Optimal Link Allocation and Revenue Maximization

Jyrki Joutsensalo and Timo Hämäläinen

Abstract: **In this paper, the maximal capacity of the data network link has attempted to be exploited by using the dynamic allocation strategy. We propose a new methodology based on the economic models for competing traffic classes (classes of sessions) in packet networks. As the demand for network services accelerates, users' satisfaction to the service level might decrease due to the congestion at the network nodes. To prevent this, efficient allocation of a networks resources, such as available bandwidth and switch capacity, is needed. By using the so-called user profile as well as the utility (e.g., data rate) functions, it is possible to allocate data rates and other utilities using the arbitrary number of QoS classes, say $\$0.01, \cdots, \$10$.**

Index Terms: **Quality of service, account management, performance management, link allocation.**

I. INTRODUCTION

This paper presents a model on how to select user applications, so that the link network revenue can be maximized. Our pricing scheme is suitable for different kinds of network situations in which the user population uses dynamically the same limited resource (e.g., switching bus, buffer, link). Here we consider a single link.

Several revenue maximization models for the multiservice networks have been presented in recent years. Our channel allocation model has similar features as used e.g., at [1]–[11]. One pricing approach proposed by several researchers is that a small number of service classes should be offered on the network [11]–[13].

A congestion-dependent pricing model is presented at [1]. They study optimal or near–optimal pricing schemes by using a decision–theoretical framework under an explicit model of users' reaction to demand functions. This work also relates to the problems of admission control in loss networks.

A method of charging bursty traffic sources based on their effective bandwidth is studied at [2], [3]. The effective bandwidth is a function of the traffic profile of a source and provides a good measure upon which to base pricing decisions. They also propose the pricing of real-time traffic with QoS requirements and provide approximations that only involve time and volume charges.

The pricing of a single network which provides multiple services at different performance levels is analyzed at [5]. They have presented a good example which shows that in comparison with flatrate pricing for all services, a price schedule based on performance objectives can enable every customer to derive a higher surplus from the service, and at the same time, generate bigger revenue for the service provider. Paper [6] describes another scheme for packet–based pricing as an incentive for more efficient flow control. The emergence of real–time traffic substantially complicates the picture and requires QoS measurements much harder to analyze [4], [9].

A single queuing model in which the network is formulated as a server or servers with finite capacity, and consumers demand the same service from the server but vary in both willingness to pay for the service and tolerance for delay is presented at [7], [8].

In the optimal pricing models mentioned, the fact that different applications may have different performance objectives was not properly considered. Also many of the above models assume that the prices are fixed and are only concerned with admission decisions. In this paper, we assume that the number of service classes can be very large, in principle infinity. Other important features are:

- Our channel allocation system accepts every call. More precisely, there is no absolute blocking, but when the network load is low (high), the data rate/user is high (low).
- Our data arrival rate depends on the current link utilization and user's payment (selected QoS class). The arrival rate is (i) increasing with respect to the offered data rate, (ii) decreasing with respect to the price, and (iii) decreasing with respect to the network load. As an example, an explicit formula obeying these conditions is given and analyzed.
- Both exponential and Pareto duration traffic models are considered.

The remainder of the paper is organized as follows. In Section II, we describe our link allocation scenario. In Section III, we introduce our traffic model. In Section IV, analytic approximation for the traffic model is represented. Section V illustrates some simulations and results with a numerical example and finally, we conclude with some remarks about open issues.

II. ALLOCATION SCENARIO

Let A be the customer and B be the operator or manager network administrator. *Price* x streams from A to B. The price paid by A is normalized to be

$$x \in (0, 1). \tag{1}$$

User profile function $n(x, t)$ depends on both price x and time t, and it has the property

$$n(x, t) > 0, \tag{2}$$

Manuscript received April 27, 2001; approved for publication by J. J. Boutros Division I Editor, November 15, 2001.

The authors are with Faculty of Information Technology, Department of Mathematical Information Technology, Agora, Mattilanniemi, P.O. Box 35, FIN-40351 Jyväskylä, FINLAND, e-mail: jyrkij@mit.jyu.fi, timoh@mit.jyu.fi.

1229-2370/02/$10.00

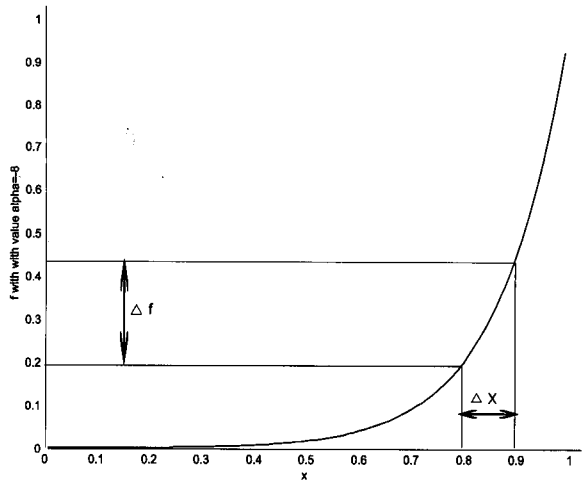


Fig. 1. Convex function f: Differential utility is large for high QoS class.

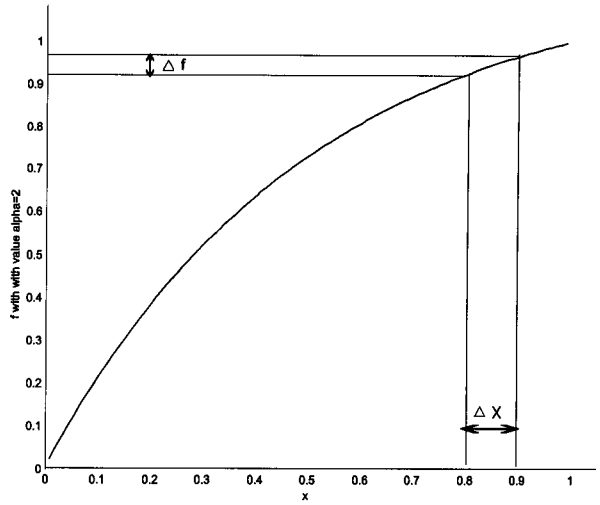


Fig. 2. Concave function f: Differential utility is small for high QoS class.

for all defined values of x. $n(x_0, t_0)$ tells us, how many users pay x_0 money units at time instant t_0. Positivity of the function means that A pays money to B, but the opposite price stream is not possible in this study. $n(x,t)$ is defined for continuous $x \in (0,1)$, but in a practical application, $n(x,t)$ is sampled and vectorized.

The definition of a *data rate function* - which belongs to the class of *utility functions* - $u(x,t)$ is the following. When the price x streams from A to B, the data rate $u(x,t)$ streams between the corresponding users A via B. Data rate $u(x,t)$ depends on x, and is a strictly increasing function of x. Generally speaking, the utility is a service rate, or Quality of Service (QoS). In telecommunications applications, the utility may include e.g., data rate, Bit Error Rate (BER), delay, or blocking probability. Here we assume that $u(x,t)$ is scalar with respect to scalar x and t.

B tries to keep all the capacity of the channel in use, so that the customers A are as satisfied as possible. By doing this B also maximizes his revenue. *Channel capacity* C is the maximal number of information that the entire network B can transfer without any errors. This is the basic definition given by the classical Shannon information theory [14]. In practice, the suboptimal estimate of C is obtained by observing the data flow history in the network. In this paper, we do not take a stance on the strict estimation procedure of the channel capacity.

In our scenario the operator offers the data rate as follows:

$$u(x,t) = \gamma(t) f(x, \alpha_1, \cdots, \alpha_n). \tag{3}$$

Here the basic function $f(x) = f(x, \alpha_1, \cdots, \alpha_n)$ has the following properties:

- $f(x)$ strictly increases with respect to price x.
- $f(0) = 0$, i.e., no utility is allocated when money is not paid.
- Without loss of generality, the maximum value is assumed to be $f(1) = 1$.
- The parameters α_i are optimized for maximizing the revenue.

There are infinitely basic functions which satisfy the above conditions. One possible basic function - which we have selected as a test case - has the form

$$f(x,\alpha) = \frac{1 - e^{-\alpha x}}{1 - e^{-\alpha}}, \tag{4}$$

where α is a parameter to be optimized for maximizing the revenue. (4) obeys $f(0) = 0$ and $f(1) = 1$, and is strictly increasing. Parameter α controls the behavior of increasing:

- when $\alpha < 0$, f is convex,
- when $\alpha > 0$, f is concave, and
- $\lim_{\alpha \to 0} f(x,\alpha) = x$.

Fig. 1 and 2 illustrate two different basic functions $f(x, \alpha = -8)$ and $f(x, \alpha = 2)$.

The first function shown in Fig. 1 is very convex, and a consequence is that the differential utility $u(\Delta x, t) = \gamma(t) f(\Delta x, -8)$ achieved by paying Δx money units more is large for high QoS classes, while in the concave case shown in Fig. 2 it is noticeably smaller. The question how to select optimal form of f depends on the distribution of the QoS classes, which is discussed in the following two sections.

We also assume that all connection requests are accepted, and the overall data rate will be kept as high as possible. In other words, the equality

$$\int_0^1 u(x,t) n(x,t) dx = 1 \tag{5}$$

is attempted to be kept at all time t. Here the capacity C is normalized to be $C = 1$. Then it follows from (3) and (5) that

$$\gamma(t) = \frac{1}{\int_0^1 f(x,\alpha) n(x,t) dx}, \tag{6}$$

and the customer, who pays the price x at the time t, gets the data rate

$$u(x,t) = \frac{f(x,\alpha)}{\int_0^1 f(y,\alpha) n(y,t) dy}. \tag{7}$$

Notice that the given data rate dynamically varies with respect to the time t due to the existence of the user profile function $n(y,t)$ in the denominator of (7). When $n(x,t)$ - i.e., the number of the users - is small, the integral in the denominator of (7) is small, and as a consequence, the offered data rate/user class is high. On the other hand, in the highly loaded systems, the offered data rate/user class is low.

III. TRAFFIC MODEL

In our traffic model, new calls arrive according to the Poisson process with parameter $\lambda(x,t)$. $\lambda(x,t)$ is a function of several QoS parameters such as data rate, delay, or BER. The general model in this study is

$$\lambda = \lambda[x,t,\alpha_1,\cdots,\alpha_n,\gamma_1(t),\cdots,\gamma_m(t)], \tag{8}$$

where

- $\lambda > 0$.
- α_i are called *hidden QoS parameters* to be optimized. That is, the customers do not observe those parameters.
- $\gamma_i(t)$ are called *observed time-varying QoS parameters.* That is, users observe those parameters, for example $\gamma(t)$ in (6). More precisely, the customer knows the offered time-variant and time-invariant utility functions which contain the information about $\gamma(t)$. For example, time-variant $u(x,t)$ and time-invariant $f(x,\alpha)$ are known, and so $\gamma(t) = u(x,t)/f(x,\alpha)$ can be observed if necessary. $f(x,\alpha)$ is available in tabular or explicit functional form.

One special case of the general form is

$$\lambda(x,t) = \frac{p(x,\alpha_1,\cdots,\alpha_n)}{1+\int_0^1 e(y,\beta_1,\cdots,\beta_m)n(y,t)dy}, \tag{9}$$

where $p > 0$ and $e > 0$ are some, perhaps very complicated, functions of the offered QoS parameters. Positivity of p and e implies positivity of λ. The integral in the denominator of (9) is a feedback term which increases with respect to $n(x,t)$, ie. number of users (due to the positivity of e). Thus it controls increasing of λ due to the fact that the data rate $u(x,t)$ in (7) depends inversely on $n(x,t)$.

In this paper, we study the simple case where the arrival rate depends on the following issues:

(i) The larger the basic function $e(x,\alpha)$ is offered, the larger the call density.
(ii) The larger the price x is, the smaller the call density.
(iii) The larger the load is, the smaller the call density.

In this special case, we select

$$p(x,\alpha,\xi) = kf(x,\alpha)h(x,\xi), \tag{10}$$

$$e(x,\alpha) = f(x,\alpha), \tag{11}$$

and $\lambda(x,t)$ could then be as in the following equation:

$$\lambda(x,t) = \frac{kf(x,\alpha)h(x,\xi)}{1+\int_0^1 f(y,\alpha)n(y,t)dy}. \tag{12}$$

Here k is a positive scaling constant, and $h(x,\xi)$ is some function which decreases with respect to the price. In practice, some *a priori* information about the behavior of the customers within different QoS classes is useful for determining the estimate of $h(x,\xi)$.

The model of $\lambda(x,t)$ is decomposed into three subfunctions. In model (12), $f(x,\alpha)$ represents feature (i), i.e., the arrival rate increases with respect to the offered data rate. Typically $h(x,\xi)$, which represents feature (ii), depends on the distribution of the richness or the willingness of pay of the users, and therefore the general properties of $h(x) = h(x,\xi)$ are:

- $h(x)$ strictly decreases with respect to the price x.
- Without loss of generality, $h(0) = 1$, i.e., the willingness of pay is maximum.
- Without loss of generality, the minimum is at $h(1) = 0$.
- The parameter ξ controls the behavior of the curve.

In this study, we have for simplicity the exponentially decreasing convex form

$$h(x,\xi) = \frac{e^{-\xi x} - e^{-\xi}}{1-e^{-\xi}}, \tag{13}$$

where $\xi > 1$. Figs. (4) and (10) show two examples of h. For large ξ, the function $h(x,\xi)$ is abrupt, ie. there are a lot of users who will pay less money, compared to those users who will pay a lot of money.

It is natural that in our scenario the denominator of (12) consists of the term $\int_0^1 f(y,\alpha)n(y,t)dy$, since that term includes information about the offered data rate per user class, as shown in (7). That integral represents the feature (iii). Notice that the arrival rate can also be written in the form

$$\lambda(x,t) = \frac{kf(x,\alpha)h(x,\xi)}{1+1/\gamma(t)}, \tag{14}$$

which shows the dependence of the arrival rate with respect to the price, time, and the given data rate per QoS class. When no users exist in the network, the maximal arrival rate

$$\lambda_{\max}(x,t) = kf(x,\alpha)h(x,\xi) \tag{15}$$

is achieved.

In our work, the duration of the connection is either exponentially or self-similarly (e.g Pareto) distributed. The cumulative distribution function for the exponential duration is

$$F(t) = 1 - e^{-\mu(x)t}, \tag{16}$$

where $\mu(x)$ is a time parameter depending on the price. For the Pareto duration, the expected connection time $E(connection)$ is

$$E(connection) = \frac{p_a p_b}{p_a - 1}, \tag{17}$$

where $p_a > 1$ and p_b specifies the minimum value that the random variable can take, i.e., the minimum duration. The cumulative distribution function for the Pareto data is

$$F(t) = 1 - (p_b/t)^{p_a}. \tag{18}$$

IV. STABLE STATE APPROXIMATION

Here we will study the case more analytically where the system is in a steady state or near it, and $\lambda(x,t)$ changes sufficiently and slowly according to the arrival model (12). The division operation is not linear, but in the stable state analysis it is assumed that the integral in the denominator changes so slowly (and it is indeed the case in the simulations), that we can make an approximation

$$\begin{aligned} & \mathrm{E}\left\{\frac{1}{1+\int_0^1 f(x,\alpha)n(x,t)dx}\right\} \\ \approx\ & \frac{1}{\mathrm{E}\{1+\int_0^1 f(x,\alpha)n(x,t)dx\}}. \end{aligned} \tag{19}$$

We apply the Little's formula [15] in two ways. The first, weaker condition for the expected number of users in the network

$$E(n(x,t)) = \frac{E(\lambda(x,t))}{\mu(x)} = \frac{\lambda(x,t)}{\mu(x,t)} \tag{20}$$

is used in the simulation, while the second, stronger condition

$$E(n(x,t)) = \frac{E(\lambda(x,t))}{\mu(x)} = \frac{\lambda(x)}{\mu(x)} \tag{21}$$

is used in the theoretical analysis. The above assumptions are valid for a large number of different types of stochastic processes, when λ is the arrival rate and $1/\mu$ is the expected value of the connection time. In the exponential scenario, $\mu = \mu(x)$, and in the Pareto scenario, $1/\mu$ is replaced by $E(connection)$. In the latter case, the steady-state equation is

$$E(n(x,t)) = \lambda(x)E(connection). \tag{22}$$

The expected revenue per time of the system is

$$\begin{aligned} revenue &= \int_0^1 xE[n(x,t)]dx \\ &= \lim_{T\to\infty}\frac{1}{T}\int_0^T\int_0^1 xn(x,t)dxdt, \end{aligned} \tag{23}$$

and by using Little's formula, it is

$$revenue = \int_0^1 x\frac{\lambda(x,t)}{\mu(x,t)}dx, \tag{24}$$

while in the theoretical steady state, if achieved, it is

$$revenue = \int_0^1 x\frac{\lambda(x)}{\mu(x)}dx, \tag{25}$$

in the exponential scenario, and

$$revenue = \int_0^1 x\lambda(x)E(connection)dx, \tag{26}$$

in the Pareto scenario. On the other hand, in the steady-state condition or near it one can approximate (12) by using formulae (20) and (21) as follows:

$$\begin{aligned} \lambda(x) &= \mathrm{E}[\lambda(x,t)] \approx \frac{p(x)}{1+\int_0^1 f(y)\mathrm{E}[n(y,t)]dy} \\ &= \frac{p(x)}{1+\int_0^1 f(y)\frac{\lambda(y)}{\mu(y)}dy}, \end{aligned} \tag{27}$$

where

$$p(x) = kf(x,\lambda)h(x,\xi). \tag{28}$$

By defining

$$s = \frac{1}{1+\int_0^1 f(y)\frac{\lambda(y)}{\mu(y)}dy}. \tag{29}$$

we get

$$\lambda(x) = sp(x). \tag{30}$$

Because

$$\frac{1}{1+\int_0^1 f(y)\frac{\lambda(y)}{\mu(y)}dy} = \frac{1}{1+s\int_0^1 f(y)\frac{p(y)}{\mu(y)}dy}, \tag{31}$$

then we obtain

$$s = \frac{1}{1+sq}, \tag{32}$$

where

$$q = \int_0^1 f(y)\frac{p(y)}{\mu(y)}dy. \tag{33}$$

(32) has the solution

$$qs^2 + s - 1 = 0. \tag{34}$$

(34) has in principle two solutions, but the positivity of λ and p in (30) implies the positivity of s, and thus

$$s = \frac{-1+\sqrt{1+4q}}{2q}. \tag{35}$$

From (28) and (33) we get

$$q = kr, \tag{36}$$

where

$$r = \int_0^1 f(x,\alpha)^2\frac{h(x,\xi)}{\mu(x)}dx. \tag{37}$$

Thus, if $h(x,\xi)$, k and $\mu(x)$ are known, we can numerically find the optimal α that maximizes revenue (25) by using formulae below.

$$revenue = \int_0^1 x\frac{\lambda(x,\alpha)}{\mu(x)}dx, \tag{38}$$

where

$$\lambda(x,\alpha) = \frac{-1+\sqrt{1+4kr}}{2kr}f(x,\alpha)h(x,\xi) \tag{39}$$

is obtained from (30), and

$$r = \int_0^1 f(x,\alpha)^2\frac{h(x,\xi)}{\mu(x)}dx. \tag{40}$$

Thus, the revenue maximization procedure in the stable state scenario is as follows:

- $h(x,\xi)$, k, and $\mu(x)$ are given.
- Vary α within some interval.
- Evaluate $f(x,\alpha)$.
- Evaluate (38)–(40).

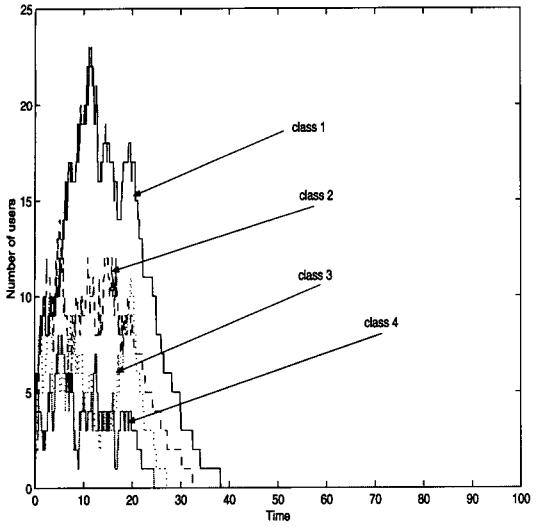


Fig. 3. Number of users in four QoS classes as a function of time. The price parameter is $\xi = 2$.

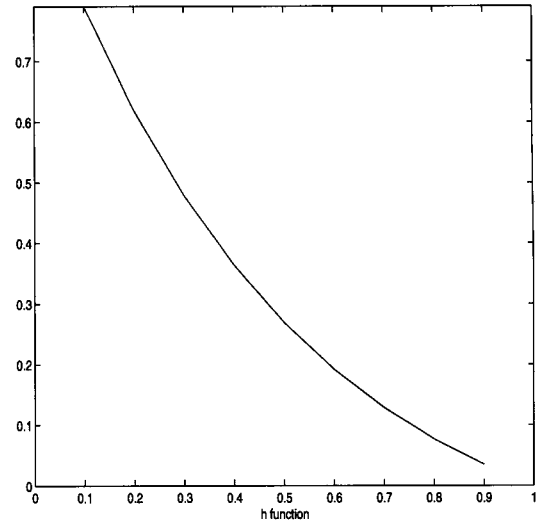


Fig. 4. The relationship of the price x and the price depending function $h(x, \xi)$ for $\xi = 2$. The curve is gently exponentially decreasing.

- Find that α_{opt} and $f(x, \alpha_{\text{opt}})$ which maximizes the revenue (38).

When more general functions $p(x)$ and $e(x)$ in the arrival model (9) are used, the stable state approximation is as follows:

$$\lambda(x) = \frac{-1 + \sqrt{1 + 4q}}{2q} p(x) \tag{41}$$

$$q = \int_0^1 e(x) \frac{p(x)}{\mu(x)} dx. \tag{42}$$

V. SIMULATIONS AND RESULTS

In the simulations, we have compared the revenue (24) given by the simulated network traffic data with that revenue (38) given by the analytical model. Our goal is to examine which kind of shape the capacity function $f(x, \alpha)$ in (4) could be optimal for maximizing the revenue under two arrival rate scenarios: In the more general case (9), and in the special case (12). Our other goal is to prove that the analytical models in the special case (38)–(40) as well as in the more general case (38), (41), (42) hold well enough. If these models hold, then one can easily construct the revenue curves corresponding to the different data traffic scenarios and utility function families, and therefore obtain estimates for the optimal data rate allocation strategies.

A. Exponentially Distributed Duration

Experiment 1. In the first simulation, the parameters were as follows:

- Positive constant from (12) was $k = 200$.
- The parameter that determines the dependence of the arrival rate on the price was $\xi = 2$. See (13). This parameter is associated with the feature (ii) in the traffic model.
- Number of service classes is 9, i.e., in our formalism it is possible to pay $x = 0.1, 0.2, \cdots, 0.9$ money units.
- Duration of each connection was exponential with the time parameter

$$\mu(x) = \frac{e^{-2x} - e^{-2}}{1 - e^{-2}}. \tag{43}$$

Thus, the larger the price is, the smaller the connection time.

- T is the simulation time, $T = 0, \cdots, 20$ time units. The time resolution was 0.02 time units, and thus 4000 events per simulation were considered.
- The number of simulations per α was 100.
- Simulations showed that the revenue had only one local (i.e., global) maximum, and thus α was varied from -10 to 10.

Fig. 3 shows the evolution of $n(x, t)$, i.e., the number of users, within different QoS classes. To clarify the figure, only four classes are shown. In Fig. 3, the price depending parameter in the function $h(x, \xi)$ is $\xi = 2$. In this figure, the parameter related to the given data rate is $\alpha = 10$. After $T = 20$ time units, the remaining calls are still connected, but no more calls are arrive, as shown in Fig. 3. The curves and parameters in the simulation study were estimated within the time interval $t \in [0, 20]$ units of time. From this figure we can see that there are lots of small users who pay less money ($x = 0.1$ in class 1), compared to the number of users that pay much money ($x = 0.9$ in class 4).

Fig. 4 shows the relationship of the price x and the price depending function $h(x, \xi)$ for the value $\xi = 2$. When $\xi = 2$, the curve is quite gentle, because $h(x, \xi)$ tells how the arrival rate depends on the price, it is assumed here that the distribution of the richness (or willingness to pay) is not sharply exponential.

Fig. 5 shows how the arrival rate converges fast to the steady state, or near it. At the beginning of the simulation, the integral in the denominator of (12) is zero (because there are no users yet), and $\lambda(x, t)$ achieves the maximum

$$\lambda(x, 0) = k f(x, \alpha) h(x, \xi) = 200 f(x, 10) h(x, 2) \tag{44}$$

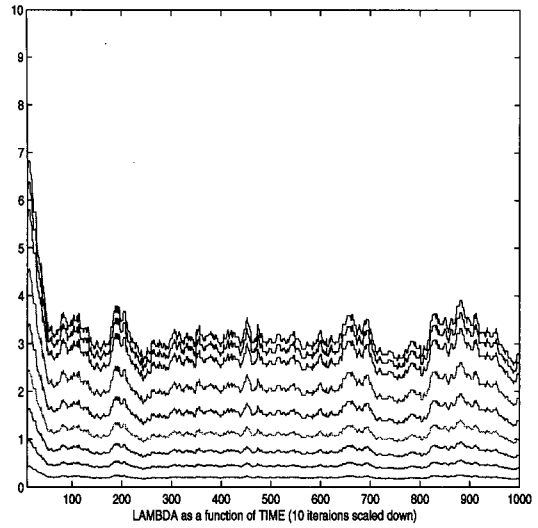


Fig. 5. Behavior of the arrival rate $\lambda(x,t)$, when the first ten samples have been cut. Different curves represent different classes and prices, the uppermost curve the cheapest class, and the lowest curve the most expensive class.

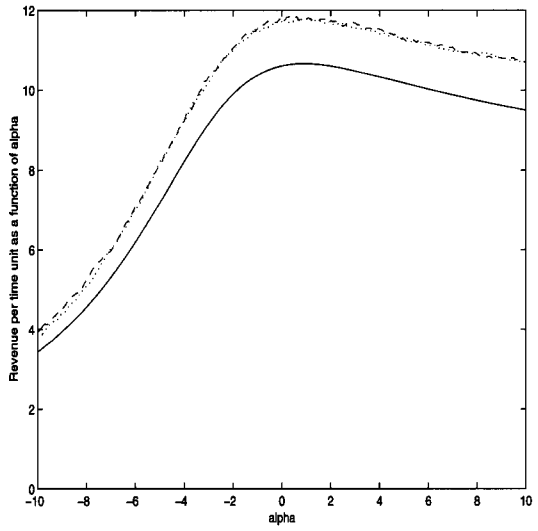


Fig. 6. Revenue per time unit as a function of α when $\xi = 2$. The optimal α is about $\alpha \approx 0.1$.

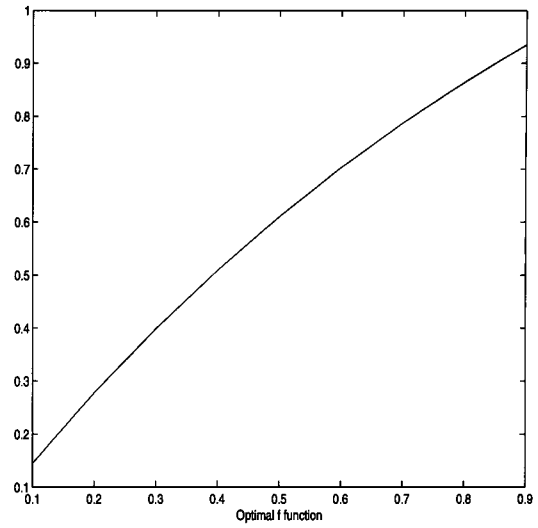


Fig. 7. Optimal f function based on the analytic model, when $\xi = 2$.

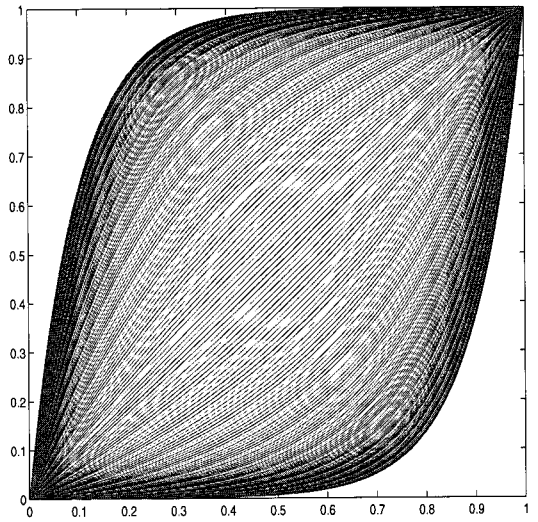


Fig. 8. The family of the basic capacity functions $f(x,\alpha)$ as a function of price x. The lowest i.e., the most convex function is corresponding to the value $\alpha = -10$, while the uppermost i.e., the most concave function is corresponding to the value $\alpha = 10$.

and the largest value is approximately 120 calls/event, but when the number of users increases, $\lambda(x,t)$ abruptly stabilizes. In Fig. 5, the first ten events have been dropped out, and the evolution of the arrival rate $\lambda(x,t)$ is quite stable, which roughly justifies the formula (21).

Fig. 6 shows the revenue per unit event for the model obtained from the simulation (24) (dashed line), steady-state equation (24) (dotted line), and analytical model (38) (solid line).

In Fig. 6, the two curves drawn by the dotted and dashed lines are approximately equal, which is a consequence of the steady-state behavior of $\lambda(x,t)$. The curves in Fig. 6 match quite well showing that the analytical steady-state model (38)–(40) holds rather well.

In these curves, α (row axis) is varied from -10 to 10. It is seen that the optimal α that maximizes the revenue is $\alpha > 0$, and therefore, the optimal curve $f(x,\alpha)$, as represented in (4), is slightly concave. The optimal f function corresponds to $\alpha = 0.1$ is represented in Fig. 7, and it is seen that for $\xi = 2$, it is nearly a straight line. The interpretation is that when the price depending curve $h(x,\xi)$, shown in Fig. 4, is gentle, it is plausible to allocate nearly linearly the resources with respect to the price. As we will see below, when we change ξ to $\xi = 8$, the behavior of the price dependence clearly changes via the function $h(x,\xi)$, and as a consequence, the shape of the optimal utility curve $f(x,\alpha)$ also changes.

It is interesting to see from Fig. 6 that for small α, the revenue is very small. The justification is that in this case the offered data rate is relatively low at low prices, i.e., at those QoS classes where there are the largest number of potential customers. This can be seen from Fig. 8. The most convex (i.e., lowest) curve of

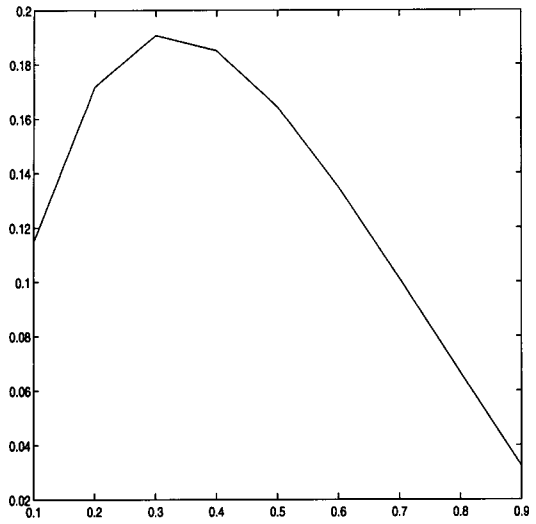


Fig. 9. Unscaled shape of the optimal arrival rate $\lambda(x) = f(x, 0.1)h(x, 2)$ obtained from the analytic model.

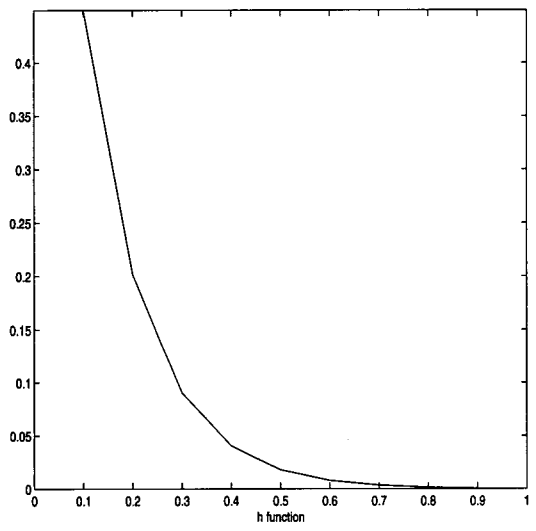


Fig. 10. The relationship of the price x and the price depending function $h(x, \xi)$ when $\xi = 8$. The function is sharply exponentially decreasing.

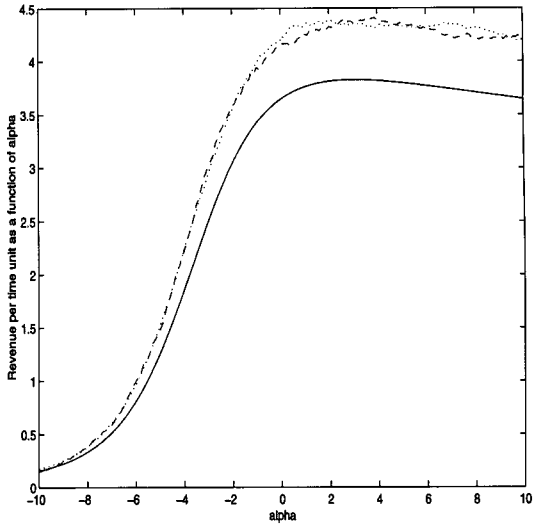


Fig. 11. Revenue per time unit as a function of α when $\xi = 8$. The optimal α is about $\alpha \approx 3.8$.

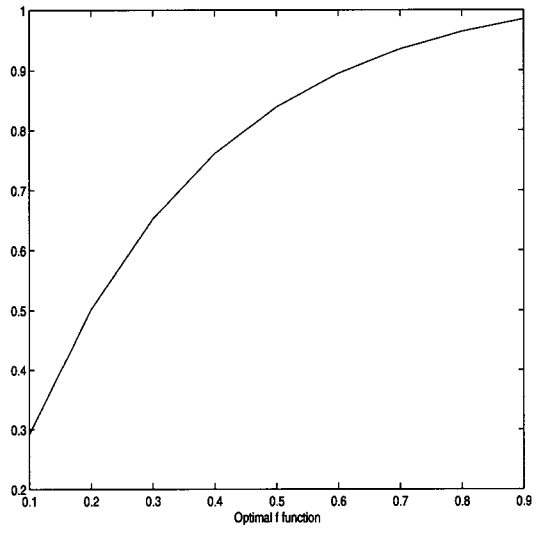


Fig. 12. Optimal f function based on the analytic model, when $\xi = 8$.

Fig. 8 represents the lowest value of α, i.e., $\alpha = -10$. Then the offered data rate for users paying the least amount of money is negligible, almost zero at the price values $x < 0.5$. That kind of capacity curve is not attractive. It is attractive only to those users who pay a lot of money, say more than $x = 0.5$ units of money, but the price depending function $h(x, 2)$ shown in the Fig. 4 is small for large x, and so there is only little number of potential customers who pay a lot of money, and thus the overall attractivity in the allocation scenario $\alpha = -10$ remains small.

On the other hand, when the most concave (i.e., uppermost) curve $f(x, \alpha)$, $\alpha = 10$, in the Fig. 8 is used for allocating the data rate $u(x, t)$, the differential benefit achieved by the best offered QoS classes is negligible. We see that for $\alpha = 10$, in practice the same data rate is allocated for QoS classes corresponding to the money units $x = 0.5$ and $x = 1$. Thus, that kind of curve does not attract the users that could pay a lot of money. From these observations, we can conclude that the optimal function $f(x, \alpha)$ is somewhere between these extremes.

Fig. 9 shows the optimal steady-state arrival rate $\lambda(x)$ as a function of the price x. The arrival rate depends on the product $f(x, \alpha)h(x, \xi)$, as seen in (12). Somewhat surprisingly, the maximal arrival rate is achieved by the QoS class corresponding to the price $x = 0.3$, not the QoS classes corresponding to the lower prices. The reason is that here the shape of the price depending curve, Fig. 4, was selected in so gentle a way. That is, it is not sharp or abrupt, as expected by considering the possible true or estimated distribution of the willingess to pay. The overall arrival rate depends on the product $f(x, \alpha)h(x, \xi)$, and when f is only a little concave, the benefit achieved by paying a small amount of money, say $x < 0.3$, is little.

Experiment 2. We made another simulation by using the value $\xi = 8$, which shows quite a sharp price depending func-

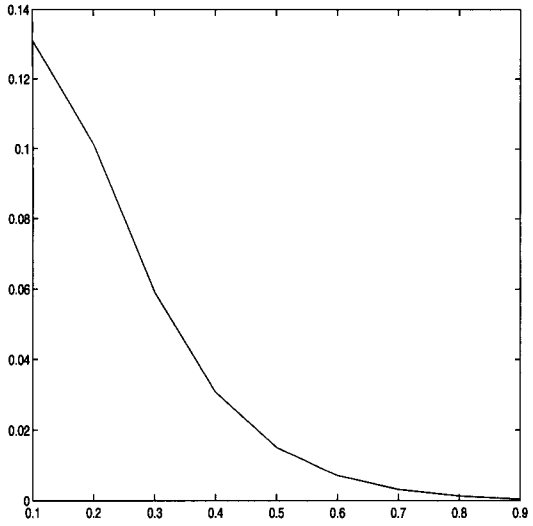


Fig. 13. Unscaled shape of the optimal arrival rate $\lambda(x) = f(x, 3.8)h(x, 8)$ obtained from the analytic model.

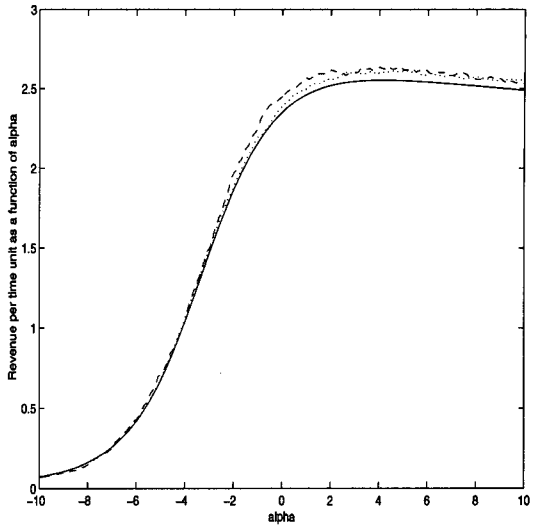


Fig. 15. Revenue per time unit as a function of α when $\xi = 8$. The optimal α is about $\alpha \approx 3.5$.

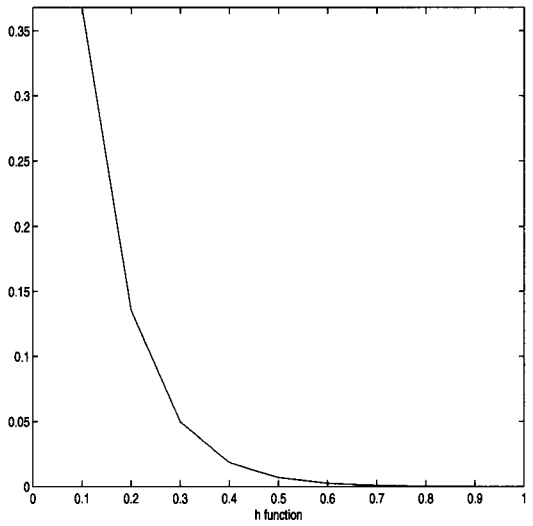


Fig. 14. The relationship of the price x and the price depending function $h(x, \xi)$ for $\xi = 10$. The curve is gently exponentially decreasing.

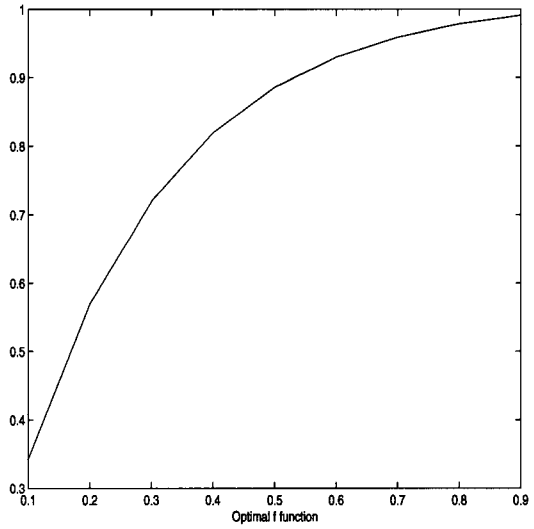


Fig. 16. Optimal f function based on the analytic model, when $\xi = 10$.

tion $h(x, \xi)$ in Fig. 10. Now there are lots of users who will pay very little money, compared to those users who pay a lot of money.

Fig. 11 shows the revenue curves for analytical as well as the simulated models, and again they match quite well. The optimal α is $\alpha = 3.8$, and the optimal utility function f is shown in Fig. 12. Now it is more concave as the curve obtained by using the value $\xi = 2$. The curve in Fig. 12 associates to the case of the airline service, where tourist, business, and first classes are available. The first class passengers will pay much more money than the tourist or business class passengers, while only receiving slightly more benefits. Thus the "capacity curve" correponding to the airline case is also concave, telling that the distribution of the willingness to pay is quite abrupt.

Fig. 13 shows the optimal $\lambda(x)$, and now the arrival rate monotonically decreases with respect to the price of the QoS class. Here the optimal f, and the offered capacity $\gamma(t)f(x, \alpha)$ is attractive enough for the cheapest classes.

From the simulations we can make the following conclusions:

- When the data arrival rate model (12) holds, the analytical solution (38–40) for calculating the revenue curve equals quite well with results given by the simulated model (24).
- In that kind of the model, the system stabilizes quite fast, which can be verified by seeing Figs. 5 and Fig. 11. Fig. 11 shows that the revenue functions (24) and (24) equal for a sufficiently large time T, and thus a steady-state behavior of the arrival rate is verified.
- Optimal α increases when ξ increases. That means: when the data arrival rate decreases steeply (e.g., exponentially) with respect to the price, the optimal capacity function becomes more concave. That is a very plausible behavior, because when the arrival rate $\lambda(x, t)$ decreases steeply

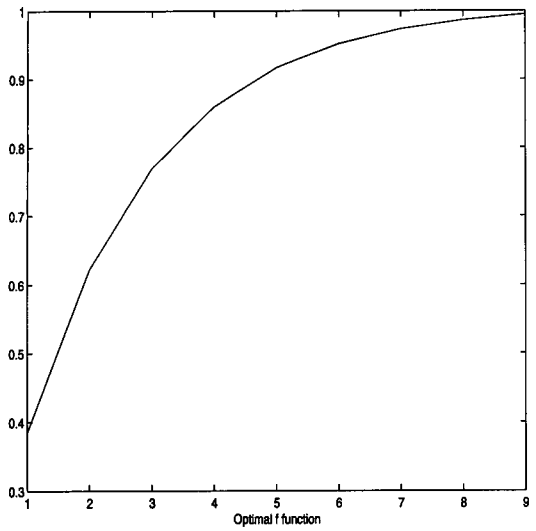


Fig. 17. Optimal f function based on the simulation model when $\xi = 10$

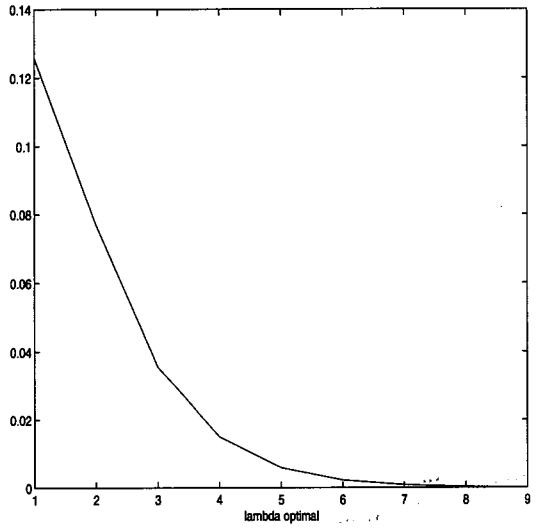


Fig. 18. Unscaled optimal λ as a function of QoS classes (analytic model).

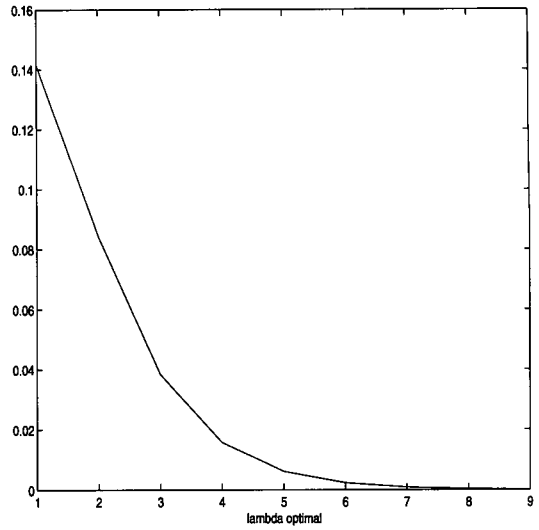


Fig. 19. Unscaled shape of the optimal arrival rate $\lambda(x) = f(x, 3.5)h(x, 10)$ obtained from the simulations.

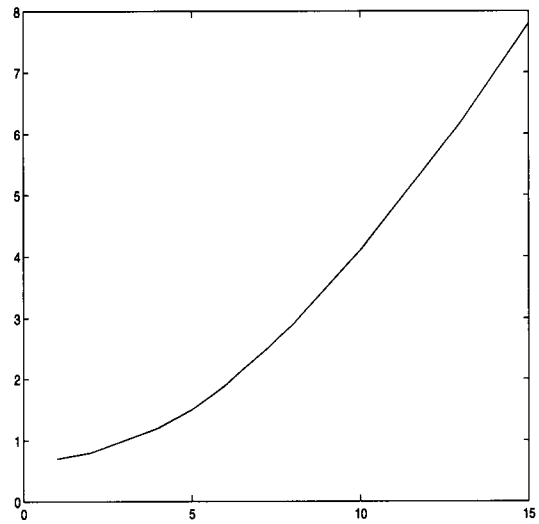


Fig. 20. The relationship of α and ξ for analytical solution when nine QoS classes exist.

with respect to the price x, the number of users who pay a lot of money is negligible, and therefore it is not sensible to allocate them much more data rate than to the users who pay less money. The clear concavity of f is a consequence of that fact.

- When the convex capacity function (i.e., $\alpha = -10$ or very small) is used, and when the model (12) holds, the revenue is relatively small, because that kind of capacity function does not match the distribution of richness of the users.

B. Pareto Distributed Duration

In the following simulations, we used the more realistic Pareto distribution, which allows the analysis of the case where very different durations in the network exist. The parameters in (17) were $E(connection) = 2$ and $p_b = 1.2$. In this simulation environment, the expected connection time $E(connection)$ was the same for each QoS class. The cumulative distribution function (18) was used to generate the Pareto distributed connection times.

The figures obtained from the simulations are the following:

- Fig. 14 represents the price depending function $h(x, \xi)$. Here $\xi = 10$, i.e., the function is sharply exponential.
- Fig. 15 shows the revenue as a function of α for analytic approximation and simulations, where $1/\mu$ has been replaced by the expected connection time $E(connection)$. It has a very similar form compared to that curve obtained by using an exponential duration model.
- Fig. 16 shows the optimal f function, which is obtained by using the value of α that maximizes the revenue in Fig. 15.
- Fig. 17 shows the optimal f based on the simulation model when $\xi = 10$.
- Fig. 18 shows the optimal arrival rate $\lambda(x)$.

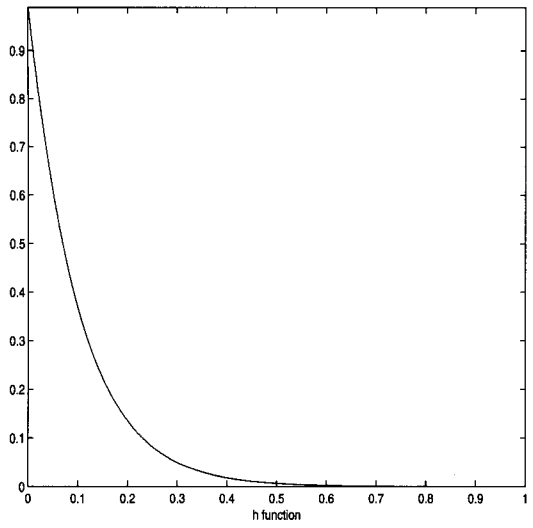


Fig. 21. The relationship of the price x and the price depending function $h(x, \xi)$ for 1000 QoS classes using the analytical solution. $\xi = 10$.

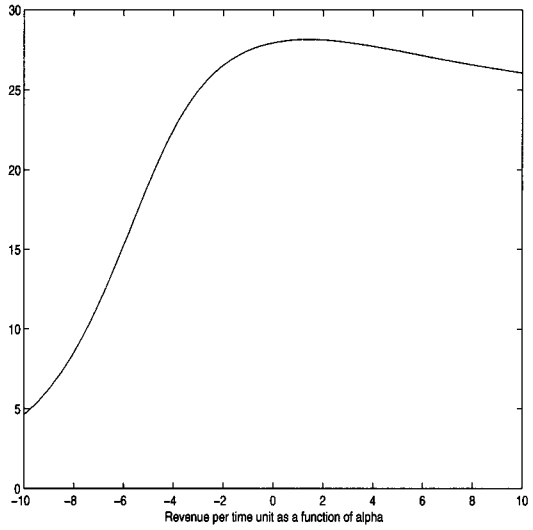


Fig. 22. Revenue per time unit as a function of α based on analytic solution for 1000 QoS classes when $\xi = 10$. The optimal α is about $\alpha \approx 2$.

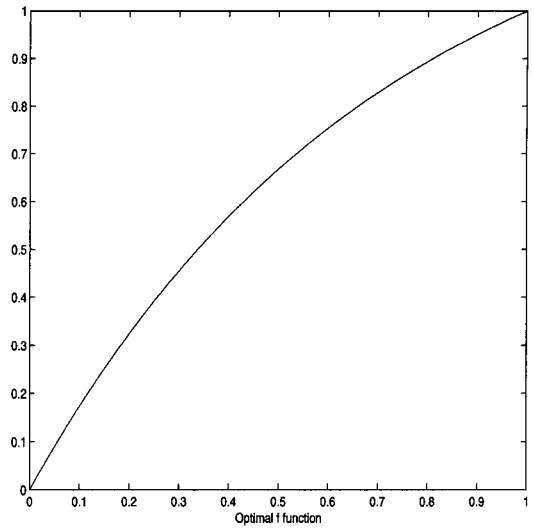


Fig. 23. Optimal f function based on the analytic model for 1000 QoS classes, when $\xi = 10$.

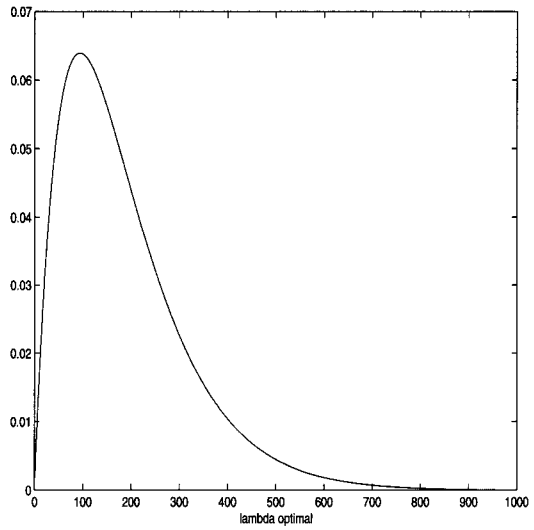


Fig. 24. Unscaled shape of the optimal arrival rate $\lambda(x) = f(x, 2)h(x, 10)$ obtained from the analytic model for 1000 QoS classes.

From these Pareto experiments we see that the behavior of the arrival rate and other quantities are quite similar to those obtained from the exponential duration model. Thus, the analytical model (38)–(40) can also be used in this study. Also the steady-state revenue solution (24) matches the simulated revenue (24) very well.

C. Experiments Using The Analytic Solution

We tested the relationship between the optimal α and ξ using the analytic model (38)–(40). The parameter ξ was varied from 1 to 15, i.e., the shape of $h(x, \xi)$ function varied from very gentle to very abrupt. Nine QoS classes existed. Fig. 20 shows that the parameter α increases with respect to ξ. Thus, the more abrupt the reaction of the users is to price x, the more concave the data rate function is.

In the following experiment, we made the test using the analytical model (38)–(40) using a very large number of QoS classes. We had 1000 classes. It can be thought for example that the cheapest class corresponds to the case where 1 cent is paid per time unit, say a minute. On the other hand, 1000 cents i.e., 10 dollars are paid per minute in the most expensive class.

Fig. 21 represents the $h(x, \xi)$ function, Fig. 22 represents the revenue function (38), Fig. 23 represents the optimal f function obtained by evaluating the revenue maximization procedure, and finally, Fig. 24 represents the optimal form of the arrival rate $\lambda(x)$. We see that for about 1 dollar/minute, we achieve the maximal arrival rate. The interpretation is that there is not very much interest to use very cheap QoS classes, say those where 1-20 cents per minute are paid, because these QoS classes offer too small data rates to be beneficial to customers. On the other

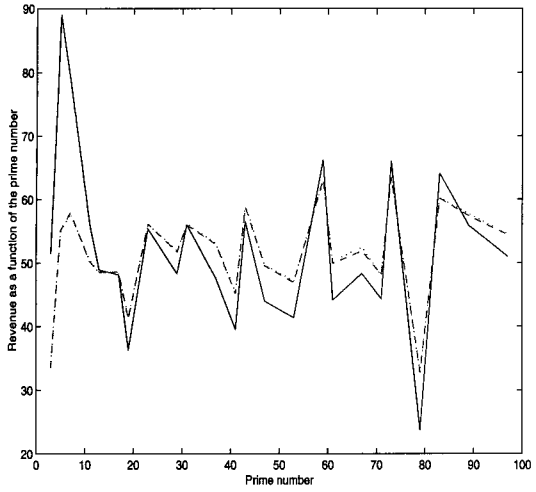


Fig. 25. Revenue per time unit as a function of random number (seed) when $\xi = 2$. The α was fixed to be $\alpha = 5$.

hand, in this simulation scenario, the users have enough money and interests to get better services from better QoS classes.

D. Experiment Using More General Model

In the last experiment, the goal is to try to prove that the stable state solution (41) and (42) holds for $p(x, \alpha_i)$ in its most general form. The arrival rate $\lambda(x,t)$ obeys (9). We generated $p(x,i)$, $i = 2, 3, 5, 7, 11, \cdots, 97$ using a random number generator, and i as a prime number being the seed. The number of classes was four. Our purpose is to simulate the very complicated reaction of the users to the different prices and QoS parameters. The function $p(x, i_0)$ for fixed i_0 can be thought of as a function of a sample taken randomly from the high-dimensional complicated QoS space, and for illustration purposes, one-dimensional figure is formed. Fig. 25 represents the revenue curves by using simulated as well as analytic approximation (41) and (42). From the figures, we see that they match quite well.

VI. REMARKS

Here remarks are given concerning the study:

- We have developed the new allocation scenario for QoS parameters and functions, when all requests are accepted. The allocation machine dynamically adjusts the QoS parameters. The detailed realization, simulation, and analysis of the technological realization is a future topic.
- The arrival rate model, which depends on the price (i.e., QoS class), as well as observed QoS parameters has been developed and analyzed using simulations and a stability analysis.
- The arrival rate λ is a time-varying function with feedback, and the simulations show that the arrival rate converges quite fast despite the stochastic behavior of the model. Small fluctuations of $\lambda(x,t)$ appear.
- Due to the stabilization of the arrival rate, the number of users is below some upper limit. This facilitates the development of the technical realization of the allocation machine.
- We have analyzed the stable state conditions of some special cases, and succeeded in reducing the arrival rate to be a function of several hidden and observed QoS functions and parameters. Still some open questions remain concerning the modeling and analyzing of a complicated relationship between λ and QoS parameters. For example, $h(x, \xi)$ is also hidden for an operator, and extensive experiments should be made for modeling that function. On the other hand, if the model of $\lambda(x,t)$ remains unknown, one can still use the allocation scenario by varying α and other hidden QoS parameters, and to try experimentally which value of α maximizes the revenue; the possibility that the use of a very fast analytic stable state approximation is lost, is not very critical.
- The benefit obtained by using the stable state model is that extensive simulations may partially be avoided.
- The use of the stable state model has been justified by showing that it matches well with the simulations. It is important to emphasize that in possible real world situations, the state $n(x,t)$, i.e., the number of users, may change much more slowly than in our simulations, depending on the applications. This means that the linearization approximation, which had a basic role in the derivation of the stable state solution, holds perhaps much better in a real world case.
- The stable state model is very flexible with respect to the choice of the function p.
- The simulation results were quite plausible, because the optimal utility functions were concave with respect to the price. Here we simply refer to the well known behavior of airplane passengers.
- In the experiments, the obtained result that α increases with respect to ξ, was also plausible.
- Simulations using one parameter α shows that the revenue curves consisted of only one local (i.e., global) maximum. That observation possibly gives us more sophisticated strategies to find the optimal parameters. However, that must be more carefully studied in the future.

VII. CONCLUSIONS

Economic models can provide new insights into resource sharing and QoS provisioning in future networks which will connect millions of users, and provide a large number of services. Pricing and competition can provide solutions to reduce the complexity of service provisioning and efficiently utilize the resources. Successful network revenue growth requires service providers to offer combinations of quality and price that match user needs, but to do this, providers must understand the structure of user demand. Here we have presented strategies for optimizing QoS parameters when the user profile density is available.

Novel contributions in this paper are as follows. The pricing mechanism is allowed to be continuous and the data arrival rate depends on the current link utilization and user's payment (selected QoS class). The mechanism is also very dynamic, because the system accepts every call. More precisely, there is

no absolute blocking, but when the network load is low (high), the data rate per QoS class is high (low). Both exponential and Pareto duration models were considered, and when the arrival rate increased with respect to the offered capacity, and decreased with respect to the price and the network load, the simulations matched with the results given by the derived analytical equations well. Thus, if one has *a priori* information about the behavior of the users within different QoS classes, one can obtain rough estimates for the optimal data rate allocation functions.

We believe that our approach can also be applied for optimizing some other QoS parameters than data rate. In the future, we will investigate market based mechanisms to admit and route sessions using resource price information in large networks, where each packet switch is the supplier. More experiments with different types of arrival rate models and the numbers of QoS classes will also be done to find out what kind of classification models are possible to implement into the switches and routers.

REFERENCES

[1] I. C. Paschalidis and J. N. Tsitsiklis, "Congestion-dependent pricing of network services," *IEEE/ACM Trans. Networking*, vol. 8, no. 2, pp. 171–183, Apr. 2000.

[2] F. Kelly , "Charging and rate control for elastic traffic," *European Trans. Telecommun.*, vol. 8, pp. 33–37, 1997.

[3] F. P. Kelly, A. M. Maulloo, and D. K. H. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *J. Operational Research Society* 49, 1998.

[4] F. P. Kelly, "Notes on effective bandwidths," in *Stochastic Networks*: Theory and Applications, S. Zachary, I. B. Ziedins, and E P. Kelly, Eds. London, U. K.: Oxford Univ. Press, vol. 9, pp. 141–168, 1996.

[5] R. Cocchi *et al.*, "Pricing in computer networks: Motivation, formulation and example," *IEEE/ACM Trans. Networking*, vol. 1, no. 6, pp. 614–627, Dec. 1993.

[6] R. J. Gibbens and E. P. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, vol. 35, no. 12, pp. 1969–1985, 1999.

[7] S. Dewan and H. Mendelson, "User delay costs and internal pricing for a service facility," *Management Science*, vol. 36, no. 12, pp.1502–1517, 1990.

[8] S. Whang and H. Mendelson, "Optimal incentive-compatible priority policy for The M/M/1 Queue," *Operations Research*, vol. 38, pp. 870–883, 1990.

[9] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, "On the large deviations behavior of acyclic networks of G/G/1 queues," *Ann. Appl. Prob.*, vol. 8, no. 4, pp. 1027–1069, 1998.

[10] J. K. Mackie-Mason and H. R. Varian, *Pricing the Internet Public Access to the Internet*, Prentice-Hall, 1995.

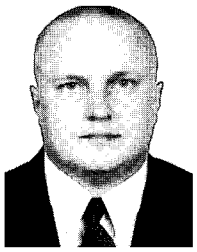
[11] A. Gubta, D. O. Stahl, and A. B. Whinston, "A stochastic equilibrium model of internet pricing," *J. Economics Dynamics Contr.*, vol. 21, no.4–5, pp. 697–722, 1997.

[12] A. M. Odlyzko, "Paris metro pricing for the internet," in *Proc of ACM Conf. Elec. Commerce*, 1999, pp.140–147.

[13] R. Gibbens, R. Mason, and R. Steinberg, "Internet service classes under competition," *J. Selected Areas Commun.*, vol. 18, no. 12, pp. 2490–2498, 2000.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley 1991.

[15] J. Banks, J. S. Carson II, and B. L. Nelson, *Discrete-Event System Simulation*, Prentice-Hall Inc. 1996.



Jyrki Joutsensalo was born in Kiukainen, Finland, in July 1966. He received diploma engineer, licentiate of technology, and doctor of technology degrees from Helsinki University of Technology, Espoo, Finland, in 1992, 1993, and 1994, respectively. Currently, he is Professor of Telecommunications at the University of Jyvaskyla. His research interests include signal processing for telecommunications, as well as data networks and pricing.

Timo Hämäläinen received the B.S in automation engineering from the Jyvaskyla Institute of technology, Finland in 1991 and the M.S and Fil.Lic degrees in telecommunication from the Tampere university of technology and University of Jyvaskyla, Finland in 1996 and 1999, respectively. His Ph.D. work studies QoS and pricing issues in broadband networks. His current research interests include intelligent wired and wireless networks (QoS and pricing).