Demand management for telecommunications services

Grigorios Zachariadis, Javier Barria *

Intelligent Systems and Networks Group, Dept. of Electrical & Electronic Eng., Imperial College London, London SW7 2AZ, UK

Abstract

In this paper we develop a novel method of controlling the demand in a multi-class, QoS-enabled network, using pricing and resource allocation for income maximisation. We first present a solution to the problem of calculating the optimal prices and QoS for a single link using a limiting regime approximation, which reduces the associated computational burden. A heuristic algorithm is then proposed that improves the limiting regime solution, achieving better results for links with small capacity. We further extend this approach to a multi-link network, where a distributed iterative algorithm is developed based on the solution of the single link model. Results from small and medium size networks show that, even when the assumptions we used do not hold, our approach yields results very close to the optimal ones (0.17% to 2.95% difference), which are computed by exhaustively searching in the decision space. Moreover, the calculation time using the proposed approach is approximately 1.5 minutes for problems which took more than 240 minutes to solve using exhaustive search.

Key words: Revenue management, pricing, utility, asymptotic approximation, resource allocation, quality of service

1 Introduction

In this paper we propose an approach to managing a telecommunication network by controlling the demand for services. In contrast to most existing approaches, which allow demand to depend either on prices or on quality, we

* Corresponding author
Email addresses: grizac@imperial.ac.uk (Grigorios Zachariadis,), j.barria@imperial.ac.uk (Javier Barria).

examine the case where demand for a particular service depends on both these characteristics. We use this approach to solve the problem of maximising the income of a multi-class network provider using pricing and resource allocation.

In the scenario here investigated the network provider is allowed to offer a limited selection of alternative classes of service to users. Each potential user, according to his/her price and quality requirements, will select the most appropriate and request a connection. Calls belonging to a specific class of service have identical characteristics. Our model is similar to the differentiated services model [31] in that: i) there is a limited number of classes, ii) packets or flows of packets of the same class are treated in the same way, and iii) state information is proportional to the number of classes rather than the number of flows. In diffserv the classes' quality of service (QoS) differs in terms of delay, jitter, reliability and other parameters. Using a value similar to effective bandwidth, our model may be a good approximation of the diffserv model. The exact mapping of other QoS requirements to bandwidth can be the subject of further research. We also note that, by making decisions for sets of calls (classes) rather than individual calls, the number of decisions to be made is greatly reduced, improving the scalability of our solutions. The tradeoff is that the decisions are not necessarily optimal for each individual user, since not all the users belonging to a class are characterised by identical requirements.

We assume that, for a given price, an increase in the resources allocated for each call of a specific class leads to an increase of the QoS for this class. If this is the case, the number of potential users willing to subscribe to this class increases, since more users will find the offered quality acceptable. However, the extra resources which will now be used by the users which were already satisfied with the lower quality could have been used for other potential users which, even after the increase of QoS, are not satisfied with it. The unsatisfied users do not request connection and hence are lost from the system. For this reason, a proportion of the resources can be considered not optimally used. The exact amount of resources which are not optimally used in this sense, is the difference between the minimum resources needed to satisfy the total active users, and the actual resources used by them. It could be argued that these resources improve the utility of the users, and therefore may be considered optimally used, but from an income maximisation point of view, these resources do not contribute directly to the income of the network provider.

In this paper we develop methods for minimising the amount of resources not used optimally, as defined earlier. These methods depend on pricing, since the willingness to pay is here considered as being related to the quality expected by a user. Setting different prices for each service is a method for ensuring that users do not request admission for a service which offers much better QoS than they need, assuming that users of low requirements are willing to pay a lower price, and in this way it partially alleviates the problem of misused resources

mentioned earlier. Therefore, the way price and QoS are determined for each class greatly affects the efficiency of a network.

We firstly explore the problem of finding optimal prices and qualities for the case of a single-link, with the aim of maximising the income. We present a solution which is based on limiting regime approximation and then propose methods to improve this solution. We then develop a distributed iterative solution for the problem of a multi-link network, a solution which is based on the single link solution.

1.1 Related work

Most published research uses pricing to control or regulate one of two things: either i) the incoming rate of new connection requests or ii) the data rate sent by existing connections or users. In this paper we use pricing to control the incoming rate of new connections.

A method considering pricing for multiservice networks is presented in [10]. The QoS of each service is given and guaranteed, and the problem of pricing is formulated as an expected income optimisation problem. In contrast, in this paper, we take into account QoS as a decision variable, which affects demand when determining the optimal prices.

Dynamic pricing policies for network services have also been widely investigated, and there is a great variety of propositions. Agent based [11], time-of-day [6], usage-based [30], threshold-based [21][7] and congestion dependent pricing [20][12] are some of them. The work of Paschalidis et al [17][18] on dynamic, congestion dependent pricing suggests that static (congestion-independent) pricing is asymptotically optimal, and therefore the benefits from dynamic pricing may not worth the extra complexity especially for large networks with many calls. A distributed mechanism for resource allocation, based on the pricing results of [17][18] is given in [19]. In contrast to the above approaches, our model allows the provider to make only static decisions.

Regarding static pricing, in the "Paris Metro Pricing" (PMP) scheme [15][22] two classes of services exist in the network, and each one has its own queue. Calls of the high class cost more than calls of the low class. In this way, fewer users request calls of the high class, and each call of the high class can thus use more resources, therefore receiving higher QoS. The PMP scheme exploits also the dependence of demand on prices. In this paper we extend the scheme proposed in PMP, as we also control and guarantee QoS.

Another study based on demand and pricing, is the work of McKie-Mason and Varian [13], which assumes that users pay a price depending on the congestion

of the network and the utilisation, which in turn depend on the demand curve of individual users and the price. In contrast, in the approach presented in this paper we focus on pricing and QoS for classes of service, rather than individual users, and in our model pricing is not congestion-dependent.

Resource allocation using auction based pricing is another popular subject of research [26][27]. Compared to such studies , our approach is different in that the network provider informs the users about available prices and qualities and users can decide whether they will request a connection or not.

Other researchers have addressed this problem resorting to a game-theoretic framework [14][1][29][28]. In such papers the users are considered players who compete against each other for resources, who adapt their demand (either by changing their rate or their priority/service class) to prices until an equilibrium is reached. This approach is not relevant in the scenarios studied in this paper were the provider sets and guarantees the quality and the price of each class, and each user just has to evaluate whether a class is satisfactory for him or not.

In the work presented in [2] and [3], the authors study the performance of the network under different bandwidth partitioning policies. It is observed via simulations that the prices and QoS offered have an effect to the income of a network provider. In [6] the combined problems of provisioning and pricing are studied. By provisioning the authors mean the procedure by which the provider purchases bandwidth, which then is divided between two classes. Time of day pricing is used, and making decisions in different timescales is evaluated. Provisioning is based on the actual cost of bandwidth which is assumed to be purchased by the provider. In our work we assume that the provider owns the resources or has purchased them in advance, and therefore provisioning is done under the constraint of limited bandwidth.

When addressing the problems of resource allocation and pricing for telecommunication networks, our approach brings revenue management ideas from operations research. Revenue management [4] aims at maximizing income through the application of tactics that predict customer behaviour and the optimisation of services availability and price. Revenue management has been successfully applied in industries like hotels or airlines, and it has been suggested that it can be applied in computer networks too [8]. We advocate that the service quality (QoS) is also a parameter which has to be considered in the case of telecommunication services.

The operation of a network provider is a complex affair where several targets have to be met simultaneously. In our work we concentrate on the target of income maximisation, which, due to the fact that the marginal costs of a network provider are minimal, can be translated to profit maximisation. Any telecom-

munications business will need to assess their income as part of their operations viability and hence we investigate this aspect here. Other targets, which have been suggested in the research community are: i) Welfare maximisation: $\sum u_i$ (The sum of all utilities of all users), and ii) Fairness: maxminu_i([25]). We note that the target of welfare maximisation can be achieved within the framework presented in this paper, where instead of price we consider the utility of the user.

The underlying aim of this paper is to develop a suitable framework to manage the offering of classes of service (CoS) assuming that the demand for a particular CoS depends on prices and QoS. We propose a framework for the calculation of the optimal static decisions regarding prices and QoS.

In Section 2 we describe how demand is modelled and its dependency to prices and quality. Then, the solution of the optimisation problem for a single link is presented in Section 3. Two solutions are developed; one based on a limiting regime approximation, and one based on a heuristic, that improves the limiting regime solution in the cases where the scale of the problem is small. Section 4 analyses the extension of the solution to a network of many links, developing a distributed solution. The paper ends with our conclusions and further research directions.

2 Demand model with respect to price and quality

Consider a provider offering services belonging to several CoS. Each class represents a different service category in resource allocation. Calls belonging to the same class are allocated the same amount of resources, and therefore experience the same QoS. We assume that each call of the i'th CoS incurs a one-off charge of p_i. The QoS of each class i is statically set and advertised. Let q_i be a fixed QoS index for the i'th CoS, indicating the QoS that a call of the i'th CoS receives.

Users interested in making a call arrive according to a Poisson process with mean rate λ_0. In order to model the satisfaction of a potential user from the price and the quality of a call, we have used utility functions. We choose a function which includes explicitly the price element, i.e. a highly priced service has a lower utility than a low priced service of the same quality. As in [24], a user's behaviour is modelled by the following utility function:

$$U_i\left(p, q, \gamma\right) = \left[\gamma(p - p_i) + (1 - \gamma)q_i\right] \theta(p - p_i)\theta(q_i - q) \tag{1}$$

where $U_i\left(p, q, \gamma\right)$ is the user utility with respect to the i'th CoS, and p_i and q_i

Fig. 1. How bandwidth affects quality for a user with minimum requirement q.

are the price and QoS for the i'th CoS, respectively. In this equation, p is the maximum price a user is willing to pay and q is the minimum QoS the user can tolerate, respectively. We assume that $q = \eta p$ (η constant) to reflect the fact that users who demand better quality are willing to pay more. We assume that $p \in [0, 1], q \in [0, \eta]$. Without loss of generality we can assume that $\eta = 1$ (scaling the maximum value of q to be 1). Then we can say about the price and the quality requirements of a specific user that $p = q$, taking into account that there is an implicit parameter $\eta = 1[bandwidth\_unit]/[price\_unit]$ which will be omitted in the rest of the paper. We can therefore use a single parameter r signifying both the price and the quality requirement. The step function $\theta(x)$ is 0 for $x < 0$ and 1 for $x \geqslant 0$. It is also assumed that γ is a weight that quantifies users' relative sensitivity to service price and service quality.

The utility function as presented in our paper is an approximation of other well known and used utility functions. Many QoS sensitive users have linear or sigmoid utility functions [32][33]. Our model captures a common characteristic of most utility functions, i.e. the fact that utility increases as bandwidth increases or as price decreases. Due to the assumption of a minimum accepted quality for each user, our actual utility function is zero up to that point, then suddenly increasing to a value (equal to q), and then linearly increasing, mimicking a sigmoid function. If the minimum requirement is relatively low, the function looks more like plain linearly increasing, and the sigmoid factor does not play an important part. In general therefore, two of the most important families of utility curves that are well known (sigmoid & linear) are captured with our model (see Fig. 1). Note that the sudden increase in utility which is characteristic of the step function is also a part of our utility function.

Current research efforts focus on the more accurate mapping of the amount of available network resources to the perceived utility, see e.g. [36]. As more

research becomes available on the issue more elaborate and accurate schemes can be employed. However, our utility function, since it captures the important characteristics of many utility functions, offers a good first approach.

The requirements of users p, q and the distribution of the users with respect to p, q play a very important role, since it is via this distribution that the variables of price and quality for a class affect the actual demand. In this paper, this distribution is assumed to be uniform, unless stated otherwise. The use of uniform distribution is essential for the discovery of tractable solutions. It is possible to find solutions for other cases but only numerically. In Section 3.4 we present simulation results which show that the uniform distribution is a good approximation to other distributions.

With respect to the call durations we assume that they are governed by an exponential distribution with mean service rate μ_i for the i'th CoS. The relaxation of this assumption, using for example a phase-type distribution, would generalise call duration, at the expense of increased complexity, and can be the subject of further work. In the remainder of the paper we use that $\mu_i = 1$ for all i. The underlying assumption is that the duration of a connection is independent of the QoS it will receive. Note that we can relax this assumption. In this case our algorithms need only minor modifications (see [35]).

A potential customer chooses to request connection for a specific CoS if his/her utility for this CoS is positive. In the event that a user's utility is greater than zero for more than one classes of service the user chooses the one with the highest utility.

Using the distribution of the potential customers with respect to their demands and maximum tolerated price we can calculate the mean arrival rate of potential customers for each class λ_k for each p_k and q_k. Denote by λ_{ab} the mean arrival rate of users with requirements r such that $a > r > b$. If λ_0 is the total mean arrival rate of potential users and $f_r(r)$ their probability density function with respect to their requirements r, then

$$\lambda_{ab} = \lambda_0 \int_b^a f_r(r) dr \tag{2}$$

The value of λ_{ab} is equal to the shaded area in Fig. 2. Note that λ_{ab} is also the mean arrival rate for a class with QoS a and price b, when users with requirements from b to a can only be satisfied by this class. Assume a user with requirement $r(= p = q)$. If a class k has a quality $q_k = a$ such that $a > r$ then the minimum quality requirement is met for this user. Similarly, if the price of a class is $p_k = b$ such that $r > b$ then the requirement regarding price is met. Therefore, a class with price and quality such that $a > r$ or $r > b$ will represent positive utility for the user with requirement r, and hence only users

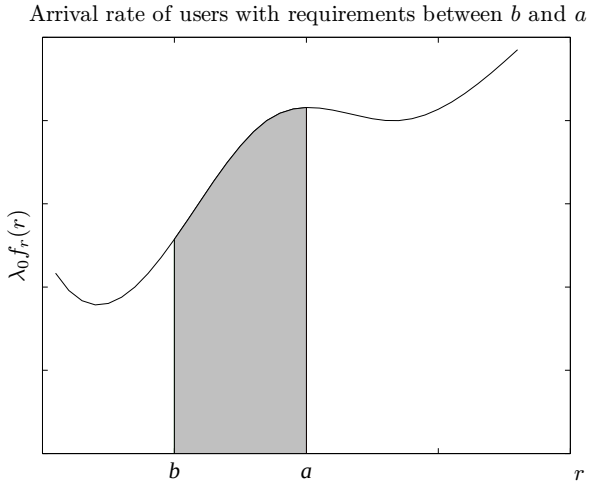

Fig. 2. Arrival rate of users with requirements between b and a.

with requirements $b < r < a$ will request admission for this class. For example, assume a user with requirement 0.8. This means that only classes for which the price is lower than 0.8 and the quality is larger than 0.8 will represent positive utility for the user. The values of p_k and q_k are the two main variables which determine the arrival rate for class k. If the prices and qualities of two classes are such that some users might be satisfied by both of them we will refer to these classes as "overlapping". In this case the arrival rates for each of the overlapping classes may also depend on the price and quality of the other class, since they are competing for the same users.

We implicitly assume the existence of another class of service (best effort), which caters for elastic traffic. This traffic exists with the lowest priority in our network, and we can consider only the QoS traffic in our analysis. A potential issue rising from this is best effort starvation, which is a fairness problem studied by [25]. A way of dealing with this issue is reserving some resources for best effort traffic.

The following parameters are considered known to the provider: λ_0 (total mean arrival rate), μ_i (mean call duration), the distribution of users with respect to their requirements and the capacity of the link/network. The change in those parameters (*not* the variance due to the stochastic nature of the problem) happens much more slowly than the algorithms presented can be executed. In addition to that, probably the most important factor of change is the time of day, or time of week, which can be studied for a period of time and thus providing with a good estimation of the expected demand on a specific time.

Finally we note that we analyse the user behaviour under the assumption that users do not have alternative choices to the ones offered by the provider. This can be the appropriate model in cases where i) there is a monopoly, ii) there is a subscription service and users are charged on top for usage,

iii) a provider targets a specific segment of the market and therefore faces virtually no competition, or iv) the providers act in cooperation. If multiple competing network providers exist, there are two options: the first option is to assume that the other providers do not react to the decisions of the provider whose income we are trying to maximise. In this case, given the complete set of services offered by other providers, and the characteristics of the total demand, the impact of the decisions of one particular provider to the demand for its services can be calculated. The second option, is that decisions made by one provider will trigger a reaction from other providers. In this case, a Pareto equilibrium will be reached, and we assume that the model we present implies that Pareto optimality has been previously reached between different providers for our solutions.

3 Optimal prices and qualities for a multi-class link

We first look at the scenario where users require service on a specific link, and the provider offers K different classes of service, for which the provider sets explicitly the price and the quality. The capacity of the link is considered given and a very simple admission control takes place: if there are enough resources to accommodate a request, the request is accepted, otherwise admission is denied. In this case we consider the quality q_i of class i, to be equal to the bandwidth r_i for each call of class i. Denote $\mathbf{q} = (q_1, q_2, \ldots, q_K)$ and $\mathbf{r} = (r_1, r_2, \ldots, r_K)$. The network provider can provide the specific level of resources to each call (or a good approximation) using a scheduling discipline like weighted fair queueing (WFQ) where the weights are set depending on the number of active calls of each class and the set quality for each class. Any requested connection is accepted with the condition that if it is accepted, the required resources for this connection do not exceed the capacity of the link.

We further assume in our model that users who request admission for a class and are rejected (blocked) do not ask admission for another class, hence the expected income per time unit is

$$I(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{K} \lambda_k(\mathbf{p}, \mathbf{q})(1 - B_k)p_k \tag{3}$$

In this equation K is the number of classes $B_k (= B_k(\mathbf{p}, \mathbf{q}))$ the request blocking probability for class k, $\lambda_k(\mathbf{p}, \mathbf{q})$ [1] the mean request arrival rate for class k, and p_k the price for class k. Note that there is no closed form solution for the

[1] Note that we will use the notations $\lambda_k(\mathbf{p}, \mathbf{q})$ and λ_k interchangeably, depending on whether the dependence of λ_k on the values of prices and qualities needs to be emphasized.

values of prices $\mathbf{p}$ and qualities $\mathbf{q}$ which maximise the above value due to the nature of B_k.

A call is admitted if the system remains in a state which belongs to the set of feasible states S. A state $\mathbf{n} = (n_1, ..., n_K)$, where n_i denotes the number of active calls of class i, belongs to S if $\sum_{k=1}^{K} n_k r_k \leq W$, where W is the total bandwidth of the link. The steady state distribution of the state of the network is given by a product-form solution [23]:

$$\pi(\mathbf{n}) = \frac{1}{G(S)} \prod_{i=1}^{k} \frac{\rho_i^{n_i}}{n_i!} \ , \ \rho_i = \frac{\lambda_i}{\mu_i} \tag{4}$$

where G(S) is defined as

$$G(S) = \sum_{\mathbf{n} \in S} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{n_i!} \tag{5}$$

If S_k is the set of states for which a call of class k is admitted, then the blocking probability for class k is $B_k = 1 - \frac{G(S_k)}{G(S)}$. In the book of Ross [23] a recursive algorithm for the calculation of blocking probabilities for the stochastic Knapsack is presented. We use this algorithm for the calculation of blocking probabilities where necessary.

3.1 Characteristics of the problem

In this section we study the effect of prices and qualities on income. We limit ourselves to the case of two classes for illustration purposes. We also assume that the users are uniformly distributed with respect to their requirements for minimum quality and maximum price. For these assumptions, according to (2) requests for calls of a class i with price p_i and quality q_i will arrive according to a Poisson process with mean rate λ_i

$$\lambda_i = \lambda_0(q_i - p_i) \tag{6}$$

That is to say that the demand for calls of a specific class of service is linearly dependent on the price and the quality of this service. This formula is valid only for the case when class i does not overlap with other classes. If this was not the case, then a part of the users who arrive with the above mean incoming rate would request services for the overlapping class. The exact proportion depends on the users' sensitivity γ in (1).

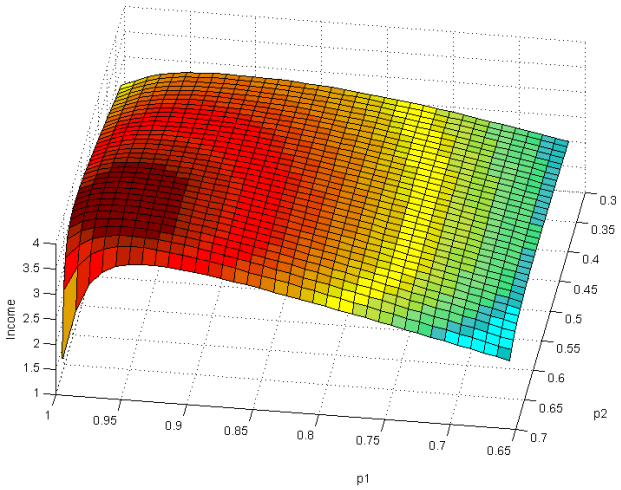

Fig. 3. Income with respect to prices

3.1.1 Behaviour with respect to prices for constant qualities

We now offer some perspective on the problem we are trying to solve and how prices and QoS of classes affect the providers income. As can be seen in Fig. 3, the income is concave with respect to the prices [2], and this observation will become apparent in the limiting regime case which is presented later in this paper. This concavity with respect to the prices could be exploited for a numerical calculation of optimal prices when the values of qualities are given.

3.1.2 Behaviour with respect to qualities for constant prices

The behaviour of $I(\mathbf{q})$ with respect to $\mathbf{q}$ is not smooth, as illustrated in Fig. 4, since there are many points where it is discontinuous. Any discontinuity of $I(\mathbf{q})$ is a result of a discontinuity in the functions B_k since λ_k is continuous.

We observe that q affects B_k in two ways: through λ and through S, the set of feasible states. As λ is continuous, any discontinuity is a result of a change in S. Note that changes in $\mathbf{q}$, and therefore, changes in $\mathbf{r}$ may change the set S, changing the blocking probabilities. The set S will change with $\mathbf{r}$ for the values of $\mathbf{r}$ for which there exist $\mathbf{n}$ such that $\sum_{k=1}^{K} n_k r_k = W$. In this case, with

[2] For some values of prices and qualities the income might not be concave, but this only happens in specific circumstances. In particular this may happen if the price of a lower class is very close to its quality, in which case the low demand for this class may cause a relatively high increase in the blocking rates of a higher class, thus leading to a convex increase in income as the price of this low class increases.

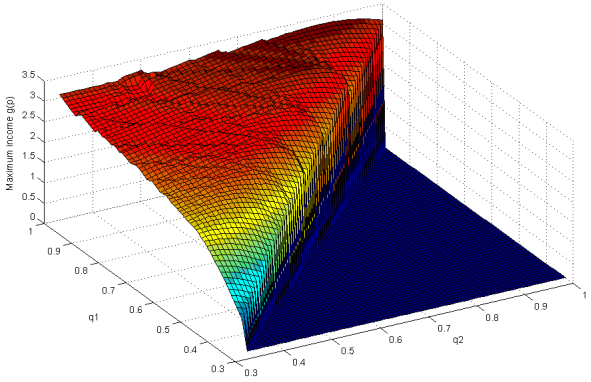

Fig. 4. Income with respect to qualities

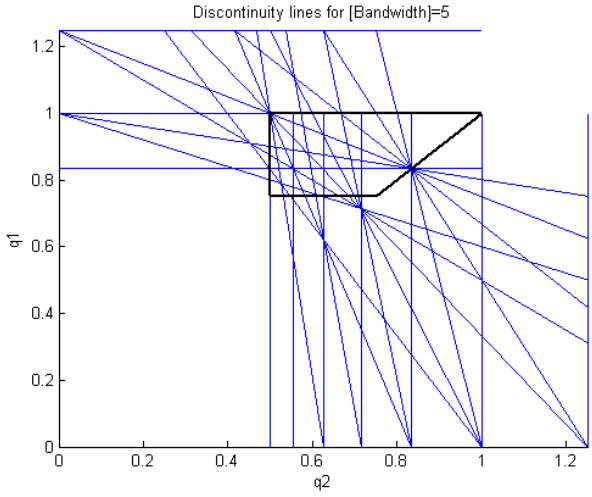

Fig. 5. Discontinuity lines. The thick lines define the surface which represents the set of allowable values for q_1, q_2.

any increase of r_k by δ for some k, for the same state $\mathbf{n}$, we have $\sum_{k=1}^{K} n_k r'_k > W$, where $r'_k = r_k$ or $r'_k = r_k + \delta$, and therefore $\mathbf{n} \notin S$. For example, in a single class link with $W = 4$ increasing the quality from 0.49 to 0.50 (for which no n exists such that $0.49n = 4$) will not alter the feasible states set, whereas an increase from 0.50 to 0.51 (for which $n = 8$ satisfies the equation $0.50n = 4$) will make the state with 8 active calls non-feasible. Therefore any points of discontinuity should be searched on the family of curves given by the equations:

$$\sum_{k=1}^{K} n_k r_k = W \tag{7}$$

for all the possible values of $\mathbf{n}$. The resulting graph of the lines on which g is discontinuous is shown on Fig. 5. The discontinuity lines separate the area of the possible q_1 and q_2 values into small sections where the function I is

continuous. Although it is possible based on the above discussion to devise a numerical method for obtaining the optimal values of prices and qualities, as the number of classes increases the method becomes intractable. To overcome this, we have developed an approximation, based on the limiting regime approximation [9].

3.2 Limiting regime (LR) approximation

A case of special interest is when the capacity of the link is very high, and therefore a large number of calls can be accommodated. In particular if we begin with a system with mean arrival rate λ_0 and capacity W and we scale the system by a parameter c, i.e. considering the system with mean arrival rate $c\lambda_0$ and capacity cW, then as $c \to \infty$ the system has certain characteristics, one of which is that if the offered load is not more than the capacity of the link the blocking probability is zero. Otherwise, the blocking probability is such that the served load is equal to the available capacity [9]:

$$B = 1 - W/\lambda_0$$

In such a system the optimal solution with respect to income is a system with blocking probability zero. If the blocking probability for a class is not zero, then by increasing its price we can decrease the offered load until it is equal to the capacity. In this way the carried load is exactly the same as in the previous case and the income is higher due to the increased prices.

Taking this into account, the problem of finding the values of $\mathbf{p}$ and $\mathbf{q}$ which maximise the income in the LR is:

$$\max_{\mathbf{p},\mathbf{q}} \sum_{k=1}^{K} \lambda_k(\mathbf{p},\mathbf{q}) p_k \tag{8}$$

subject to:

$$\sum_{k=1}^{K} \lambda_k(\mathbf{p},\mathbf{q}) q_k \leq W$$

where W is the total bandwidth and r_k the resources provided for a call of class k. For completeness, we summarize next some useful observations regarding the LR solution [35].

a) It is optimal not to have overlapping classes.

Suppose that we have two classes i, j with prices p_i, p_j and QoS q_i, q_j respectively. Suppose that i is the highest class of the two (meaning it has higher

quality and price). If $q_j > p_i$ (i.e. the quality of the lower class is higher than the price of the higher class) then the users with requirements r such that $q_j > r > p_i$ is satisfied by both classes. If the proportion of the users whose requirements can be covered by both classes and select class i is ζ, then users who are satisfied only by the class i arrive with mean rate $\lambda_{q_i q_j}$, users who are satisfied only by class j arrive with mean rate of $\lambda_{p_i p_j}$, users who are satisfied by both classes and choose i arrive with mean rate $\zeta\lambda_{q_j p_i}$ and users who are satisfied by both classes and choose j arrive with mean rate $(1-\zeta)\lambda_{q_j p_i}$. Therefore the income I and the occupied bandwidth W by those two classes are given by the following equations:

$$I = \lambda_{q_i q_j} p_i + \zeta\lambda_{q_j p_i} p_i + (1-\zeta)\lambda_{q_j p_i} p_j + \lambda_{p_i p_j} p_j$$
$$W = \lambda_{q_i q_j} q_i + \zeta\lambda_{q_j p_i} q_i + (1-\zeta)\lambda_{q_j p_i} q_j + \lambda_{p_i p_j} q_j$$

Let σ be the value for which $\lambda_{q_j \sigma} = \zeta\lambda_{q_j p_i}$. From the definition of σ, note that $q_j \geq \sigma \geq p_i$. It can be easily shown that if we set a new price for class i, $p_i' = \sigma$ and a new quality for class j, $q_j' = \sigma$ the new resulting income I' is higher than I and the occupied bandwidth W' is lower than W, therefore the new solution is better than the previous one. Hence, setting the prices and qualities in a way that more than one classes may be satisfying for some users is not optimal.

b) If $r^2 f_r(r)$ is strictly increasing, the quality of each class should the highest possible, i.e. equal to the price of the immediately higher class.

Consider the problem of finding what the optimal price and quality is for a given amount of bandwidth for a class. If the quality for the class is q_0 and the price p_0 and assuming that every user can only be satisfied by one class, the income I is:

$$I = p_0 \lambda_0 \int_{p_0}^{q_0} f_r(r) dr \tag{9}$$

Suppose that q_m is the maximum quality that the class in question can get (either because this q_m is the highest quality we can offer in our network, or because setting the quality higher than q_m for this class creates overlapping classes). Then, using q_m we can find a solution where the income obtained is higher, without using more bandwidth [35]. Let p_m be the price for which the following condition is achieved:

$$q_m \lambda_0 \int_{p_m}^{q_m} f_r(r) dr = q_0 \lambda_0 \int_{p_0}^{q_0} f_r(r) dr (= W) \tag{10}$$

The income in this case is:

$$I' = p_m \lambda_0 \int_{p_m}^{q_m} f_r(r) dr \tag{11}$$

For $r^2 f_r(r)$ strictly increasing, it can be easily proven that $I' > I$, provided that $f_r(r)$ is continuous, positive and bounded [35]. Note that the condition that $r^2 f_r(r)$ is strictly increasing includes the case of uniform distribution, for which $f_r(r)$ =constant.

c) Optimal prices and qualities

For a uniform distribution of users with respect to their requirements ($f_r(r) = 1$, $r \in [0, 1]$), the difference between the optimal quality of a class and the optimal price of the same class should be the same for all classes (for a proof see [35]), i.e.

$$q_k - p_k = \alpha, \forall k \in K \tag{12}$$

where $\alpha \in \Re$ is constant.

3.3 Limiting regime solution

The optimal quality and price for the k'th class is [35]:

$$q_k = 1 - (k-1)\alpha \tag{13}$$
$$p_k = 1 - k\alpha \tag{14}$$

There are two cases for α, depending on whether the capacity constraint is violated by the solution for infinite resources or not.

Case a) The capacity constraint is not violated. Then:

$$\alpha = \frac{1}{K+1} \tag{15}$$

Case b) The capacity constraint is violated. Then:

$$\alpha = \begin{cases} \frac{W}{\lambda_0} & \text{, if } K = 1 \\ \frac{\lambda_0 K - \sqrt{\lambda_0^2 K^2 - 2\lambda_0 K(K-1)W}}{\lambda_0 K(K-1)} & \text{, if } K \geq 2 \end{cases} \tag{16}$$

The capacity constraint is violated if:

$$\lambda_0 \frac{(K^2+3K)}{2(K+1)^2} < W$$

To get an indication of how the number of classes affects the income we observe that for the unconstrained case (low load) the income is $\frac{\lambda_0 K}{2(K+1)}$, therefore for a larger number of classes the income is improved, asymptotically reaching $\lambda_0/2$.

3.4 Improved Limiting Regime approximation

We note that the LR presented in Section 3.2 always underestimates the blocking probability, hence it overestimates the carried traffic for a given level of prices and qualities. In addition to this, according to the LR approach, increasing the offered traffic above the level which can be carried by the link does not result in an increase of the carried traffic. The carried traffic is always such that the utilisation is 100% and the exceeding demand is blocked. A more realistic assumption is, however, that the carried traffic never reaches 100% of the capacity (only asymptotically) and an increase in the offered traffic leads to an increase in the carried traffic. Hence, in this section we propose an improved solution to the problem of setting price and quality optimally.

The above observations suggest that the LR approach yields solutions where the offered traffic is less than the optimal. Therefore, in order to increase the income in this section we modify the LR solution in a way that increases the overall offered traffic. This can be done either by increasing the quality level of one or more classes and/or by lowering their price. In the LR solution, however, the quality of each class is equal with the price of its immediately higher class (due to observation b in Section 3.2). Therefore an increase in its quality can only transfer some of the demand from one class to another. For the LR we have shown that such a transfer is not beneficial (due to observation a in Section 3.2). Similarly, decreasing the price of any class, except of the lowest, does not increase the total number of users who request admission, because the extra users which now have positive utility for this class already had positive utility for the immediately lower class and would request admission to the network even without the change. Therefore the only way to increase the users requesting admission by changing a parameter (quality or price for a class), is by lowering the price of the lowest class. This is, therefore, the first part of the modified algorithm, after the calculation of the optimal LR solution. We calculate the blocking probabilities for each class and the partial derivatives of the blocking probabilities with respect to the price of the lowest class (for the prices and qualities that resulted from the LR solution) and use a linear approximation model for the blocking probabilities based upon which

we conduct an iterative optimisation procedure. The problem to be solved is

$$\max_{p_K} I(\mathbf{p}, \mathbf{q}) = \max_{p_K} \sum_{k=1}^{K} \lambda_k(\mathbf{p}, \mathbf{q})(1 - B_k)p_k \tag{17}$$

where the values of the blocking probabilities for each class i are linearly approximated by

$$B_i = B_i^* + \frac{\partial B_i}{\partial p_K}(p_K - p_K^*) \tag{18}$$

In the above equation p_K^* is the solution using the LR approximation (see Sec. 3.2) and B_i^* and $\frac{\partial B_i}{\partial p_K}$ are calculated using the values for prices and QoS that result from the LR approximation. B_i^* is calculated using the product-form formula or the recursive algorithm of Ross [23]. $\frac{\partial B_i}{\partial p_K}$ is calculated numerically, using the definition of derivatives. The index K refers to the lowest class. The function to be maximised is a 3rd degree polynomial (irrespective of the total number of classes) and can easily be maximised by setting the first derivative of I with respect to q_K equal to zero. Once an iteration is over we substitute p_K^* with the value for p_K which results from the above optimisation, calculate B_i^* and $\frac{\partial B_i}{\partial p_K}$ for the new value and perform the maximisation of the new resulting function. In this way we find the price of the lowest class which maximises the income.

Taking into account the characteristics of the optimal solution of the limiting regime approach we can make further adjustments to all prices and qualities (except the quality of the first class which remains at the maximum and the price of the lowest class which remains at the value calculated in the first step of this algorithm) so that the quality of each class is equal to the price of its immediately higher class, taking into account observation c in Section 3.2) from the LR solution. In this way a further improvement of the results is achieved. The above sequence of actions can be summarisedas follows:

Algorithm 1

(1) Calculate the LR solution (15) or (16).
(2) Solve (17), assuming that B_i is a linear function of p_K (18). Repeat this step until $|p_K^n - p_K^{n-1}| < \epsilon$, where ϵ a small positive real and p_K^n the resulting p_K of iteration n.
(3) For the resulting value of p_K, correct prices and quality of all classes according to observation c) of Section 3.2, i.e. for $k = 1, ..., K$ set $q_k = 1 - (k-1)\alpha'$ and $p_k = 1 - k\alpha'$, where $\alpha' = (q_1 - p_K)/(K-1)$.
(4) End.

The above algorithm achieves results within 1-2% of the optimum (Fig. 6) for all values tested, for a fraction of the computational time compared to an exhaustive search algorithm. The method we use for the calculation of blocking probabilities has a complexity of $O(CK)$, where C is the size of the link and K the number of classes. The complexity of Algorithm 1 is, therefore, $O(CK)$, since the complexity of the LR solution is $O(1)$. The complexity of an exhaustive search algorithm is $O(CK\Xi^{2K})$, where Ξ is the number of different values tested for the price and quality of each class, and it depends on quantisation. It is worth noting that the heuristic algorithm for a link with bandwidth 10 took 0.01 seconds to run [3], whereas the exhaustive search method for two classes took 5 minutes, (with a quantisation step of the prices and quality of 0.001), using a fast iterative method for the calculation of the blocking probabilities[23]. The difference in computational time increases exponentially as the number of classes increases. Our algorithm has converged for all tested simulation settings. The formal proof of convergence is the subject of further research.

In Step (2) of Algorithm 1, an alternative way of calculating the optimal reduction of p_K is using a brute force exploration (exhaustive search) over p_K only. The complexity of this is $O(CK\Xi)$. The results for this case are marginally ($< 1\%$) improved compared to Algorithm 1 and the calculation time is 0.25 seconds for a link of bandwidth 10.

We observe that as the mean arrival rate of potential customers increases the improvement of the income is higher. In the case where the offered traffic is very small (and therefore the blocking probability is low) the modified solution does not present significant improvement over the original LR solution. In Fig. 7, for each value of c we assume that the mean incoming rate of potential users is $c\lambda_0$ and the capacity cC_0. We observe that as the scale of the network increases, the improvement of the heuristic solution over the LR solution decreases. This is expected because the LR is a more accurate approximation for larger-scale networks. Figure 8 shows the income improvement of the LR and the heuristic solution compared to the income using the values from the solution for unlimited resources. Notice that for each case the income reaches a maximum, which for the case of LR and improved LR solution is equal to the income when all resources are sold at the maximum price per resource unit, i.e. the price for each class is very close to its quality. This can happen only when demand is so high that the prices can be set very high, which is usually not the case. A repeated occurrence of such a situation may be an indication that the network infrastructure is insufficient compared to the demand and it is probably beneficial to expand it, depending of course on the cost of such an

[3] All simulations in this paper were conducted using MATLAB (TM) v. 7.0.1.24704 (R14) Service Pack 1 on a Dell (TM) PC with Intel Pentium (TM) 2500MHz processor and 512MB RAM, under Windows XP Professional (TM) operating system

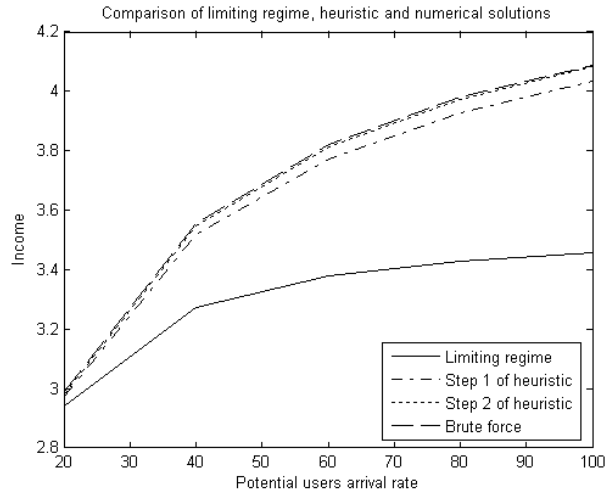

Fig. 6. Income using different methods for calculating an approximation of the optimal values of prices and qualities.

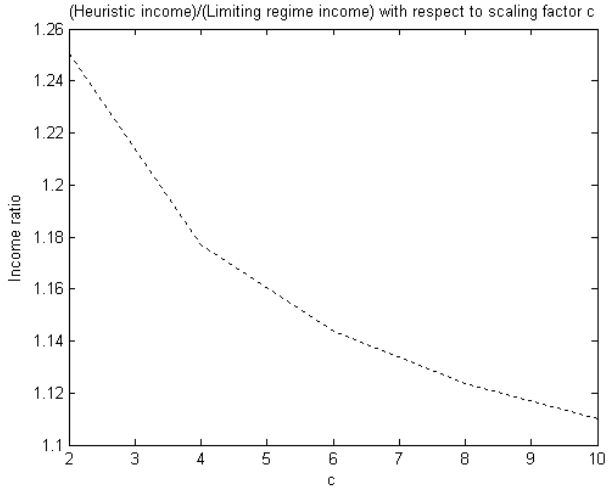

Fig. 7. Improvement of heuristic over LR as the scale of the network increases.

expansion.

We conducted further tests for non-uniform distribution of the users with respect to their price and quality requirements. The results are encouraging with an achieved income of no more than 3% lower than the optimal income for all cases tested. The distribution functions which were tested are the normal distribution (Fig. 9) and the exponential distribution (Fig 10). The optimal solutions on the graphs were obtained using an exhaustive search algorithm, whereas the heuristic ones have been calculated using Algorithm 1.

The relaxation of the linear relationship of p and q has similar characteristics to the relaxation of the uniform distribution assumption, since it mainly affects the linear relationship of the arrival rate for a class with its price and quality

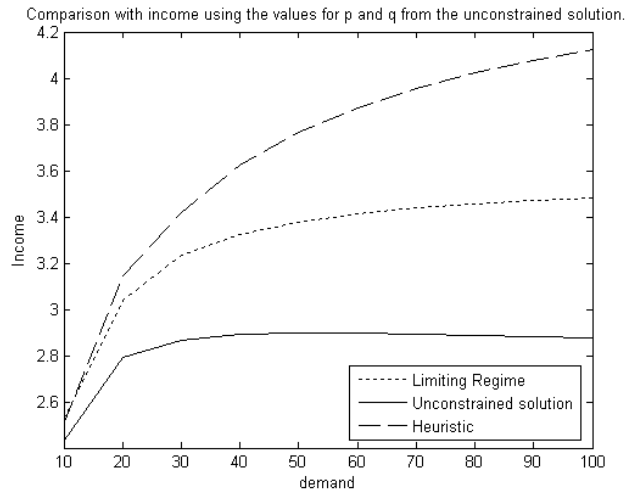

Fig. 8. Comparison of income using LR and heuristic solution with income using solution for unlimited bandwidth.

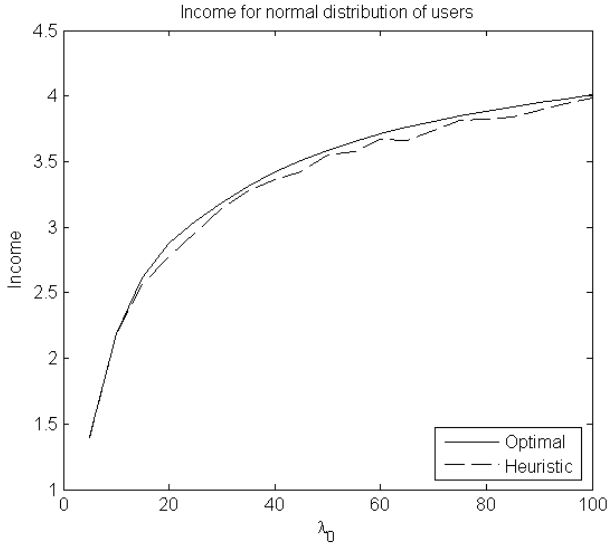

Fig. 9. Comparison of results between heuristic and optimal solutions for normal distribution of users with respect to their requirements.

($\lambda_k = \lambda_0(q_k - p_k)$). To the best of our knowledge this relaxation can only be solved numerically.

3.5 Remarks on maximisation using LR approximation

As the problem (8) is a case of constrained optimisation, we can approach the problem using Lagrange multipliers, which can offer us some further insights in the nature of the problem studied in this paper.

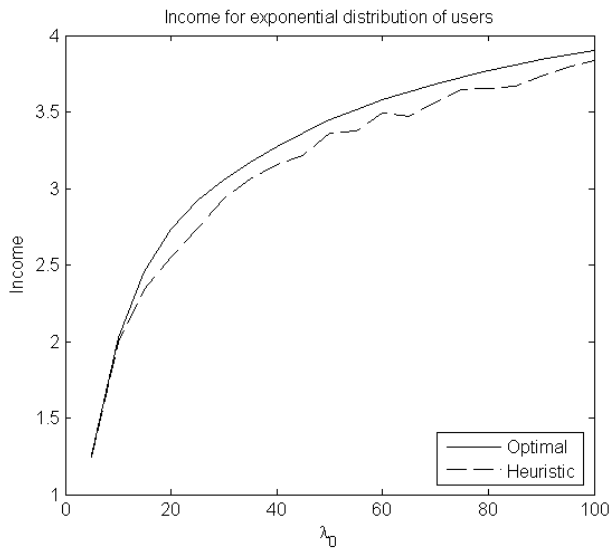

Fig. 10. Comparison of results between heuristic and optimal solutions for exponential distribution of users with respect to their requirements.

The function to be maximised is $\sum_i \lambda_i p_i$ and the constraints are:

- The total capacity constraint: $\sum_i \lambda_i q_i \leq W$
- The constraint that the price of a call is lower than its quality: $p_i \leq q_i$
- The constraints from observation a) in Section 3.2 that the quality of a class should not be higher than the price of the immediately better class: $q_i \leq p_{i-1}$
- The maximum allowable value of the quality of the first class should be 1 and the lowest value of the price of the last class should be greater or equal to zero: $q_1 \leq 1$, $p_K \geq 0$, thus ensuring that all prices and qualities $\in [0, 1]$

Then the Lagrangian function to be maximised is

$$\mathcal{L} = \sum_i \lambda_i(\mathbf{p}, \mathbf{q}) p_i - \Lambda[\sum_i \lambda_i(\mathbf{p}, \mathbf{q}) q_i - W] - \sum_i \Phi_i(p_i - q_i) - \sum_i \Psi_i(q_i - p_{i-1}) \quad (19)$$

The Lagrangian multipliers Λ, Φ_i, Ψ_i are either equal to zero, for the case when the corresponding inequality is not active for the optimal solution, or equal to the marginal value of relaxing the corresponding inequality, if the inequalities apply as equalities for the solution.

It can be shown that for any distribution of the users, it is not optimal to have the price and the quality of a class equal. This can be proven by using a similar method as we used for observation a) in Section 3.2. (Recall that when the price and the quality of a class are equal there is no demand for this class.) Therefore, the Lagrange multipliers regarding this subset of constraints (Φ_i) are equal to zero. Therefore, the Lagrangian is:

$$\mathcal{L} = \sum_i \lambda_i p_i - \Lambda \sum_i (\lambda_i q_i - W) - \sum_i \Psi_i (q_i - p_{i-1}) \tag{20}$$

where Λ represents the scarcity (or marginal value) of bandwidth, as classes compete for it. The values Ψ_i represent the scarcity of the demand. Note that from observation a) in Section 3.2 the prices and qualities of the classes should be such that each potential customer may have at most one class satisfying him/her. Hence, demand becomes also a scarcity, and classes compete with each other for it. Each of the above Lagrangian multipliers Ψ_i represents the marginal value of demand for each pair of neighbouring classes.

Taking the first derivatives of (20) with respect to p_i and q_i we obtain:

$$\frac{\partial \mathcal{L}}{\partial p_i} = \frac{\partial I}{\partial p_i} - \Lambda \frac{\partial W_0}{\partial p_i} + \Psi_{i+1} \tag{21}$$

$$\frac{\partial \mathcal{L}}{\partial q_i} = \frac{\partial I}{\partial q_i} - \Lambda \frac{\partial W_0}{\partial q_i} - \Psi_i \tag{22}$$

Here $W_0 = \sum_i \lambda_i q_i$ is the expected total occupied bandwidth, and $I = \sum_i \lambda_i p_i$ is the expected income per time unit. In order to find the optimal solution to the problem, we need the values in (21),(22) to be equal to zero. We now focus on what would these equations mean. For the case of (21), it would mean that the increase in income ($\frac{\partial I}{\partial p_i}$), plus the decrease in utilised bandwidth multiplied by the marginal value of bandwidth, plus the marginal value of demand (referring to users with requirements around the price p_i) is equal to zero. The second and the third terms of this equation are positive, therefore the first term will be negative. The physical interpretation of this is the following: the optimal price p_i is such that the decrease in income which would result by further increasing the price would be equal to the value of resources that would be freed plus the value of demand which would be available for the immediately lower class.

Similarly, from (22), we can state that the solution regarding q_i is such that the increase in income which would result by increasing the q_i would be equal to the value (or cost) of extra resources that would be now engaged plus the value of demand which would no longer be available to the immediately higher class, since increasing the q_i would mean increasing p_{i-1} according to observation a) in Section 3.2.

The above observation provides further insight into what our approach accomplishes using prices and QoS. Apart from the usual aim of allocating resources to classes, we also allocate demand to each class at the same time.

4 Network-wide models

Consider a network with a set L of links and a set T of possible call routes. Potential calls for route $R \in T$ arrive according to a Poisson process with mean rate $\lambda_{0,R}$. Each potential user has specific quality and price requirements. If a specific class of service on route R is satisfying the potential user, the user will request admission. If enough resources are available on all links, the request will be accepted, otherwise it will be blocked. Each call belongs to a specific class, which is the same for all the links it crosses. This means that a call cannot belong to class i of a link and class j, $j \neq i$ of another link.

In designing a pricing structure for our network, we assume that for each class k and each link l there is a one-off charge of p_{lk} which has to be paid by all calls of class k which pass from link l [37]. Therefore, the price of a call of class k on route R, p_{Rk} is the sum of the prices paid on each link of route R:

$$p_{Rk} = \sum_{l \in R} p_{lk} \tag{23}$$

With this approach, a call is charged for all the resources it is using. We also consider that the quality index of a call is equal to the data rate. Therefore, the quality of a call for class k on route R, q_{Rk}, is equal to the minimum quality of the links traversed:

$$q_{Rk} = \min_{l \in R} q_{lk} \tag{24}$$

where q_{lk} is the quality of calls of class k on link l. We assume uniform distribution of users with respect to their price and quality requirements, and that users are prepared to pay prices proportional to the number of links of their route, otherwise the provider will have a strong preference for users whose routes traverse less links, since it will probably be very profitable to effectively shut out users who require longer routes. Then the demand (mean arrival rate of users requesting admission) for a class of service on a specific route will be:

$$\lambda_{Rk} = \lambda_{0,R}(q_{Rk} - p_{Rk}/|R|) \tag{25}$$

where $|R|$ is he number of links on route R. The overall problem to be solved is the maximisation of the total income on all routes:

$$I = \sum_{R \in T} \sum_{k=1}^{K} \lambda_{Rk}(1 - B_{Rk}) p_{Rk} \tag{26}$$

In order to solve the stated network problem, we resort to link independence approximation [9] which assists in the decomposition of the problem to several local subproblems, one for each link of the network. This approximation is more accurate if a link is carrying many calls of low bandwidth. In order to find the characteristics of each single link problem to be solved, we use the low blocking approximation [9], where thinning of demand on a link due to blocking on other links of a route is not taken into account. Therefore, the blocking probability for class k on route R, B_{Rk}, is approximated by the blocking probability for class k on link l, B_{lk}, when we solve the local subproblem of link l. Under the low blocking approximation, potential users for a link arrive according to a Poisson process, with mean rate $\lambda_{0,l}$ equal to the sum of the mean rates for each route that passes through that link [5].

$$\lambda_{0,l} = \sum_{R: l \in R} \lambda_{0,R} \tag{27}$$

When solving the problem for each link, the approximation is made that the proportion of potential users who request admission for routes passing from this link, depends only on the price and the quality of this link. Hence, we approximate q_{Rk} with q_{lk} and p_{Rk} with p_{lk} for each route R when doing the calculations on each link l. The difference between the locally (on link $l \in R$) estimated mean arrival rate of users for a specific class k of a specific route R and the real mean arrival rate of users (assuming low blocking and link independence) is then:

$$\lambda_{0,R}(q_{lk} - \min_{m \in R} q_{mk} - p_{lk} + \sum_{m \in R} p_{mk}/|R|) \tag{28}$$

This difference is zero when the quality and the price for class k are the same on all links. Using the LR approximation analysis of Section 3.2, we observe that this will be true when the value of α in (16) is the same for all links. It is easily shown that α will depend only on the fraction λ_{0l}/W_l (potential call arrival rate/capacity), which we will call relative load of the link. This means that the decentralised solution will be optimal (under the low blocking and link independence assumptions) if the relative load on each link on the network is the same. The more diverge the relative load on the links, the more the decentralised solution differs from the optimal.

The demand for calls of class k on route R, when calculated for the solution of the income maximisation problem on link l is approximated by:

$$\lambda_{Rk} = \lambda_{0,R}(q_{lk} - p_{lk}) \tag{29}$$

Therefore, using (27), the total demand for calls of class k on link l is:

$$\lambda_{lk} = \sum_{R:l\in R} \lambda_{0,R}(q_{lk} - p_{lk}) = \lambda_{0l}(q_{lk} - p_{lk}) \tag{30}$$

which is identical to (6), the single link case. Hence, the problem that we solve on each link l is the maximisation of the income from link l, I_l:

$$I_l = \sum_{k=1}^{K} \lambda_{lk}(1 - B_{lk})p_{lk} \tag{31}$$

For the solution of each single link problem we employ the methods developed in Section 3. If we substitute p_{Rk} in (26) according to (23), and taking into account the low blocking approximation [5] ($B_{Rk} \approx B_{lk}$) when considering a link l, after rearranging and substituting we get $I \approx \sum_{l\in L} I_l$, and due to the low blocking and link independence approximations we assume that by solving each local maximisation problem, we achieve a good approximation of the global maximisation, since the local decisions on each link do not greatly affect the income from other links.

$$\max I \approx \sum_{l\in L} \max I_l \tag{32}$$

The accuracy of this approximation depends on the blocking probabilities on the network. The lower the blocking probabilities (for example, in a non-overloaded system in the LR of many small calls), the more accurate the approximations (29) and (32) and therefore the more accurate the final result.

The decentralised solution of (31) is an approximation of the optimal solution (26). The three basic parameters which affect the sub-optimality of the decentralised solution are the blocking probabilities, the difference in relative load on the links and the relative size of the calls to the size of the links. Calculating the solutions independently for each link is scalable, as opposed to a centralised optimal solution which would require to take into account the prices and quality of all links and routes when estimating the demand.

We further note that, although in cases of low traffic the low-blocking approximation does not cause significant difference from the optimal solution, when the offered load is very high, a decentralised solution based on the low blocking approximation will differ greatly from the optimal. A well known method

for dealing with the problem of calculating the blocking probabilities on a network without solving the corresponding system of equations is the Erlang fixed point iterations [9][23]. Therefore, in order to enhance our algorithm, at each iteration we assume that the offered load on each link is thinned by the blocking probabilities on other links. Denote by $Er_k(\boldsymbol{\lambda}_l)$ the Erlang blocking probability of a multiple-class link when the mean incoming rate of potential calls is $\boldsymbol{\lambda}_l = \lambda_1, ..., \lambda_K$. In a multiclass network, for the t'th iteration we would have:

$$\lambda_{th,l,k}^{(t)} = \sum_{R:l\in R} \lambda_{0,R} \prod_{m\in R}^{m\neq l} (1 - B_{m,k}^{(t-1)}) \tag{33}$$

$$B_{l,k}^{(t)} = Er(\boldsymbol{\lambda}_{th,l}{}^{(t)}) \tag{34}$$

where $B_{l,k}^{(t)}$ the estimation of the blocking for class k on link l for iteration t, and $\lambda_{th,l,k}^{(t)}$ is the estimation of mean incoming rate of potential calls for class k on link l after taking thinning into account, for iteration t.

It is not trivial to incorporate a fixed point iteration in our algorithm when multiple classes need to be considered. In Section 3, we do not take into account demand for different classes independently, but we assume that potential users arrive with mean rate λ_0 and then are distributed to classes according to their requirements. Although it could be possible to take into account the different thinning for each class throughout the following algorithm, the complexity would increase and the simple solution of Section 3 would not apply. For this reason, we employ the following heuristic/hybrid algorithm:

Algorithm 2

(1) For all $l \in L$, calculate the optimal prices and qualities without taking thinning into account, using Algorithm 1.
(2) Calculate the resulting blocking probabilities using fixed point iteration, (33).
(3) For all $l \in L$, define as h_l the highest blocking probability among the classes. Calculate the LR solution (15) or (16) using $\lambda_{h,l} = \lambda_{0,l}(1 - h_l)$ instead of λ_0. Call $\lambda_{h,l,k} = \lambda_{h,l}(q_{lk} - p_{lk})$ the incoming rate for each class k.
(4) Improve the solution by taking into account the thinning of each class independently:

$$q_{lk} = p_{l(k-1)} \tag{35}$$

$$p_{lk} = q_{lk} - \lambda_{h,l,k}/\lambda_{th,l,k} \tag{36}$$

(5) Using q_{lk} and p_{lk} perform steps (2) and (3) of Algorithm 1, where in (17)

λ_k is substituted with $\lambda_{th,l,k}(q_{lk} - p_{lk})$.

(6) Go to step (2) of this algorithm unless for all k and l $|q_{lk}^{(t)} - q_{lk}^{(t-1)}| < \epsilon_1$ and $|p_{lk}^{(t)} - p_{lk}^{(t-1)}| < \epsilon_2$, for small ϵ_1, ϵ_2.

In Step (3), we solve the LR problem assuming that the incoming rate has thinned by the same proportion for all services. For each link we use the highest thinning among the different classes, which can be locally estimated as the number of requests that arrived divided by the number of requests expected to arrive if there where no thinning. Step (4) is based on the observation that the result from step (3) gives a good indication of the optimal distribution of demand among classes in the optimal solution. Since the highest thinning has been used for all of the classes in step (3), the thinning of all classes, except the one which had the highest thinning, has been overestimated. Beginning from the class of highest quality according to the LR solution and considering each class with decreasing order, set the quality q_{lk} of each class to be the highest possible without overlapping with the immediately higher class (see Section 3) and set the price p_{lk} in such a way that the offered load of this class remains the same as it was using the solution in step (3). In step (5), (17) should be written taking into consideration that the offered traffic for each class is thinned by a different parameter (the blocking probability specific to each class). This will not increase the complexity (the equation to be solved remains a 3^{rd} degree polynomial) and will provide more accurate results.

The main advantage of Algorithm 2 is the fact that in each local maximisation on the links, the decisions on the other links are now taken into account through the blocking which results in thinning of the load. Note that, for this improvement, no information needs to be explicitly exchanged between nodes. The thinning of the demand on each link can be estimated through measurements, by comparing the theoretical demand (if there was no thinning) with the actual demand on each link, or by measuring the proportion of calls which the link accepted but did not actually take place due to blocking from another link. In such a case the calculation of blocking probabilities using the fixed point iteration (33) would be substituted by measurement of the blocking probabilities.

We have conducted simulations for the simplest case of a network consisting of two links and 3 possible routes, as shown on Fig. 11. The results for several different cases regarding the values of the capacity of each link (W_1,W_2) and the overall demand on each route (λ_1,λ_2,λ_3) are shown on Table I. We can see that our decentralised algorithm obtains income results very close to the optimal, with a difference between 0.17% and 2.7%. The worst case of 2.7% difference, corresponds to heavy blocking (50-60%) and severe link dependence (since most of the traffic comes from route 3 which crosses both links). At the same time, for the cases studied, the computational time is reduced from hours (for exhaustive searching) to 1-2 minutes, the exact computational time

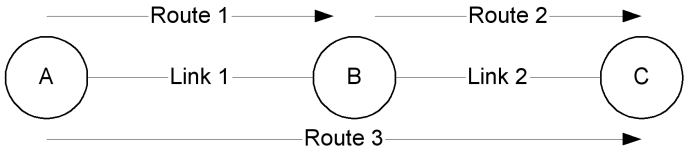

Fig. 11. A simple multi-link case.

	Settings					Results		
	W_1	W_2	λ_1 calls/s	λ_2 calls/s	λ_3 calls/s	Income using solution (31)	Optimal income (26)	Difference
Case 1	6	4	11	11	11	5.69	5.70	0.17%
Case 2	6	4	11	11	21	6.08	6.13	0.82%
Case 3	10	3	11	11	21	6.64	6.71	1.01%
Case 4	3	5	11	5	15	4.23	4.27	0.94%
Case 5	3	5	5	11	15	4.66	4.67	0.26%
Case 6	3	5	2	2	20	3.98	4.10	2.7%

Table 1
Results Comparison for a 2-links network.

depending mainly on the size of the links, which affects the computational complexity of calculating blocking probabilities ($O(CK)$, where C is the size of the link and K the number of classes). In particular, the exhaustive search is $O(CK\Xi^{2KL})$, where Ξ is the number of values tried for each price and quality, which depends on the quantisation step and L is the total number of links. Algorithm 2 is $O(CKL)$, since it depends on the number of links L and the calculation of blocking probabilities which is $O(CK)$ for each link.

Further simulations have been conducted for the four-node network of Fig. 12. We assume that the potential user mean arrival rate for each of flows 1-4 is λ_1, for flows 5-8 λ_2, for flows 9-12 λ_3 and for flows 13 and 14 λ_4. The capacity of links A-D is W_1 and links E and F have capacity W_2. Due to symmetry, the solutions for links A-D will be identical, as well as the solutions for links E-F, making the exhaustive search more tractable. The results show that in all cases the decentralised solution (31) is very close to the optimal (26) as discovered using exhaustive search algorithm. Note that the total calculation time for the 6 cases was 4 hours and 16 minutes for the exhaustive search algorithm, and only 1 minute and 38 seconds for the decentralised hybrid algorithm. The difference is computational time is expected to be even larger for the case of more than two CoS's due to the exponential increase of the size of the decision space.

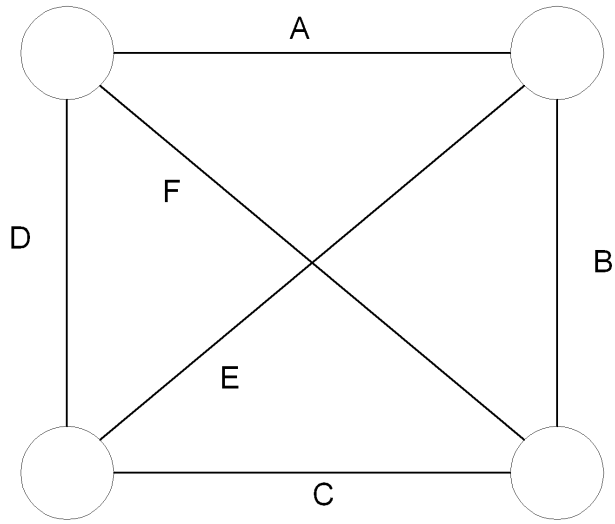

Fig. 12. The square network used for simulations.

	Settings						Results		
	W_1	W_2	λ_1 calls/s	λ_2 calls/s	λ_3 calls/s	λ_4 calls/s	Income using solution (31)	Optimal income (26)	Diference
Case 1	5	5	5	6	7	3	15.971	16.218	1.52%
Case 2	8	3	4	5	6	4	18.475	19.036	2.95%
Case 3	5	3	10	3	2	5	14.208	14.323	0.80%
Case 4	3	8	4	5	3	5	12.147	12.459	2.50%
Case 5	10	10	4	2	3	1	18.359	18.711	1.88%
Case 6	4	4	3	5	3	3	11.61	11.804	1.64%

Table 2
Results Comparison for 4-node network.

5 Final remarks

In this paper we analysed the problem of resource allocation and pricing, with the aim of income maximisation. The role of prices in a telecommunications environment can be twofold. Apart from being the well-known reward mechanism for the provider of the telecommunications resources, it is also a mechanism of selection and classification of users according to their utility requirements. The fact that users who demand higher services are prepared to pay more, can be used by the provider in any task that requires (or would be more efficiently performed with) segmentation or categorisation of the users.

The problem we have addressed in this paper attempts to aid the decisions of a network provider regarding the quality and the price of calls belonging to different classes. After analysing the problem and its parameters, an approximation was presented, for the case when the number of calls of the route is very large. This approximation simplified all equations to a great extent. We developed a numerical method for improving the LR solution of [35], which would be more relevant for access links of smaller capacity. For the case of a network of many links, under the assumptions of link independence and similar load across different links, we have found an approximate solution which is within 3% from the optimal for all the cases studied, even when the assumption of uniform distribution, or the assumption of balanced load does not hold. For the approximate decentralised solution we use the Erlang fixed point iteration method. A matter of further research is to compare our algorithm to more general existing global optimisation methods. Another relevant extension of our work is to conduct a sensitivity analysis of our solution to the relationship between q and p.

We are currently considering only static decisions in our model, but dynamic/adaptive versions of the algorithms presented in our paper are currently under investigation, where prices and quality change depending on the state of the network, and potential users are informed of the offered services and their prices. In general, although adaptive solutions offer better results in terms of income, this improvement comes at a cost, since the complexity of the decisions is increased, and users also face a more complicated system, where they cannot be sure about the prices in advance. Especially for the case of dynamic adjustments of prices and quality, preliminary results show that it is possible to improve the income from a network, but the complexity of the problem increases significantly. Another direction of research, which is currently under investigation, is balancing the relative load among alternative routes using pricing. This would improve the accuracy of our algorithm which, as we show, partially relies on the assumption that the relative load is similar on all links.

References

[1] E. Altman, D. Barman, R. El Azouzi, D. Ros, B. Tuffin, "Pricing differentiated services: A game-theoretic approach", Computer Networks, Vol. 50, Issue 7, 15 May 2006, pp 982-1002.

[2] H. Che, S. Zheng, X. Hong: "Integrated model for performance analysis of multiple class-of-service Internet", IEE Proceedings in Communications. 2002. pp.139-146.

[3] H. Che, S. Zheng, X. Hong:"A model analysis of pricing and link bandwidth allocation in a multiple class-of-service network", Proceedings of IEEE

International Conference on Computer Communications and Networks 2000, pp. 510-516.

[4] R. G. Cross, "Revenue Management, Hard Core Tactics for Market Domination", Broadway Books, New York, 1998, ISBN 0- 7679-0033-2.

[5] A. Farago, "A general method for the blocking analysis of networks with dependent links", High Performance Switching and Routing, 2001 IEEE Workshop on , vol., no.pp.124-129, 2001

[6] E. W. Fulp, D. S. Reeves, "Bandwidth provisioning and pricing for networks with multiple classes of service", Computer Networks, vol. 46, Issue 1, 16 September 2004, pp 41-52.

[7] Y. Hayel, B. Tuffin, "A mathematical analysis of the cumulus pricing scheme", Computer Networks, vol. 47, Issue 6, 22 April 2005, pp 907-921.

[8] S. Humair, "Yield Management for Telecommunication Networks: Defining a New Landscape", PhD Thesis, MIT, 2001.

[9] F. P. Kelly, "Loss networks", Annals of Applied Probability, vol. 1, no.3, pp. 319378, 1991.

[10] N. J. Keon, G. Anandalingam, "Optimal pricing for multiple services in telecommunications networks offering quality-of-service guarantees", IEEE/ACM Transactions on Networking, vol.11, no.1, pp. 66- 80, Feb 2003.

[11] J. O. Kephart, J. E. Hanson and A. R. Greenwald, "Dynamic pricing by software agents", Computer Networks, vol. 32, Issue 6, 30 May 2000, pp 731-752.

[12] T. Li, Y. Iraqi, R. Boutaba, "Pricing and admission control for QoS-enabled Internet", Computer Networks, vol. 46, Issue 1, 16 September 2004, pp 87-110.

[13] J. K. MacKie-Mason, H. R. Varian, "Pricing congestible network resources", IEEE Journal on Selected Areas in Communications, 1995. 13(0733-8716), p. 1141-1149.

[14] M. Mandjes, "Pricing strategies under heterogeneous service requirements", Computer Networks, vol. 42, Issue 2, 5 June 2003, pp 231-249.

[15] A. Odlyzko, "Paris Metro Pricing for the Internet", Proceedings of the 1st ACM conference on Electronic commerce, pp. 140 - 147.

[16] A. Odlyzko, "Internet pricing and the history of communications", Computer Networks, vol. 36, Issues 5-6, August 2001, pp 493-517.

[17] I. Ch. Paschalidis, J. N. Tsitsiklis, "Congestion-dependent pricing of network services", IEEE/ACM Transactions on Networking, 2000. 8(1063-6692): p. 171-184.

[18] I. Ch. Paschalidis, L. Yong, "Pricing in multiservice loss networks: static pricing, asymptotic optimality and demand substitution effects", IEEE/ACM Transactions on Networking, vol.10, no.3, pp.425-438, Jun 2002.

[19] I. Ch. Paschalidis, L. Yong, "Distributed resource allocation in multiservice communication networks using pricing", in Proceedings of the 41st IEEE Conference on Decision and Control, 2002.

[20] J. M. Peha, "Dynamic pricing and congestion control for best-effort ATM services", Computer Networks, vol. 32, Issue 3, March 2000, pp 333-345.

[21] P. Reichl, D. Hausheer and B. Stiller, "The Cumulus Pricing model as an adaptive framework for feasible, efficient, and user-friendly tariffing of Internet services", Computer Networks, vol. 43, Issue 1, 16 September 2003, pp 3-24.

[22] D. Ros and B. Tuffin, "A mathematical model of the Paris Metro Pricing scheme for charging packet networks", Computer Networks, vol. 46, Issue 1, Internet Economics: Pricing and Policies, 16 September 2004, pp 73-85.

[23] K. Ross, "Multiservice Loss Models for Broadband Communication Networks", Springer, 1995.

[24] J. Sairamesh, J. O. Kephart, "Price Dynamics of Vertically differentiated information markets", Proceedings of the first international conference on Information and computation economies, Charleston, South Carolina, United States, 1998, pp 28-36.

[25] Ronaldo M. Salles and Javier A. Barria, "Fair and efficient dynamic bandwidth allocation for multi-application networks", Computer Networks, vol. 49, Issue 6, 19 December 2005, pp 856-877.

[26] N. Semret, R. R.-F. Liao, A. T. Campbell, A. A. Lazar, "Pricing, provisioning and peering: dynamic markets for differentiated Internet services and implications for network interconnections", IEEE Journal on Selected Areas in Communications, Dec. 2000.

[27] N. Semret, R. R.-F. Liao, A. T. Campbell, A. A. Lazar, "Peering and provisioning of differentiated internet services", INFOCOM 2000.

[28] C. Xi-Ren, S. Hong-Xia, R. Milito, P. Wirth, "Internet pricing with a game theoretical approach: concepts and examples", IEEE/ACM Transactions on Networking, vol.10, no.2, pp.208-216, Apr. 2002.

[29] H. Yaiche, R. R. Mazumdar, C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks", IEEE/ACM Transactions on Networking, vol.8, no.5, pp. 667-678, Oct. 2000.

[30] J. Youngmi, G. Kesidis, "Dynamics of usage-priced communication networks: the case of a single bottleneck resource", IEEE/ACM Transactions on Networking, vol.13, no.5, pp. 1041-1053, Oct. 2005.

[31] X. Xiao and L.M. Ni, "Internet QoS: A Big Picture," IEEE Network, Vol. 13, No. 2, Mar.-Apr. 1999, pp. 8-18.

[32] S. Shenker, "Fundamental design issues for the future Internet", IEEE Journal on Selected Areas in Communications, vol. 13, no. 7, pp. 11761188, Sep. 1995.

[33] Jang-Won Lee, R.R. Mazumdar, N.B. Shroff, "Non-Convex Optimization and Rate Control for Multi-Class Services in the Internet", IEEE/ACM Transactions on Networking, vol.13, no.4, pp. 827- 840, Aug. 2005.

[34] The M3I consortium, "Deliverable 3, Pricing mechanisms Pt II, Price reaction design", retrieved online from www.m3i.org/results/m3idel03_2.pdf at 24 July 2006.

[35] G. Zachariadis, J. Barria, "Income maximisation using prices and QoS for a multi-class telecommunications system", accepted for publication in IEEE Communications Letters.

[36] Samir Mohamed, Gerardo Rubino, Martin Varela: "Performance evaluation of real-time speech through a packet network: a random neural networks-based approach". Performance Evaluation 57(2): 141-161 (2004).

[37] Stidham, S., Jr., "Pricing and congestion management in a network with heterogeneous users", IEEE Transactions on Automatic Control, vol.49, no.6, pp. 976- 981, June 2004.