A tariff model to charge IP services with guaranteed quality: effect of users' demand in a case study

N. Blefari-Melazzi
DIE-University of Rome "Tor Vergata"
Viale del Politecnico,1
Rome, Italy
blefari@uniroma2.it

D. Di Sorte, M. Femminella, G. Reali
DIEI-University of Perugia
Via G. Duranti, 93
Perugia, Italy
{disorte,femminella,reali}@diei.unipg.it

Abstract

In this paper, we consider a per-call, usage-based tariff model to charge for IP services with guaranteed quality. This model is based on the virtual delay, which is a Quality of Service (QoS) index that describes an improved IP service provided by a network domain. We show how to compute the virtual delay, and how to make it dependent on the service demand. Then, we demonstrate the effectiveness of our tariff model to tune revenues, blocking probability, and resource utilization in a meaningful application scenario. Our goal is to give some directions for network resource dimensioning and pricing purposes, which depend on the service demand.

Keywords

IP-QoS, usage-based charging, service demand, numerical analysis

1. Introduction

The role of the network service charging, in IP-based networks, is not clear yet. As regards the widest deployment of the IP protocol, which is the Internet, the most common charging method is based on the flat-rate model, i.e., subscribers pay a fee for accessing the network, independent of their effective use of the service. If the Internet is to become a network supporting differentiated application and transfer services, then it would be advisable (and perhaps necessary) to deploy architectures and protocols providing QoS guarantees and allowing efficient and flexible charging, billing, and accounting functions [1][2]. In this future scenario, the flat-rate model might be inefficient from an economic point view, since it does not enable to charge services according to their type and quality. From this point of view, flat-rate pricing may be inadequate for QoS-enabled networks, which require admission control and resource reservation mechanisms (e.g., see [2]), and the management of which entails expensive investments. Thus, usage-based tariffs are needed to guarantee an additional income to network service providers, as suggested in [2].

For the time being, there are no standardised pricing models which account for network resources actually reserved and consumed; however, several such pricing models have been proposed in the literature (see [1][3][4] for a thorough analysis).

Our proposal consists of a tariff model to charge for network services with performance guarantees offered by an administrative domain. In this regard, we use the concept of "traffic flows": a flow is a set of packets traversing an administrative domain, belonging to the same application session (also referred to as a "call"), running between two hosts, and receiving the same QoS treatment.

We start from the tariff model to charge for IP guaranteed services proposed in [5], with the following characteristics:

- it charges the edge-to-edge service of an administrative domain on a per-flow basis;
- it depends on (i) the QoS level of the transfer service, (ii) the type of traffic to be supported, (iii) the amount of resources of the domain, and (iv) the service demand;
- it is based on the duration of the connection and/or the traffic volume exchanged.

Such a tariff is based on the novel concept of virtual delay, introduced in a qualitative way in [5][7]. The virtual delay value is a comprehensive, all-inclusive appraisal of the QoS parameters characterizing the edge-to-edge transfer service within a domain.

The virtual delay may be assumed to be a standard measure of the QoS level and we have shown the potentials of this descriptor, under this assumption, as basis of a inter-domain routing algorithm [7]. In [6] and [20], we changed the perspective from an end-to-end scope (i.e., minimization of the overall network service price) to an intra-domain scope (i.e., maximization of the network operator revenue). From this viewpoint, the virtual delay is regarded as a QoS index computed by each domain in its own way. This parameter is visible only within the administrative domain and is the starting point on which to build a usage-based tariff model on a per-call basis. Below we recall how the virtual delay is computed, and how it affects the tariff. Then, we analyse our model in a case study, in which we include the willingness of users to pay, in a market characterized by tariffs that fluctuate according to the amount of service demand. To this end, we extend the basic virtual delay computation to make the tariff dependent on the status of resource availability. Starting from the assumption that the network operator is able to estimate the users' service demand, we show the effectiveness of our pricing approach to tune revenues, blocking probability, and resource utilization. Our final goal is to give some hints for network dimensioning and pricing purposes.

The results presented in this paper will be exploited as a potential building block of an all-encompassing vision of future personalized and easy-to-use services, pursued in the framework of the Simplicity project (www.ist-simplicity.org). The aim of Simplicity is to provide each user with a personalized profile, stored in a so called Simplicity device. Ideally, by plugging this device (which can be a physical or a virtual one) into the terminal, each user will personalize both terminal and services alike. In this vision, the charging approach is clearly an important component. In addition, the availability of the Simplicity device offers an interesting implementation venue for the concepts described in this paper, since it helps the network operator to promptly estimate the users' service demand and willingness to pay.

The paper is organized as follows. The next section recalls the concepts of network commodity and virtual delay, together with the tariff model. In section 3, which is the core of the paper, we extend the virtual delay concept. Then, we describe an application of our pricing approach in a specific scenario, and present the relevant numerical results. Finally, some remarks conclude the paper.

2. The pricing approach

In this section, we first briefly recall the concept of the "commoditization" of a network service and a related pricing law, used to charge for IP guaranteed services [5][6][7][20]; this concept is based on the so-called virtual delay, which is a QoS index of an edge-to-edge transfer service offered by an administrative domain. Then, we show how to compute the virtual delay, in a specific scenario [6][20].

To provide QoS over IP networks, admission control and resource reservation functions are necessary. We remark that we consider the administrative domain as a black box, in the sense that we are not interested in the QoS architecture, nor in the specific Admission Control (AC) scheme and resource reservation strategy, which are implemented within it. Our approach is compliant with both per-flow (Integrated Services [9]) and per-aggregate (Differentiated Services [11]) traffic management within the core network.

The following assumptions, relevant to a given administrative domain, hold:

- flows entering the domain are regulated by Dual Leaky Buckets (DLBs) [8][9], with parameters: peak rate, P_S, sustainable rate, r_S, and burst tolerance, B_{TS}. These parameters define the extreme traffic profile associated with the considered traffic class, which cannot be violated by users. Then, DLB parameters cannot be negotiated (in accordance with [6] and differently from [20]). This means that the operator sets the extreme traffic profile associated with a given service (e.g., VoIP);
- the QoS of the port-to-port IP service provided to the specific traffic class is described by the following service parameters: the maximum transfer delay, D_{max}, the maximum delay jitter, D_{jitter}, and the loss probability, P_{loss}, due to buffer overflow. They can be negotiated between network operators and users. This means that the operator offers a number of service classes for a given service (e.g., VoIP);
- other QoS parameters, such as channel reliability, resilience and connection set-up time (network parameters), characterize the intrinsic quality of the network, do not depend on the specific flow, and cannot be negotiated;
- resource reservation and admission control functions are implemented to respect the service level agreement. Consequently, an amount of network resources is reserved for the flow from the ingress to the egress point of the domain.

2.1 The network commodity and the pricing law

We identify the network commodity offered by network operators as the transfer of information units from a node A to a node B in the network. It is mainly described by the service parameters, which can be summarized by the so-called virtual delay, *d*. This quantity is a comprehensive and all-inclusive appraisal of the transfer delay, the delay jitter and the loss probability.

The following considerations are in order. A network service is modeled as a hypothetical equivalent service with the virtual delay, *d*, which gives a measure of the QoS level: the higher the level of the service, the lower the value of *d*. Moreover, we consider a monotonic, non-increasing function of the virtual delay, *f(d)*, which associates the port-to-port transfer of an information unit with a technical measure, expressed in commodity units. Each domain is free to choose the function which best fits its own requirements. The cost (or value) of the transfer of an information unit from a point A to a point B with a virtual delay *d* is $S(d) = \alpha_{A \to B} f(d)$, where $\alpha_{A\to B}$ is the cost of each commodity unit, which depends on (i) network parameters; (ii) the two nodes A and B (e.g., their distance); (iii) the policies of the relevant domain. Clearly, when the commodity is on the market, its price can fluctuate according to factors that are beyond technical considerations. Thus, we define the price of the transfer of an information unit as $P(d) = \gamma S(d) = \gamma \alpha f(d) = \beta f(d)$, where γ is a price variation factor that accounts for market fluctuations, and $\beta = \alpha\gamma$ is the market commodity price (i.e., the price per commodity unit). We assume that both β and the QoS level are constant during the connection.

Let *T* be the duration of a call and t_0 its starting time. The per-call tariff applied to a flow entering the domain with an instant bandwidth equal to $B_{ist}(t)$ is

$$Q = \beta f(d) \int_{t_0}^{t_0+T} \max[B_{ist}(t) - B_{res}, 0] dt + \beta f(d) B_{res} T \,, \qquad (1)$$

where B_{res} is the bandwidth value that a domain charges on a per-time basis. Such a value ranges from zero to the peak rate of the flow. According to the tariff (1), extra-usage of bandwidth (with respect to the component B_{res}) is charged on a per-volume basis. Thus, the tariff consists of a component depending on the duration time of the connection (allocation charge) and a component depending on the amount of traffic volume exchanged (effective usage charge). The weights of the two components can be arbitrarily set, by varying B_{res}, which, in this view, may be regarded as a tunable knob. On the one hand, the allocation charge ensures a minimum amount of revenue to network operator, whereas, on the other hand, users would like to be charged according to their actual use of the service (effective usage charge). It is worth noting that if the value of B_{res} increases, the price charged to network users increases as well.

In principle, each network is obviously free to choose the value of B_{res}, according to its own pricing policy. From the technical point of view, the quantity B_{res} may be properly chosen equal to the effective bandwidth used for admission control purposes ([8][9]). It is our opinion that this is the most reasonable choice, since, in this way, operators charge users in proportion to the reserved resources, and protect themselves

against unfair user behavior. Since the typical range of this parameter lies within the interval [r_S, P_S], a higher value of effective bandwidth always implies a higher tariff.
It is clear that the tariff strongly depends on the instant bandwidth, $B_{ist}(t)$, of the flow entering the domain (note that we are assuming that the accounting devices operate at the ingress of the domain).
If r_S is the average transmission rate of the flow, for a DLB shaped flow we can identify two extreme cases of traffic emission, compliant with the DLB operation: constant rate and extremal ON/OFF rate. Correspondingly, it is possible to identify the extreme tariff values as follows [20]:

$$Q_{\max} = T[\beta(B_{res}(1 - r_S / P_S) + r_S)]f(d) \text{, and } Q_{\min} = T[\beta B_{res}]f(d) . \qquad (2)$$

The quantities in square brackets in (2) represent the price per-commodity unit per-time unit. It is not surprising and it is correct that the highest tariff corresponds to the maximum burstiness of the transmission rate. In fact, it is well known that bursty flows (in particular ON/OFF shaped ones) stress resources more than flows with a smoothed transmission rate [8].

2.2 Computation of the virtual delay

As mentioned above, the service parameters that we use (negotiable between a network user and the administrator of the network service provider) are the maximum edge-to-edge delay, D_{max}, the maximum delay jitter, D_{jitter}, and the loss probability, P_{loss}. Please, note that the selection of this set of parameters is due to our view of the network service. Nevertheless, it does not prevent the introduction of other parameters, if regarded as necessary and compliant with the approach of the virtual delay by the network service provider.
In our case, the total edge-to-edge delay includes transmission time at the source, propagation delay, processing and transmission time, and queuing delay at network nodes. Since delay jitter is mainly caused by queuing in nodes, it is possible to assume that D_{jitter} is a component of the maximum transfer delay. In other words, the maximum queuing delay is equal to the magnitude of the maximum delay jitter. Similarly, in an equivalent "virtual" model, it is possible to compensate the packet loss probability, due to buffer overflow in nodes, by increasing the amount of buffer allocated to the flow. From this point of view, the loss probability may be traded for queuing delay, and therefore it may be represented by a contribution in the virtual delay evaluation.
Now, our goal is to find the component of the virtual delay related to the loss probability, D_P, so that the value of virtual delay can be expressed by ([6][20])

$$d(D_{\max}, P_{loss}) = D_{\max} + D_P(P_{loss}) . \qquad (3)$$

We associate a port-to-port service described by (D_{max}, D_{jitter}, P_{loss}) with an equivalent node that introduces an (almost) constant delay D_C=D_{max}-D_{jitter}, a maximum queuing delay equal to D_{jitter}, and a loss probability equal to P_{loss}. We call C the amount of capacity from the input port to the output port, and consequently $B=CD_{jitter}$ is the buffer space of the equivalent node.

In [8], the Authors have shown that the maximum buffer occupancy of a single flow, characterized by a set of DLB traffic descriptors (P_S, r_S, B_{TS}), which feeds a buffer with unlimited size, served at a transmission capacity c is given by:

$$b = B_{TS}(P_S - c)/(P_S - r_S)\,. \qquad (4)$$

The pair (b,c) belongs to an infinite set of values of effective buffer and bandwidth, respectively, relevant to each shaped source. In addition, if we set the maximum queuing delay of an information unit as $D_{jitter} = B/C$, then the following additional relation applies: $b/c = B/C$. From this equation and (4), it is straightforward to determine the pair (b_0,c_0) to be assigned to each flow at the equivalent node in order to guarantee a service characterized by a maximum queuing delay equal to D_{jitter} and no packet losses. It results that

$$c_0 = P_S B_{TS}/(D_{jitter}(P_S - r_S) + B_{TS})\text{, and } b_0 = c_0 D_{jitter}\,. \qquad (5)$$

Then, let c_p be the value of effective bandwidth and let b_p be the effective buffer corresponding to a service characterized by a value of queuing delay equal to D_{jitter} ($b_p/c_p=D_{jitter}$) and by a value of packet loss probability equal to P_{loss}. We undertake to calculate the (c_p,b_p) pair by using the novel approach illustrated in [9], in particular with the small buffer approximation ([9], section III).

Our goal is to associate the network service with losses to a virtual service without losses, by increasing the effective buffer space allocated to the flow. From (4), once the amount of bandwidth c_p reserved to the flow has been determined, the buffer needed to avoid losses would be equal to $\overline{b_p} = B_{TS}(P_S - c_p)/(P_S - r_S)$. It means that the buffer allocated to the flow should be increased by a value equal to $\overline{b_p} - b_p$. This would imply an additional queuing delay. Consequently, the virtual component of delay associated to the loss probability is equal to

$$D_P = (\overline{b}_p - b_p)/c_p\,. \qquad (6)$$

This leads to the final equation of the virtual delay:

$$d(D_{\max}, P_{loss}) = D_C + D_{jitter} + D_P = D_C + (B_{TS}/c_p)[(P_S - c_p)/(P_S - r_S)]\,. \qquad (7)$$

Therefore, this parameter can be considered as an index describing the QoS level of the service, and it is also related to traffic descriptors and network resources. Note that it is possible to assign different weights to the different contributions (D_C, D_{jitter}, D_P) to the value of virtual delay, according to the pricing policy of the domain or the type of application service to be supported by the network. For simplicity reasons and without loss of generality, in the following we assume that all weights are set at 1.

The analysis of the sensitivity of the virtual delay and of the tariff to the edge-to-edge capacity, C, and to the loss probability, P_{loss}, is shown in [6][20]; the virtual delay increases with both C and P_{loss} (since c_p decresases), whereas the tariff decreases.

3. Application of the tariff model: a case study

In this section, our goal is to test the effectiveness of the proposed pricing approach in an application scenario characterized by the users' demand curve (i.e., the willingness of users to pay for a given service), in a market characterized by dynamic tariffs. It is worth noting that we assume to deal with a set of network services offering hard QoS guarantees, and meant to support real-time/streaming multimedia application services (inelastic applications). These applications cannot adapt the transmission rate to either varying network conditions or users' willingness to pay. Then, we assume that each communication session is charged by network administrator on a per-call basis, according to the edge-to-edge QoS level negotiated between users and the network service provider. Such a QoS level must be maintained throughout the duration of the call. In order to make the pricing scheme and its implementation simple, we assume that the price charged for a call is constant throughout the duration of the connection and is set at the call set-up. In our opinion, this model simplifies the tasks of the users and makes the entire pricing system more transparent to them. In fact, users would immediately perceive both the quality of the service they will be given and the relevant price. Thus, we allow negotiation of price and QoS only at the beginning of the communication session. However, we emphasize that this is only a case study, and our assumption does not prevent network operator from using different approaches.

In our model, the price charged depends on the amount of traffic (i.e., active calls) present in the network, and it increases with the current amount of service demand. Please note that this approach is different from the "congestion pricing" [12][13][14][15][16][17], since here congestion is avoided by a AC scheme.

In such a scenario, the next step in the development of the model is the introduction in the tariff of a dynamic factor, which takes into account the current service demand. For this purpose, we identify an elegant and viable solution, which consists of making the virtual delay value, which characterizes the network service, also depending on the service demand. To this end, let us consider the model used to compute the value of the virtual delay and let us make a small change, as described in the following. As shown in [6][20], the virtual delay increases with the amount of total capacity C of the equivalent node. Consequently, if we compute the value of virtual delay at each time using the amount of capacity currently available, $C_{ava}(t)$, instead of the total capacity C, then the tariff charged would increase together with the bandwidth demand. Considering the function $f(d)=e^{-md}$ (as done in [6][20]), then this dependence can be tuned by acting on the parameter m. It is worth noting that the virtual delay depends on the amount of network capacity through the effective bandwidth (see (7)). Consequently, in this time-depending computation of virtual delay, also the effective bandwidth has to be determined using the amount of available capacity $C_{ava}(t)$ instead of the total capacity C. From now on we will refer to the effective bandwidth computed from C as c_p, and to the effective bandwidth computed from $C_{ava}(t)$ as c_p^*. Therefore, the new relation to compute the value of virtual delay is

$$d(D_{\max}, P_{loss}) = D_C + D_{jitter} + D_P = D_C + \left(\frac{P_S - c_p^*}{P_S - r_S} \right) \frac{B_{TS}}{c_p^*} \tag{8}$$

3.1 Network scenario

We consider an administrative domain that offers a transfer service (VoIP-like) characterized by the following QoS service parameters: D_{max}=175 ms, (D_C=140 ms and D_{jitter}=35 ms), and P_{loss}=10^{-3}.

Flows are assumed to be homogeneous and described by the following DLB traffic descriptors: P_S=32 Kbps, r_S=13.6 Kbps, and B_{TS}=5300 bytes.

The transfer capacity associated with the port-to-port service is equal to 2.048 Mbps.

Calls are generated according to a Poisson arrival process with parameter λ, and each call duration is modeled as an exponentially distributed, random variable with mean value equal to $1/\mu$. The average call duration is set to 240 s (μ =1/240 s^{-1}), whereas the call arrival rate λ is a variable parameter.

As regards the tariff charged by the domain, we assume that the amount of bandwidth charged on a per-time basis is equal to the effective bandwidth computed on the equivalent node, i.e., $B_{res} = 18.12$ Kbps. In addition, we have chosen $f(d)=e^{-md}$, with m variable, and the price per-commodity unit equal to β=10^{-5} \$.

In general, the number of flows supported by the port-to-port service depends on the (proprietary) AC policies deployed by the network operator within its domain. Without loss of generality, we have chosen N_{max} equal to $\lfloor C / B_{res} \rfloor$ =113. Consequently, the maximum value of resource utilization factor is $\rho_{\max} = N_{\max} r_S / C = 0.75039$.

The virtual delay $d(t)$ associated with the service requested at the time t is computed considering the amount of bandwidth currently available C_{ava}, which is equal to $C_{ava}(t) = C(1 - N(t) / N_{\max})$, where $N(t)$ is the number of active flows at the time t.

In the following, we carry out the analysis relevant only to Q_{min}, the expression of which is shown in (2). It is possible to extend this an analysis not only to the case of Q_{max}, but also to the effectively charged tariff, if a statistical knowledge of the rate emission pattern of flows is known.

According to the assumptions described above, the per-second tariff $Q_{N(t)}$ set by the network operator for a service request at time t depends on the number $N(t) \in \{0,1,2,..., N_{\max}\}$ of the currently active connections. This is due to the fact that

$d(t)$ depends on $C_{ava}(t)$. Figure 1(a) shows the virtual delay d as a function of the number of active calls. As expected, the virtual delay is decreasing with the number of active calls up to the value of active calls equal to N_t=103. From this value on, the virtual delay remains constant. This is due to the fact that when the amount of network resources $C_{ava}(t)$ decreases, then the statistical multiplexing gain decreases as well. In other words, the value of c_p^* increases up to the maximum value c_0, which corresponds to the effective bandwidth in the lossless case (5). Since the virtual delay depends on the amount of network resources through the value of effective bandwidth c_p^* , then, once the maximum value c_0 is reached, the virtual delay reaches its minimum value and remains constant for values of N higher than the threshold value N_t. This constant value of the virtual delay is $d_{min}=D_C+D_{jitter}$ and, consequently, D_P=0. Note that the value of $d(t)$ for $N(t) = N_{max}$ can not be computed, since C_{ava} is zero. However, we extend the function $d(t)$ up to N_{max} with the value d_{min}.

Clearly, the per-second tariff is increasing with N up to N_t, and then it remains constant. This behavior is shown in Figure 1(b), where m is a variable parameter. The higher m, the lower the per-second tariff charged. In addition the dependence of the tariff on the virtual delay increases with the parameter m.

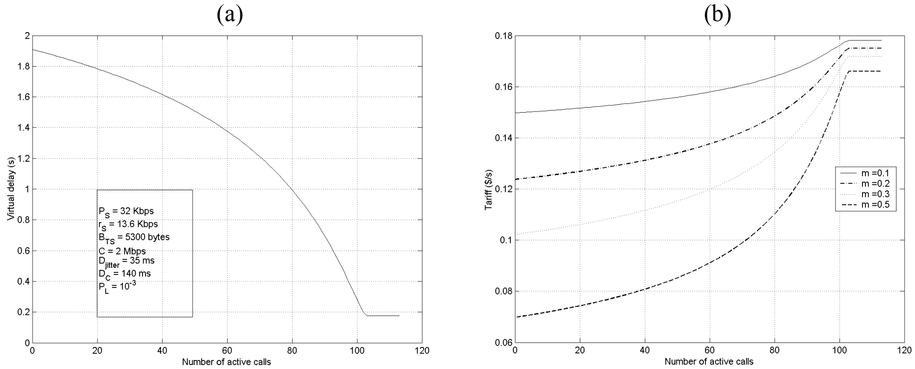

Figure 1: Virtual delay (a) and per-second tariff (b) as a function of the number of active calls.

As regards the users' willingness to pay to get the transfer service, U, we model this parameter as a random variable, with probability density function $P_U(u)$. U is expressed in currency units per-second, $/s. We assume that the network operator knows the behavior of consumers and, therefore, the distribution of this random variable (see the final comment in the introduction). For the sake of simplicity and for obtaining numerical results, we assume that the willingness of a user to pay is a random variable uniformly distributed between 0.12 $/s and 0.20 $/s. However, the generality of the model clearly allows the analysis for whatever distribution.

This system can be modeled, at the call level, as an $M/M/N_{max}/N_{max}$ queue with discouraged arrivals [18]. Each state i is characterized by arrival and departure rates:

$$\lambda_i = \lambda \cdot g_i = \lambda \int_{Q_i}^{+\infty} P_U(u)du,\ \mu_{i+1} = i\mu\ . \qquad i=0,1,2,\ldots,N_{max} \tag{9}$$

Note that $g_i = \int_{Q_i}^{\infty} P_U(u)du$ is the probability that a user will accept the price charged by the network operator when the system is in the state *i*. The probability P_i to have a number *i* of active calls in the system is easily found to be [18]:

$$P_0 = 1/(1 + \sum_{k=1}^{\infty} \prod_{j=0}^{k-1} (\lambda_j / \mu_{j+1})), \quad P_i = P_0 \prod_{j=0}^{i-1} (\lambda_j / \mu_{j+1}), \qquad i = 1,\ldots,N_{\max}\ . \tag{10}$$

The traffic load offered to the system can be identified by the average call arrival rate equal to $\lambda_{ave} = \sum_{i=0}^{N_{\max}} \lambda_i p_i$ (i.e., the arrival rate of the calls that overcome only the price acceptance control). The average number, N_{ave}, of active calls is equal to $N_{ave} = \sum_{i=0}^{N_{\max}} iP_i = (\lambda_{ave} - \lambda_{N_{\max}} P_{N_{\max}}) / \mu = \lambda_{ave}^{*} / \mu$, where λ_{ave}^{*} is the average rate of call arrivals that actually enter the system (i.e., those calls that overcome both the price acceptance control and the subsequent AC). Consequently, the average utilization factor of the edge-to-edge link is given by $\rho_{ave} = N_{ave} r_{S0} / C$.

The call blocking probability, P_{block}^{res}, due to the lack of resources is equal to

$$P_{block}^{res} = P_{N_{\max}} \lambda_{N_{\max}} / \lambda_{ave}\ , \tag{11}$$

The call blocking probability, P_{block}^{price}, due to a too high price charged is

$$P_{block}^{price} = \lambda \sum_{i=0}^{N_{\max}} P_i (1 - g_i) \Big/ \lambda = \sum_{i=0}^{N_{\max}} P_i (1 - g_i)\ . \tag{12}$$

The structure of the system model that we consider is depicted in Figure 2.

In addition, the average total revenue obtained by the network operator in a given time period T_O is given, in the steady-state condition, by $Q_{T_O} = N_{T_O} Q_{call,ave}$, where N_{To} is the number of calls that are served in the period T_O (equal to $\lambda_{ave}^{*} T_O$, if T_O is long enough), and $Q_{call,ave} = \frac{1}{\mu} \sum_{i=0}^{N_{\max}-1} Q_i P_{Q_i}$ is the average revenue per call. P_{Q_i} is the probability that an incoming call is charged with the per-second tariff that characterizes the state *i*, Q_i \$/s. Such a probability is given by [18]:

$$P_{Q_i} = \lim_{\Delta t \to 0} \Pr[N(t) = i/A(t, t+\Delta t)] = \lim_{\Delta t \to 0} \frac{\Pr[A(t, t+\Delta t)/N(t) = i]P_i}{\Pr[A(t, t+\Delta t)]} = \frac{\lambda_i P_i}{\lambda^*_{ave}}, \qquad (13)$$

where $A(t,t+\Delta t)$ is the event that an arrival occurs in the interval (t, $t+\Delta t$). Thus, the total (Q_{T_O}) and the average per-second (Q_s) revenues can be rewritten as

$$Q_{T_O} = T_O/\mu \sum_{i=0}^{N_{\max}-1} Q_i P_i \lambda, \quad Q_s = \frac{1}{\mu} \sum_{i=0}^{N_{\max}-1} Q_i P_i \lambda_i\ . \qquad (14)$$

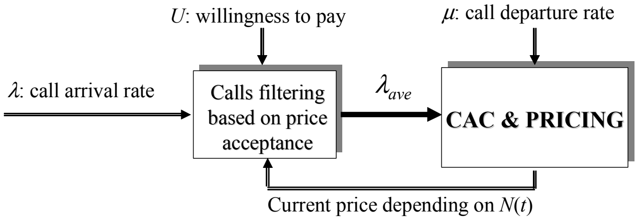
U: willingness to pay
μ: call departure rate
λ: call arrival rate
Calls filtering based on price acceptance
λ_{ave}
CAC & PRICING
Current price depending on $N(t)$

Figure 2: System model.

3.2. Numerical results

Our goal is to verify the behavior of the system under different traffic load conditions, when the parameter *m* in the exponential function *f*(*d*) varies. In particular we are interested in analyzing the following points: (i) the call blocking probability due to a too high price; (ii) the call blocking probability due to lack of resources; (iii) the amount of per-second revenue obtained by the network operator; (iv) the resource utilization factor. We first verify the goodness of the theoretical approach by simulations under different offered traffic loads and different tariffs. Table 1 reports results for *m*=0.2. The values reported in Table 1 are obtained by averaging the outputs of the simulations in the steady state. An excellent match of simulation results with theoretical ones can be observed.

Now, we analyze the behavior of the system under different traffic loads and when different tariffs are charged. Figure 3(a) shows the utilization coefficient versus the traffic load, normalized by the maximum number of admitted calls, $(\lambda/\mu)/N_{\max}$, with *m* set as parameter. As expected, the utilization factor increases with *m*, i.e., decreases with the price charged, and it clearly increases with the amount of offered traffic. For low traffic load, the utilization is almost constant when *m* ranges from 0.4 to 0.7. This is compliant with the fact that the corresponding call blocking probabilities due to high prices and lack of resources remain almost constant as well (see Figure 3(b) and Figure 3(c)). As expected, the price-induced call blocking probability decreases with *m*, i.e., it decreases when prices are lowered (Figure 3(b)), and increases with the amount of offered traffic. On the other hand, the call blocking probability due to lack of resources increases with *m*, since the utilization factor grows up (a larger amount of calls overcome the price acceptance function), and call rejection due to lack of resources is clearly more frequent. In addition, it increases with the traffic load. Note that the values of call blocking probability for lack of resources are negligible with respect to the values of the price-induced one.

λ (s^{-1})	ρ_{ave}		Q_s (\$/s)		P_{block}^{res}		P_{block}^{price}	
	Simulated	Theoretical	Simulated	Theoretical	Simulated	Theoretical	Simulated	Theoretical
0.4	0.46009	0.46124	9.8041	9.8563	0	$5.46 \cdot 10^{-15}$	0.27725	0.27649
0.5	0.52314	0.52093	11.628	11.557	0	$3.61 \cdot 10^{-11}$	0.34928	0.34628
0.6	0.56034	0.56336	12.846	12.920	0	$1.20 \cdot 10^{-8}$	0.41065	0.41086
0.7	0.59401	0.59382	14.075	14.008	0	$7.09 \cdot 10^{-7}$	0.46999	0.46773
0.8	0.61462	0.61645	14.878	14.891	0	$1.44 \cdot 10^{-5}$	0.51471	0.51650
0.9	0.63215	0.63424	15.612	15.629	$1.1 \cdot 10^{-4}$	$1.44 \cdot 10^{-4}$	0.55705	0.55776
1	0.64641	0.64931	16.230	16.272	$8.1 \cdot 10^{-4}$	$8.77 \cdot 10^{-4}$	0.59294	0.59223

Table 1: Simulation results versus theoretical results.

As regards the revenues, the situation is just a little more complicated when the offered traffic and tariffs are variable. The per-second total revenue obtained by network operators versus the normalized traffic load is shown in Figure 3(d). We can see that, for large values of offered traffic, low tariffs correspond to increased revenue. This is due to the fact that a larger number of calls are served. Therefore, the effect due to a higher number of calls served by the system is stronger than the one due to a lower amount of price charged per-call. On the other side, when the offered load is low (thus the number of active calls is small), charging very low rates (m = 0.6 and 0.7) implies an excessively low revenue. In addition, for low loads, the highest per-second revenue corresponds to a value of m equal to 0.3, which, together with m=0.4, are the best choices for what concerns the revenue when the traffic load is highly variable. With m=0.4 we obtain higher revenue for realistic values of normalized offered traffic (in the range [0.7, 1]) and lower values of price-induced blocking probability (see Figure 3(b)). Thus, m=0.4 is the best choice in this specific case study. In this way, the network operator can both maximize revenue and guarantee service provisioning (i.e., the quality of the service in terms of price blocking probability).

To sum up, it is our opinion that the choice of the parameter m made by network operator should be driven by the following considerations. When the user behavior is not influenced by previous experiences, then the operator is interested in maximizing the revenue and does not take into account the call blocking probability due to prices too high with respect to the willingness of users to pay. This could happen when the operator exercises a monopoly, and also when users (or a user agent [19] or a broker [5][7] on their behalf) mechanically scan all network offers and choose the one that currently maximizes their benefit, without considering previous failed/successful price negotiations (this is the case we have analyzed). On the other side, when users are "with memory", then the network operator could also sacrifice part of the current potential revenues with the aim of guaranteeing service provisioning, and consequently to safeguard future revenues.

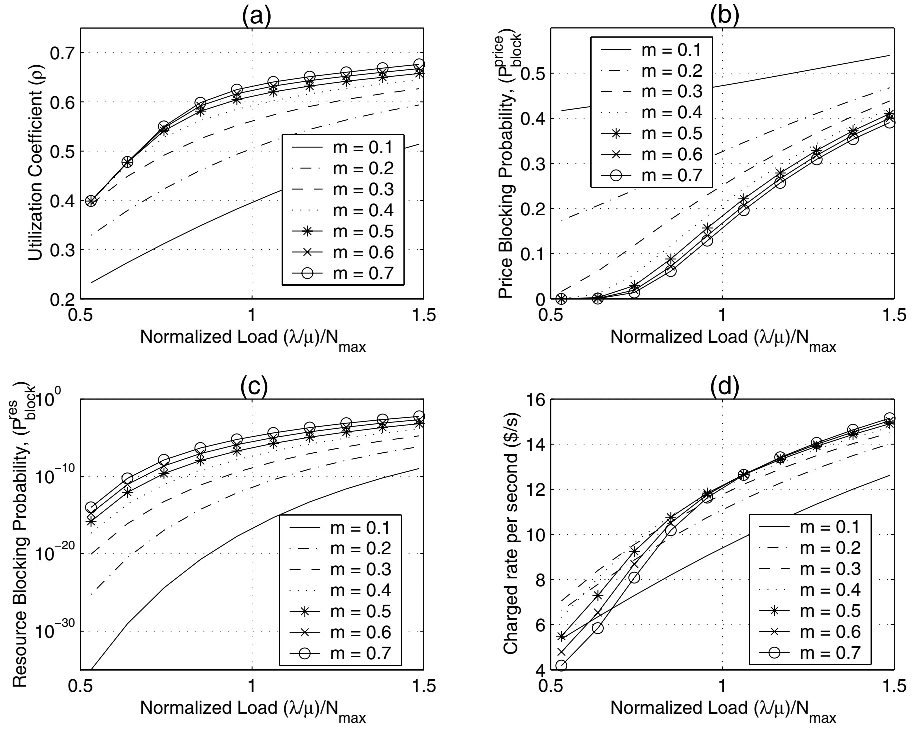

Figure 3: Utilization factor (a), blocking probability due to too expensive tariffs (b), blocking probability due to lack of resources (c), and total per-second revenue (d) vs. normalized offered traffic.

4. Conclusions

An innovative per-call tariff model to charge for improved IP services is based on the concept of the virtual delay, which identifies the QoS level. In this work, we have extended the concept of virtual delay, by making it dependent on the status of resource availability. We have analysed the resulting pricing model in a case study by means of a flow level study. The essential result of this analysis is that the price can be an effective network control tool, which allows an operator to finely tune the trade-off among revenues, blocking probability, and resource utilization. Once the service demand is estimated, the operator has to configure only few initial pricing parameters, on the basis of the performance desired. This task is simplified by our analysis tool. After this initial configuration, the ongoing price value is automatically obtained, according to the current network status.

ACKNOWLEDGMENT

This work is co-funded by the European Union under the IST project SIMPLICITY.

References

[1] N. Blefari-Melazzi, D. Di Sorte, G. Reali, Accounting and pricing: a forecast of the scenario of the next generation Internet, *Computer Communications, Elsevier Science*, 26(18), December 2003.

[2] G. Huston, Next steps for the IP QoS architecture, *IETF RFC 2990*, Nov. 2000.

[3] M. Falkner, M. Devetsikiotis, I. Lambadaris, An overview of pricing concepts for broadband IP networks, *IEEE Communications Surveys*, Second Quarter 2000.

[4] L.A. DaSilva, Pricing for QoS-enabled networks: a survey, *IEEE Communications Surveys*, Second Quarter 2000.

[5] D. Di Sorte, M. Femminella, G. Reali, S. Zeisberg, Network service provisioning in UWB open mobile access networks, *IEEE JSAC*, 20(9), December 2002.

[6] D. Di Sorte, M. Femminella, G. Reali, A QoS index for IP services to effectively support usage-based charging, *IEEE Communications Letters*, 8(11), Nov. 2004.

[7] D. Di Sorte, G. Reali, Minimum price inter-domain routing algorithm, *IEEE Communications Letters*, 6(4), April 2002.

[8] A. Elwalid, D. Mitra, R. H. Wentworth, A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM node, *IEEE JSAC*, 13(6), August 1995.

[9] K. Kumaran, M. Mandjes, Multiplexing regulated traffic streams: design and performance, *IEEE INFOCOM 2001*, Anchorage, USA, April 2001.

[10] R. Braden, D. Clark, S. Shenker, Integrated services in the Internet architecture: an overview, *IETF RFC 1633*, June 1994.

[11] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An architecture for Differentiated Services, *IETF RFC 2475*, December 1998.

[12] W. Wang, H. Schulzrinne, Pricing network resources for adaptive applications in a differentiated services network, *IEEE INFOCOM 2001*, Anchorage, USA, April 2001.

[13] A. Ganesh, K. Laevens, R. Steinberg, Congestion pricing and user adaptation, *IEEE INFOCOM 2001*, Anchorage, USA, April 2001.

[14] F. Kelly, Charging and rate control for elastic traffic, *European Transactions on Telecommunications*, 8, 1997.

[15] M. Caesar, D. Ghosal, R.H. Katz, Resource management for IP telephony networks, *IWQoS 2002*, Miami Beach, USA, May 2002.

[16] I.C. Paschalidis, J.N. Tsitsiklis, Congestion-dependent pricing of network services, *IEEE/ACM Transactions on Networking*, 8(2), April 2000.

[17] E.W. Fulp, D.S. Reeves, Optimal provisioning and pricing of Internet differentiated services in hierarchical markets, *IEEE International Conference on Networking*, Colmar, France, July 2001.

[18] L. Kleinrock, *Queueing Systems, Volume I: Theory*, Wiley Interscience, New York, 1975.

[19] J. Altmann, P. Varaiya, Managing usage-based pricing in a future telecommunication market, *4th PAAM 1999*, London, UK, April 1999.

[20] N. Blefari-Melazzi, D. Di Sorte, M. Femminella, G. Reali, Theoretical analysis of a virtual delay based tariff model, *IEEE ICC 2003*, Anchorage, USA, May 2003.