that the optimal solution for the problem (P-discount) can be obtained from a set of optimization problems, one for each period, $k = \{1, 2, \dots\}$, given by:

$$(P-k) \quad \text{Min } \lambda_k T_k d_k b'_k d_k \quad (24)$$

subject to,

$$b'_k d_k \leq 1 \quad (25)$$

$$\lambda_k (1 - b'_k d_k) + \lambda_k^d \leq \lambda^* \quad (26)$$

$$0 \leq d_k \leq p \quad (27)$$

Theorem: Suppose a feasible optimal solution exists for (P-discount) and is given by the vector $d^*_{discount}$. Let the feasible optimal solutions for (P-k) be given by the scalar d^*_k ($k = 1, 2, \dots, k, \dots$). Then $d^*_{discount} = \{d^*_1, d^*_2, \dots, d^*_k, \dots\}$.

Proof: See Appendix A.2.

Before we give the optimal price discounts for the problem (P-k), it should be noted that in some instances, the problem could be infeasible. Given that the price discount cannot exceed the price itself, there may be situations in which insufficient users are shifted to other periods, and the call blocking specification set by the service provider is too low to be achieved. In such cases, we suggest the best one can do is to offer a sufficient price discount to delay as many users as possible, subject to the requirement that the price discount be less than the price.

Thus, the complete statement of the optimal price discount is as follows:

$$d^*_k = \begin{cases} 0 & \text{if } \frac{1}{b'} \left(1 - \frac{\lambda^* - \lambda_k^d}{\lambda_k} \right) \leq 0 \\ \frac{1}{b'} \left(1 - \frac{\lambda^* - \lambda_k^d}{\lambda_k} \right) & \text{if } 0 < \frac{1}{b'} \left(1 - \frac{\lambda^* - \lambda_k^d}{\lambda_k} \right) < \text{Min} \left(p, \frac{1}{b'} \right) \end{cases} \quad (28)$$

$$\text{Min} \left(p, \frac{1}{b'} \right) \quad \text{if } \frac{1}{b'} \left(1 - \frac{\lambda^* - \lambda_k^d}{\lambda_k} \right) \geq \text{Min} \left(p, \frac{1}{b'} \right) \quad (29)$$

$$(30)$$

The expressions contained in (28), (29) and (30) can be understood as follows:

- If the net arrival rate at a particular price (which is set exogenously) is less than the maximum allowable arrival rate dictated by the designed blocking probability and given by condition (28), the decision is not to offer a discount, i.e. $d_k = 0$.

- If all the constraints in problem (P- k) are satisfied, given by the conditions in (29), then the price discount is simply set at the lowest feasible value. Thus, the lowest price discount is calculated using constraint (26), which has to be binding (see the proof in Appendix A.2).
- The situation where the feasible region of problem (P- k) is empty is given by the condition in (30). In this case, the lowest price discount offered will be dictated by either the exogenously set price, p , or the logical constraint of not being able to shift more than 100% of new arrivals, given by constraint (25). Clearly, the price discount cannot exceed the price itself. Thus, if $p \leq 1/b'$, the optimal discount would be p , and less than 100% of the new users will be diverted. Conversely, if $1/b' \leq p$, then 100% of the new users will be diverted by providing an optimal discount, $1/b'$, less than the price. We offer this part of the “optimal solution” only as the best of a bad situation, since in fact no feasible, much less optimal, solution exists for the situation just described.

If the calculation of solutions for the problems (P- k) is performed on-line, the number of delayed users returning is historical information in any period. Thus, we can use an exact calculation of the delayed arrival rate:

$$\lambda_k^d = \frac{nd_{k-1}}{\Delta t} \quad (31)$$

nd_{k-1} = number of users who accepted discounts in period $k-1$

Δt = length of time elapsed over one period

We use the exact number of users who accept delays in the previous period, nd_{k-1} , in (19) for the net arrival rate calculation. Note that $E[nd_{k-1}] = b'_{k-1}d_{k-1}$.

5 Simulation

We now present an implementation of the discounting scheme, applied to a few examples, similar to the situation illustrated in Figure 1, at the beginning of the paper.

5.1 Implementation

In order to implement the adaptive discounting scheme, we must set a traffic intensity, ρ^* , which is determined by the Erlang Loss formula for a given blocking specification. We will consider 10 possible values, corresponding to maximum blocking probabilities of 1% to 10%:

Parameter	Value	Comment
c	250	Max. number of connections
ρ^*	228.3	Max. traffic intensity for $P_b = 1\%$
	235.8	Max. traffic intensity for $P_b = 2\%$
	241.4	Max. traffic intensity for $P_b = 3\%$
	246.2	Max. traffic intensity for $P_b = 4\%$
	250.5	Max. traffic intensity for $P_b = 5\%$
	254.6	Max. traffic intensity for $P_b = 6\%$
	258.4	Max. traffic intensity for $P_b = 7\%$
	262.2	Max. traffic intensity for $P_b = 8\%$
	265.9	Max. traffic intensity for $P_b = 9\%$
	269.6	Max. traffic intensity for $P_b = 10\%$

Table 1. Range of maximum traffic intensities considered in simulation experiments.

We choose the traffic intensity ρ^* in Table 1 according to the Erlang loss formula, (16). The maximum arrival rate, λ^* is calculated using as a function of ρ^* , $\lambda^* = \rho^*/T$. The offered discounts in the simulations, given in expressions (28) – (30), are in turn determined as a function of λ^* .

5.2 Example Problems

We will consider four scenarios with variable levels of demand, $\lambda(k)$, and time-value of consumption, $\beta(k)$. The distribution of users' valuations of immediate service, F_{WTP} , is unknown to the service provider and the individual users, but is assumed to be a uniform random variable, distributed on the interval (0,40), throughout the simulations. The fixed price is assumed to be \$0.10 per minute (the uniform F_{WTP} in the preceding sentence is denominated in cents) per connection and the holding times are exponentially distributed with a mean holding time of 5 minutes for all four scenarios. We always use a capacity, c , of 250 connections. The first three scenarios range from an easy problem, with a short-term peak and subsequent low demand, to a difficult problem, with persistently excess demand followed by demand that is near the maximum permitted level, λ^* . The fourth scenario is randomly generated. The problem scenarios are summarized below:

Scenario	Time (hours)	Arrival Rate per Minute (λ)	Time Valuation of Consumption (β)	Proportional Acceptance Rate of Discounts (b')
Short Peak	0 – 2	55.0	0.5	0.067
	2 – 4	35.0	0.4	0.056
	> 4	35.0	0.3	0.047
Moderate Persistence	0 – 2	50.0	0.5	0.067
	2 – 4	55.0	0.4	0.056
	> 4	35.0	0.3	0.047
Extreme Persistence	0 – 2	50.0	0.5	0.067
	2 – 4	55.0	0.4	0.056
	4 – 6	45.0	0.3	0.047
	> 6	35.0	0.3	0.047
Random	0-1	54.77	0.5	0.067
	1-2	54.21	0.5	0.056
	2-3	44.19	0.4	0.047
	3-4	54.51	0.4	0.047
	4-5	49.19	0.3	0.047
	5-6	46.42	0.3	0.047
	>6	35	0.3	0.047

Table 2. Four example demand scenarios.

The demand scenarios presented in Table 2, include three scenarios designed to illustrate the potential for shifting demand between periods, as well as a randomly generated scenario. In the “Short Peak” scenario, the peak period lasts for 2 hours and is followed by low demand. The “Moderate Persistence” scenario has a slightly excessive demand for the initial 2 hours, followed by a high peak of 2 hours, but we can easily accommodate delayed users’ requests after 4 hours, when demand falls. The “Extreme Persistence” scenario again has a slightly excessive demand for the initial 2 hours, followed by a high peak of 2 hours. The demand from hours 4-6 is near peak capacity and leaves little room to accommodate the delayed users, requiring delaying more users again into the low demand periods after 6 hours. Finally, the “Random” scenario is randomly generated for hours 0 through 6 to provide an example of an unpredictable series of changes in demand. In addition to the changing demand, β decreases from 0.5 to 0.3 in each scenario, as users place more importance on consuming in the current period as the simulation progresses, reflecting an end effect, such as users may not wish to delay consumption towards the end of a day.

5.3 *An Illustrative Example*

First, we include the sample paths for a number of variables from a single simulation run. We will use the “Extreme Persistence” demand scenario contained in Table 2. This example is included to illustrate the system performance under the discounting scheme.

It is first interesting to observe the optimal discount offers.

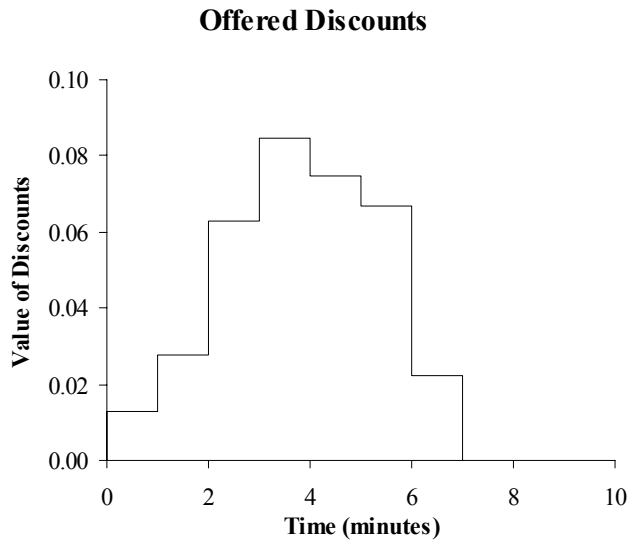


Figure 4. Optimal discount offers.

In Figure 4, we see that the price discounts offered change over time, increasing between hours 2 and 4, and decreasing between hours 5 and 7. This is due to the return of delayed users from the earlier period in each of these two-hour periods. For instance users delayed between hours 2 and 3 return between hours 3 and 4. This persistence of the peak requires even more users be delayed during the second hour of the peak.

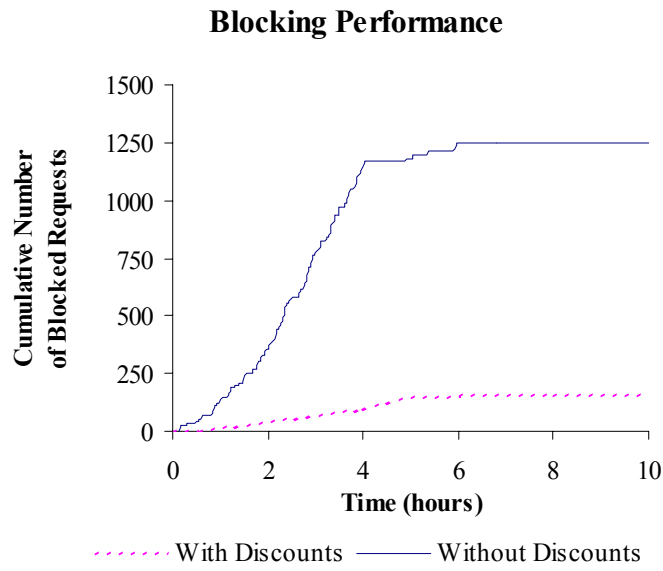


Figure 5. Comparison of Call-Blocking (Extreme Persistence Case)

Using the price-discounting scheme, it turns out that very few users are blocked even in the Extreme Persistence case (Figure 5). Between hours 0 and 6 without discounting, 6.86% of requests are blocked versus 0.84% with the discounting scheme activated. The target blocking rate in this example was set at 1%. Finally we examine the change in the system occupancy under the discounting scheme. The results are shown in Figure 6:

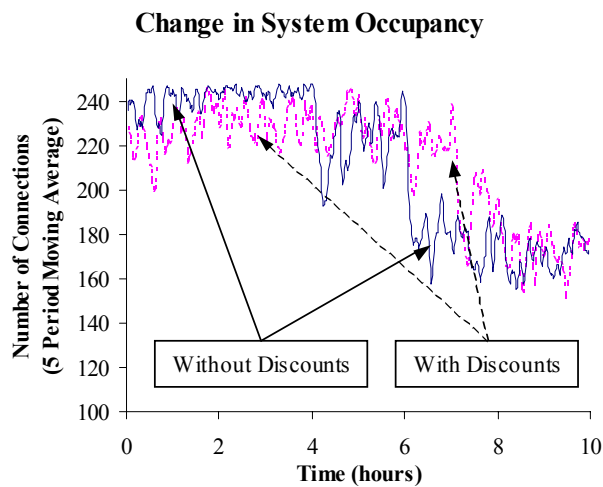


Figure 6. Comparison of number of connections with and without price discounts.

In Figure 6, we observe that without the price discounts, the system operates at a slightly higher occupancy until roughly hour 6. As demand subsides, in the later periods, the number of

connections decreases accordingly. The discounting scheme successfully shifts demand from the high demand periods to the low demand periods. Connections are slightly reduced from hours 0 to 4, but are significantly higher than without discounting in the latter portion of the simulation. The aggregate effect on system performance is to accommodate more demand and observe a more consistent level of demand over time.

5.4 Effects of Price-Discounting on Performance

We simulated a number of problems for each of the demand scenarios in Table 2. We varied the prescribed blocking probability from 1% to 10% for each of the problems above, running 1000 simulations of each again to get average performance measures. For each demand scenario we simulated the case with no discounting as a measure of baseline performance. Every simulation was initialized with a six-hour period of simulated time at the maximum arrival rate of requests per minute (e.g. for 1% cases the simulations were initialized with an arrival rate of $228.3/5 = 45.66$). This initialization begins each simulation with the system operating in a state exhibiting the maximum tolerated blocking. As stated before, we use a capacity of 250 and expected holding time of 5 minutes for all simulations. For each demand scenario under all 10 blocking specifications, we present the improvement in blocking performance, the total discounts paid out to users as a percentage of total revenue and the net revenue effect. The results are illustrated in Figure 7 - Figure 10 below:

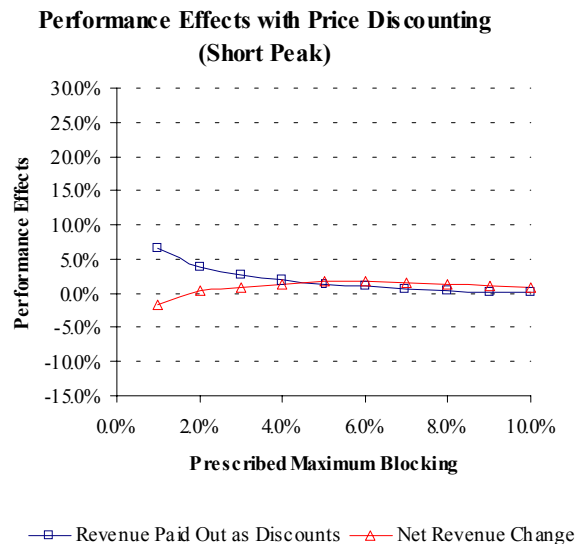


Figure 7. Performance of Price Discounting for Short Peak Demand Scenario.

As Figure 6 shows, when the blocking specification is eased from 1% to 10%, the total discounts paid out as a percentage of revenue decrease significantly, to almost zero under the 10% specification. For the short-lived peak, the discounting scheme is essentially self-financing, with almost no decrease in net revenue under any specification and small increases in net revenue in the relatively easier cases where the blocking specification is over 2%. This is because the additional revenue from serving users that would otherwise be blocked offsets the cost of providing the discount as an incentive to users to delay consumption. Figure 7 demonstrates that the service provider is better off with a discounting scheme in place than when a simple flat price is used, because better service is being offered with little or no costs to the service provider.

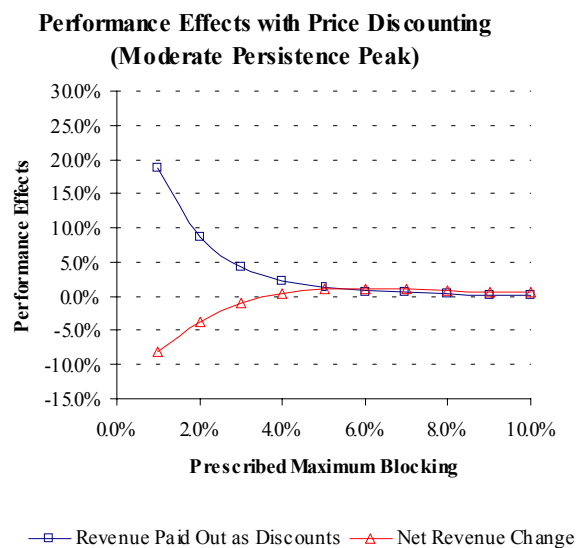


Figure 8. Performance of Price Discounting for Moderate Persistence Demand Scenario.

With the moderate persistence of the peak periods, the increased revenue derived through reduced blocking offsets more than half of the discounts paid out in the stringent blocking specification cases (1% and 2%), and the scheme becomes self-financing, even yielding small improvements in net revenue at less stringent blocking specifications. Keep in mind the relative magnitude of the excess demand is quite large (~ 20% in excess of the maximum arrival rate). The higher revenue observed in the non-discounting cases is not considered acceptable by the service provider, who wishes to regulate blocking. Indeed in the higher cost cases (1% and 2% blocking specifications) the blocking is reduced roughly 10%, which is a large improvement in performance at a reasonable cost.

The extreme persistence demand is the most difficult case considered. For low blocking specifications, revenue is severely decreased by the need to delay many users and hence offer large discounts to accommodate them later. For blocking rates roughly 4% and higher, the scheme is essentially revenue neutral. In such a case of extreme persistence of the peak demand, the problem is really that the capacity of the system is not sufficient for the level of demand over several periods. In this case the long-term solution is to either increase capacity or raise the price to restrict demand, if this is possible. However, in the short-term, the price-discounting scheme offers a method to satisfy blocking criteria.

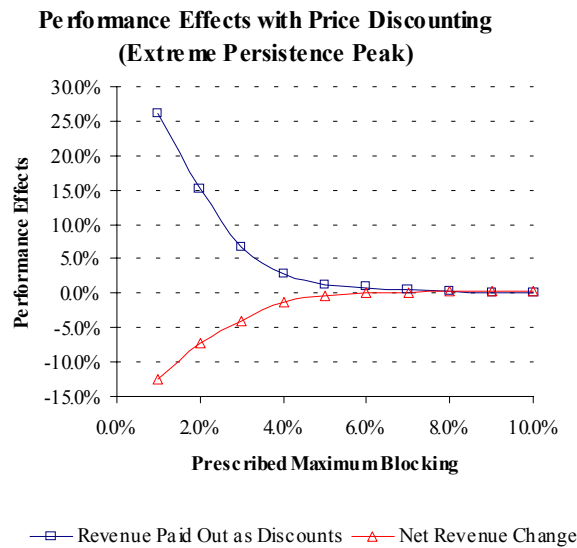


Figure 9. Performance of Price Discounting for Extreme Persistence Demand Scenario.

Performance Effects with Price Discounting (Random)

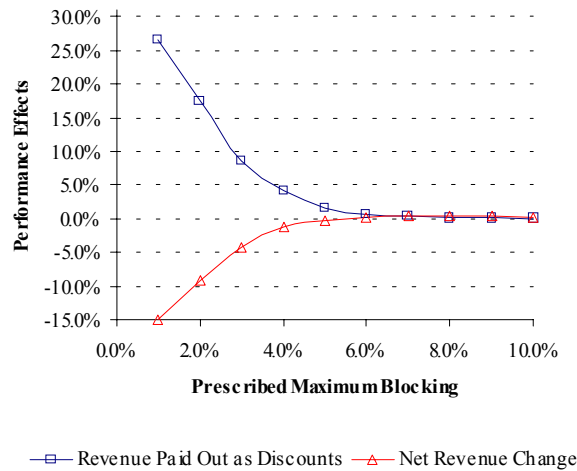


Figure 10. Performance of Price Discounting for Random Demand Scenario.

The random demand scenario is included as an example of a volatile demand scenario, where unpredictable changes occur from one period to the next. The demand range considered was uniformly distributed between 45 and 55 requests per minute for the first six periods. Again we observe a significant control effort made through the discounts when the blocking criteria is stringent, i.e. 1 or 2%. However, the scheme does become self-financing as the blocking objective becomes more realistic given the capacity and the underlying demand profile. Furthermore, desired improvement in blocking performance is achieved in each case, and in the 1 and 2% cases the demand profile considered represents a large excess demand (up to 20% in excess of capacity)

The simulation results indicate three important results:

- The discounting scheme can be used to admit more connections in every case, i.e. block fewer connections, as we expect. *The blocking specification was satisfied in every case simulated.* This represents higher market share in a competitive market.
- Revenue increases under the discounting scheme where the persistence and magnitude of peak demand is reasonable.
- In cases of extreme persistence it may happen that revenue is penalized by offering discounts in order to limit the call-blocking to a prescribed level, P_b . However, even in this case less connections are blocked, and the higher blocking observed without the discounting scheme is strictly considered an infeasible outcome, according to the provider's prescribed blocking.

We have not addressed blocking of the returning users explicitly. Delayed users who are then denied service are likely to require significant compensation. We suggest a small number of slots could be held in reserve by the provider to prevent such unfortunate occurrences.

6 Concluding Remarks

We have presented a method for calculating optimal price discounts, which are used to shift demand from congested to uncongested periods in a telecommunications system. We develop a user model of behavior so we may predict the proportion of users who will accept a price discount and delay use of the service. For problems where the peaks do not persist significantly, the discounting scheme actually increases revenue. In more difficult cases where demand persists over a long period, many users must be delayed to admit relatively few additional requests. In these cases, it may not be possible to increase revenue by the price-discounting scheme. However, we have also shown how to trade-off blocking specification with a revenue enhancing discounting scheme. The discounting scheme also provides an alternative to capacity expansion, when capacity is sufficient in all but a few periods.

The results presented here reflect the assumption that the problem parameters can be directly observed by the service provider and then used to calculate the discounts. Implementation under uncertainty, where the problem parameters are not known to the service provider, is the subject of further research with the price-discounting scheme, and will be presented in a subsequent publication.

A.1 Appendix: Proof of Discount Acceptance Theorem

Users decide to purchase service according to the decision described by (3) in section 2.3. If a user, who has purchased service, delays consumption between one and three periods, the value of the service is reduced by a factor, β , so that the individual user surplus is now $\beta(WTP - p)$. Customers drawn at random from the population, are characterized by a willingness-to-pay (WTP), with cumulative distribution function $F_{WTP}(p)$. The probability of the user accepting a discount, d , in return for delaying consumption depends on the user's decision, given in (7) in section 2.3:

$$P(A) = P\left(WTP \leq p + \frac{d}{1-\beta} \mid WTP > p\right) \quad (32)$$

$$P(A) = \frac{P\left(WTP \leq p + \frac{d}{1-\beta} \cap WTP > p\right)}{P(WTP > p)} \quad (33)$$

$$P(A) = \frac{P\left(p \leq WTP \leq p + \frac{d}{1-\beta}\right)}{1 - P(WTP \leq p)} \quad (34)$$

$$P(A) = \frac{F_{WTP}\left(p + \frac{d}{1-\beta}\right) - F_{WTP}(p)}{1 - F_{WTP}(p)} \quad (35)$$

This proves the theorem. To prove the Lemma, consider the case where WTP is uniformly distributed over the interval $[a, b]$, $a \leq p \leq b$:

$$F_{WTP}(p) = \begin{cases} 0 & , p < a \\ \frac{p-a}{b-a} & , a \leq p \leq b \\ 1 & , p > b \end{cases} \quad (36)$$

Note that the formulation of problem (P-discount) uses the definition $b' = 1/((1-\beta)(b-p))$ and requires that $b'd \leq 1$, making it easy to show that $p + d/(1-\beta) \leq b$. This is a technical requirement for the validity of the Lemma. Thus, the probability of accepting the discount is found to be proportional to the discount offered:

$$P(A) = \frac{\frac{p + \frac{d}{1-\beta} - a}{b-a} - \frac{p-a}{b-a}}{1 - \frac{p-a}{b-a}} \quad (37)$$

$$P(A) = \frac{d}{(1-\beta)(b-p)} \quad (38)$$

A.2 Appendix: Proof of Single Period Formulation (P-k) Theorem

We assume a feasible solutions exist for both problems (P-discount) and (P-k). Throughout the appendix we use the substitution $\lambda_k^d = \lambda_{k-1}b'_{k-1}d_{k-1}$ to make the relationships between periods clear.

First, we consider the set of single period problems (P-k), restated for convenience:

$$(P-k) \quad \text{Min } \lambda_k T_k d_k b'_k d_k \quad (39)$$

subject to,

$$d_k \leq \frac{1}{b'_k} \quad (40)$$

$$\lambda_k(1 - b'_k d_k) + \lambda_{k-1} b'_{k-1} d_{k-1} \leq \lambda^* \quad (41)$$

$$-d_k \leq 0 \quad (42)$$

$$d_k \leq p \quad (43)$$

Recall that that there is one problem (P-k) for each period, and the problems must be solved in order for $k = 1, k = 2, \dots, k = N$. Therefore, in each period k , the discount from the previous period, d_{k-1} , is a problem parameter and no longer a decision variable.

Consider the following solution for problem (P-k):

$$d_k = \frac{1}{b'_k} \left(1 - \frac{\lambda^* - \lambda_{k-1} b'_{k-1} d_{k-1}}{\lambda_k} \right) \text{ if } \lambda_k + \lambda_{k-1} b'_{k-1} d_{k-1} > \lambda^* \quad 1 \leq k \leq N \quad (44)$$

$$d_k = 0 \text{ if } \lambda_k + \lambda_{k-1} b'_{k-1} d_{k-1} \leq \lambda^* \quad 1 \leq k \leq N \quad (45)$$

The expression we obtain for the optimal solution to d_k , given in (44) – (45), comes from constraint (41), which requires that the discount offer, d_k , be used to regulate new arrivals, λ_k , so that the maximum acceptable arrival rate, λ^* , is not exceeded. Intuitively, the solution is easy to see. If total arrivals are below the prescribed limit the service providers does not offer a discount leaving system performance unchanged at no cost. If total arrivals exceed the prescribed limit, then the minimum discount that achieves a feasible solution is offered, satisfying the constraints and minimizing the objective, i.e. the expected payout of discounts.

The Karush-Kuhn-Tucker conditions, assuming feasibility, are:

$$2\lambda_k T_k b'_k d_k + u_k(1) + v_k(-\lambda_k b'_k) + w_k(-1) + z_k(1) = 0 \quad (46)$$

$$u_k \left(d_k - \frac{1}{b'_k} \right) = 0 \quad (47)$$

$$v_k (\lambda_k(1 - b'_k d_k) + \lambda_{k-1} b'_{k-1} d_{k-1} - \lambda^*) = 0 \quad (48)$$

$$w_k d_k = 0 \quad (49)$$

$$z_k(d_k - p) = 0 \quad (50)$$

$$u_k \geq 0, v_k \geq 0, w_k \geq 0, z_k \geq 0 \quad (51)$$

The problem, given by (39) – (43), has a quadratic and convex objective and linear constraints, so that (46) – (51) are sufficient optimality conditions. Note that the optimality conditions given by (46) – (51) apply for each problem denoted by k .

For problems where the solution is given by (50), we set $u_k = w_k = z_k = 0$, and the resulting $v_k = 2T_k d_k$ from (46) is positive. For problems where the solution is given by (51), we set $u_k = v_k = z_k = w_k = 0$. The reader may verify that these Lagrange multipliers satisfy all the K.K.T conditions.

Now consider the multi-period problem (*P-discount*), where are the above problems (*P-k*) are solved simultaneously. Again, we restate the problem under consideration:

$$(P\text{-discount}) \quad \text{Min} \sum_{k=1}^N \lambda_k T_k d_k b'_k d_k \quad (52)$$

subject to,

$$d_k \leq \frac{1}{b'_k} \quad 1 \leq k \leq N \quad (53)$$

$$\lambda_k(1 - b'_k d_k) + \lambda_{k-1} b'_{k-1} d_{k-1} \leq \lambda^* \quad 1 \leq k \leq N \quad (54)$$

$$-d_k \leq 0 \quad 1 \leq k \leq N \quad (55)$$

$$d_k \leq p \quad 1 \leq k \leq N \quad (56)$$

The K.K.T. conditions for (*P-discount*) are:

$$2\lambda_k b'_k d_k + q_k(1) + (r_{k+1} - r_k)\lambda_k b'_k + s_k(-1) + t_k(1) = 0 \quad 1 \leq k \leq N-1 \quad (57)$$

$$q_k \left(d_k - \frac{1}{b'_k} \right) = 0 \quad 1 \leq k \leq N \quad (58)$$

$$r_k (\lambda_k (1 - b'_k d_k) + \lambda_{k-1} b'_{k-1} d_{k-1} - \lambda^*) = 0 \quad 1 \leq k \leq N \quad (59)$$

$$s_k d_k = 0 \quad 1 \leq k \leq N \quad (60)$$

$$t_k (d_k - p) = 0 \quad 1 \leq k \leq N \quad (61)$$

$$q_k \geq 0, r_k \geq 0, s_k \geq 0, t_k \geq 0 \quad 1 \leq k \leq N \quad (62)$$

Similarly to (*P-k*), the K.K.T conditions, (57) – (62) are both necessary and sufficient, since (*P-discount*) is made up of a quadratic and convex objective function and linear constraints. We adopt the convention that $r_{N+1} = 0$ to keep the notation as simple as possible, rather than writing the special case for condition (57) when $k = N$.

The optimal solution to (P-discount) is calculated sequentially for $k = 1, 2, \dots, N$, using (44) – (45). However, we use a recursive argument to show that the greedy single period solution for (P-k), offered in (44) – (45), satisfies the sufficient K.K.T conditions in (57) – (62) for (P-discount). The argument below is similar to a dynamic programming argument, except we have not defined value functions to restructure the problem in the dynamic programming context. We choose to use K.K.T again, for the consistency with the argument for optimality in the single period case, (P-k). We trust that the reader finds the solution to (P-k) intuitive and that extending the optimality of the same solution to (P-discount) is therefore intuitive as well.

First we set $q_k = t_k = 0$ for all k , satisfying (58) and (61). Furthermore, for any period k , if $d_k > 0$, then $r_k > 0$ and $s_k = 0$. Alternatively, for any period k , if $d_k = 0$, $r_k = 0$ and $s_k \geq 0$. Given the form of (44) – (45) all the remaining K.K.T conditions, (57), (59), (60) and (62), will be satisfied when we show the above properties for the Lagrange multipliers to hold in every period k .

For $k = N$, and assuming a feasible solution exists for the overall problem, $d_N = 0$, since there is no later period to which demand can be delayed. As a result, $r_N = s_N = 0$ and (57), (59), (60) and (62) are all satisfied for $k = N$. For $k = N-1$, if $d_{N-1} = 0$ is calculated using (44), we set $r_{N-1} = s_{N-1} = 0$ and all the K.K.T. conditions, especially (57) are satisfied for $k = N-1$. On the other hand, if $d_{N-1} > 0$ is calculated using (45), we set $s_{N-1} = 0$ and $r_{N-1} = 2T_{N-1}d_{N-1}$, again satisfying all conditions, especially (57).

Now assuming that the K.K.T conditions are satisfied for period $k = n+1$, we will show that the general form of the Lagrange multipliers stated above holds for $k = n$ and the inductive proof of optimality for the solution (44) – (45) will be complete.

First, consider if $d_{n+1} > 0$, and hence $r_{n+1} > 0$ and $s_{n+1} = 0$. If $d_n > 0$, we set $s_n = 0$ and $r_n = r_{n+1} + 2T_n d_n > 0$ follows from (57). If $d_n = 0$, we set $r_n = 0$ and (57) implies $s_n = r_{n+1} \geq 0$. Therefore, the general form we have offered for the Lagrange multipliers is maintained when $d_{n+1} > 0$.

Now consider if $d_{n+1} = 0$, and hence we must use $r_{n+1} = 0$ and $s_{n+1} \geq 0$. If $d_n > 0$, we again set $s_n = 0$ and $r_n = 2T_n d_n$ from (57). Finally, if $d_n = 0$, setting $r_n = s_n = 0$ (≥ 0) still satisfies (57) and all the remaining K.K.T conditions. Hence the optimality of (44) – (45) for (P-discount) is proven.

References

- [1] Bhat, U.N. and Basawa, I.V., *Queuing and Related Models*, Oxford University Press, New York, New York (1992).
- [2] Cocchi, R., S. Shenker, D. Estrin and L. Zhang, "Pricing in Computer Networks: Motivation, Formulation, and Example," *IEEE/ACM Transactions on Networking*, vol. 1, No. 6, 614-27 (1993).
- [3] Courbetis, C., V. Siris and G.D. Stamoulis, "Integration of Pricing and Flow Control for Available Bit Rate Services in ATM Networks," Department of Computer Science, University of Crete, (submitted for publication) (1996).
- [4] Franklin, G.F, J.D. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, Addison-Wesley Publishing Company, Reading, Massachusetts, 3rd edition (1994).
- [5] Gibbens, R.J. and F.P. Kelly, "Resource Pricing and the Evolution of Congestion Control," draft paper (1999). (available at <http://www.statslab.cam.ac.uk/~frank/evol.html>)
- [6] Gupta, A., D.O. Stahl and A. Whinston, "A Stochastic Equilibrium Model of Internet Pricing," *Journal of Economic Dynamics and Control*, vol. 21, 699-702 (1997).
- [7] Gupta, A. D.O. Stahl and A. Whinston, "The Economics of Network Management," *Communications of the ACM*, vol. 42, No. 9, 57-63, September (1999).
- [8] Jiang, H. and S. Jordan, "The Role of Price in the Connection Establishment Process," *European Transactions on Telecommunications – Economics of Telecommunications*, November 9 (1994).
- [9] Kelly, F.P. A.K. Maullo and D.K.H. Tan, "Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability," *Journal of the Operational Research Society*, vol. 49, 237-252 (1998). (available at <http://www.statslab.cam.ac.uk/~frank/rate.html>)
- [10] Keon, N. and G. Anandalingam, "Real Time Pricing of Multiple Telecommunications Services Under Uncertain Demand," Systems Engineering Department, University of Pennsylvania, Working Paper 98-13 (1998).
- [11] Korilis, Y.A., T.A. Varvarigou and S.R. Ahuja, "Pricing Noncooperative Networks," submitted to the *IEEE/ACM Transactions on Networking*, May 1997.
- [12] Ljung, L. and T. Söderstrom, *Theory and Practice of Recursive Identification*, The MIT Press, Cambridge, Massachusetts (1983).
- [13] Low, S.H. and P. Varaiya, "A New Approach to Service Provisioning in ATM Networks," *IEEE/ACM Transactions on Networking*, vol 1, no. 5, 547-553 (1993).
- [14] Mackie-Mason, J.K. and H. Varian, "Pricing the Internet," in: *Public Access to the Internet*, eds. B. Kahin and J. Keller, Cambridge and London: MIT Press, 269-314 (1995).

- [15] Mackie-Mason, J.K. and H. Varian, "Some Economics of the Internet," in: *Networks, Infrastructure and the New Task for Regulation*, eds. W. Sichel and D.L. Alexander, Ann Arbor: University of Michigan Press, 107-36 (1996).
- [16] Masuda, Y. and S. Whang, "Dynamic Pricing for Network Service: Equilibrium and Stability," *Management Science*, vol. 45, no. 6, 837-859 (1999).
- [17] Mendelson, H. "Pricing Computer Services: Queuing Effects", *Communications of the ACM*, vol. 28, pp 312-321 (1985)
- [18] Mendelson, H. and S. Dewan, "User Delay Costs and Internal Pricing for a Service Facility", *Management Science*, Vol. 6, pp 1502-1517 (1990).
- [19] Papazian, E. (ed.), *TV Dimensions '99*, Media Dynamics, Inc. New York, NY, (1999).
- [20] Patek, S., "Regulation of Packet-Switched Networks: Stochastic Optimal Control Models and Computational Methods," NSF CAREER Grant Proposal, August (1998).
- [21] Ross, S.M., *A First Course in Probability*, MacMillan College Publishing Company, New York, New York, 4th ed. (1994).
- [22] Steiner, P.O., "Peak Loads and Efficient Pricing," *Quarterly Journal of Economics*," Vol. 71, No. 4, 585-610 (1957).
- [23] Takagi, H., *Queuing Analysis: A Foundation of Performance Evaluation*, Elsevier Science Publishers, New York, New York (1993).
- [24] de Veciana, G. and R. Baldick, "Resource Allocation in Multi-Service Networks via Pricing: Statistical Multiplexing," *Computer Networks and ISDN Systems*, vol 30, 951-962 (1998).
- [25] Varian, H., *Microeconomic Analysis*, W.W. Norton & Company, New York, New York, 3rd ed. (1992).
- [26] Wang, Q., J.M. Peha, and M.A. Sirbu, "The Design of an Optimal Pricing Scheme for ATM Integrated Services Networks," Special Issue: Internet Economics, *Journal of Electronic Publishing*, University of Michigan Press, vol. 2, no. 1 (1996).
- [27] Williamson, O. "Peak Load Pricing and Optimal Capacity Under Indivisibility Constraints", *American Economic Review*, Vol. 56, No. 4, pp 810-827 (1966).
- [28] Wilson, R. B. *Nonlinear Pricing*, Oxford University Press (1992).