

# COVER PAGE

**Title:**

**“A Novel Approach to Charge for IP Services with QoS support”**

**Authors:** Dario Di Sorte, Mauro Femminella\*, Gianluca Reali

Department of Information and Electronic Engineering (DIEI)  
University of Perugia  
Via G. Duranti, 93  
I-06125 Perugia, Italy  
Email: {disorte,femminella,reali}@diei.unipg.it

**\* Corresponding Author:**

Mauro Femminella  
Department of Information and Electronic Engineering (DIEI)  
University of Perugia  
Via G. Duranti, 93  
I-06125 Perugia, Italy  
Ph.: +39 075 585 3630  
Fax: +39 075 585 3654  
e-mail: femminella@diei.unipg.it

## **Abstract**

If the Internet is to become a network supporting differentiated application and transfer services, advanced architectures must be deployed to efficiently support hard Quality of Service (QoS) and usage-based charging. In this paper we present a novel pricing scheme for IP services with guaranteed quality. Our approach is built on the basis of the *virtual delay*, which is a novel, simple and effective QoS index that describes an advanced IP service. We propose a model to compute the virtual delay from a purely technical point of view, taking into account not only guaranteed performance, but also traffic and system parameters. We then analyze the sensitivity of both the virtual delay and the tariff towards the involved parameters, taking into account both the users' benefit and the operators' income. We also extend the pricing model to make it dependent on service demand. Finally, we also present an economic analysis, the aim of which is to establish a model to set the QoS level and the relevant price, taking into account revenue, social fairness, and service availability.

**Keywords:** QoS index, usage-based charging, service demand, adaptive system, price setting.

**Suggested running header:** To charge for IP services with QoS support

## 1. INTRODUCTION

The role of network service charging in IP-based networks is not yet clear. As regards the Internet, the most common charging method is based on the flat-rate model, i.e., subscribers pay a fee to access the network, independent of their effective use of the service. A number of pricing models have been proposed in literature (see [1,2,3,4] for an overview) for both *elastic* applications (typically TCP-based), which tolerate frequent variations of the transmission rate [5], and *inelastic* applications (typically UDP-based), which require a network support with strict guarantees in terms of throughput, delay, and packet losses [5]. Pricing issues have also been widely investigated recently within the framework of international research projects (e.g., INDEX [6], CA\$MAN [7], CATI [8], Internet Next Generation [9], WHYLESS.COM [10], MOBIVAS [11]) and IETF and IRTF working groups (AAAARCH WG and AAA WG). If the Internet is to become a network which supports differentiated application and transfer services, the flat-rate model might be considered inefficient from an economic point of view, since it does not enable charges to be made according to the type and to the quality level of the transfer service [12]. From this point of view, flat-rate pricing may be considered inadequate for networks supporting Quality of Service (QoS), whose management entails expensive investments to support service differentiation. In this respect, it is well known (e.g., see [12]) that the need to implement admission control mechanisms and resource reservation protocols has become urgent. In fact, the accomplishment of these tasks can avoid degradation of network services under network congestion, differentiate network services, and satisfactorily support real time applications (e.g., audio and video services) with hard QoS guarantees. Thus, usage-based tariffs can guarantee network service providers an additional income [12]. This clearly implies a large number of issues in the field of network management. Advanced architectures to efficiently support not only QoS, but also pricing, charging, billing, and accounting must be deployed [12]. More specifically (see Fig. 1), (i) accounting is the act of collecting data concerning resource consumption [1,13,14,15], (ii) pricing is the process which determines the tariff model to be adopted [1,2,3,4], (iii) charging is the function which computes the cost in monetary units from accounting and pricing data, and (iv) billing is the act of preparing and sending invoices to users [11].

In this work, we focus especially on aspects of intra-domain pricing, and in particular on the definition of a tariff model to charge for QoS-enabled network services supporting inelastic applications. The proposed tariff model is based on the concept of virtual delay, which is a novel, simple, flexible and effective QoS index, that describes an

advanced edge-to-edge IP service provided by a single administrative domain. The virtual delay was introduced qualitatively in [16,17], whereas the model to compute it was briefly sketched in [18]. The virtual delay is the input to set the charges in terms of price per unit-of-volume and unit-of-time in the usage-based tariff model presented in [16] to charge for intra-domain added-value IP services <sup>1</sup>.

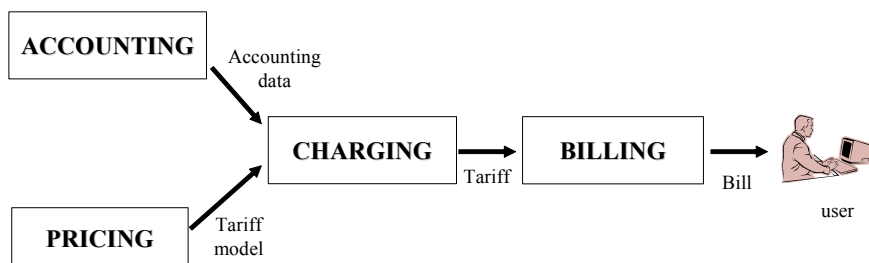


Fig. 1 – Usage-based charging: main functions.

In this paper, our objectives are:

1. to investigate the sensitivity of the overall pricing system to model parameters (traffic descriptors, QoS parameters, system capacity) and to extend and refine the preliminary quantitative results presented in [18], by providing a more detailed, theoretical analysis, which also includes an evaluation of the sensitivity of the model to variation in traffic parameters;
2. to extend the virtual delay computation to make the tariff dependent on the status of resource availability;
3. to present an economic analysis, the aim of which is establish a model to identify the most appropriate QoS level and the relevant price to be charged. Our objective is to maximize the total income, while taking into account some constraints such as social fairness and service availability (i.e., blocking probability due to lack of network resources). We assume that the operator is able to estimate the user service demand. We provide a numerical results in a realistic working environment.

The paper is organized as follows. The motivations of our research activity are illustrated in the following section. In section 3, we recall the main concepts related to the virtual delay based pricing approach, with the aim of giving a complete overview of the overall system. Section 4 presents the model to calculate the virtual delay. The results obtained from a theoretical analysis are described in section 5. In section 6, we extend the tariff model to take into

---

<sup>1</sup> We are aware that a traffic flow may pass through a number of different domains and a model to build the end-to-end tariff is needed. A number of models have been presented in literature: the edge pricing model [19], RNAP protocol [20], and brokering-based procedures [16,17,21]. Also, issues relevant to QoS-and-price based inter-domain routing are very challenging [17,21,22]. In this regard, the virtual delay could also be the basis of minimum price inter-domain routing algorithms [17,21], and it should be a standard (and not proprietary) measure of the QoS level of a service transmission, since the perspective is moved from a local intra-domain to an inter-domain scope. However, these aspects are beyond the scope of this work, in which we propose and analyze a model to compute the virtual delay commencing with per domain technical issues.

account also the current service demand. Section 7 reports a model to set the final price and numerical results relevant to a specific application scenario. In section 8, we provide the reader with considerations on the applicability of the proposed pricing model. The paper ends with some conclusive remarks.

## **2. RELATED WORKS AND RESEARCH MOTIVATIONS**

Some pricing models, such as, e.g., priority pricing [23], Paris metro pricing [24], expected capacity pricing [25], cumulus pricing scheme [26] based on the current use of network resources, have been proposed in literature so far. They have a common characteristic: they do not claim to exploit a network architecture supporting hard QoS guarantees. In fact, the models aim to provide privileged treatment (soft QoS guarantees) to those customers who pay more, but they do not in any way guarantee a pre-defined network performance. We believe that usage-based charging is correct whenever a service differentiation is provided. On the other hand, we also believe that the approaches described above do not completely support the development of a multi-service Internet, and therefore it would not be convenient in the long-term for a network operator to create expensive accounting mechanisms to support them. However, they could represent an intermediate solution towards a platform which fully supports hard QoS. The interested reader should refer to [1,2,3,4] for an overview of these pricing models according to network, economic, market, and social criteria.

In this paper our goal is to investigate intra-domain pricing issues regarding IP guaranteed services. The idea is that network operators provide users with best effort services for a periodical flat-rate fee, and with IP guaranteed services to support special application services at per-time and/or per-volume tariffs on a per-call basis. One could argue that usage-based charges may discourage the use of the Internet. Nevertheless, some studies [6] show that customers are willing to pay an additional per-usage charge in order to improve network performance.

An important need for both customers and operators has emerged for the identification of a clear, objective and convincing pricing criterion to charge for value-added network services supporting real-time applications. Although there are, as yet, no standard criteria, some pricing models, based on those network resources that are actually consumed and reserved, have been proposed [27,28,29]. These approaches propose to charge the effective bandwidth [30,31], which is a measure of the resource allocation within the network to obtain a given performance level. Charging network services on the basis of the effective bandwidth is reasonable, since operators charge users in proportion to the amount of resources consumed by their traffic. In addition, the computation of the effective

bandwidth addresses a number of parameters (traffic parameters, QoS parameters, network resources) from well-established technical models.

The concept of effective bandwidth for pricing purposes was first used in [27], where the Authors present a pricing law per-unit of effective bandwidth, so that the user is charged a fixed price per-packet plus an amount based on the burstiness of the flow. In [28], the Authors assume a linear tariff charge depending on time ( $T$ ) and volume ( $V$ ) equal to  $a_0 T + a_1 V$ , where  $a_0$  and  $a_1$  are functions of traffic contract parameters and the operating point of the network (known a priori), including the user-declared value of the mean rate  $X=V/T$ . The model is such that it charges lower tariffs if users declare an accurate estimate of  $X$  at set-up. This is also advantageous for the network, since users are encouraged to behave fairly, and resources can be efficiently allocated. This model develops the general pricing approach proposed in [29], according to which the user is charged for a previously calculated bandwidth value  $\alpha(X)$ , which is a linear function tangent to the effective bandwidth curve  $f(X)$  typical of the flow. Such a model assumes a knowledge of the characteristics of multiplexed traffic and link resources.

However, in our opinion, the effective bandwidth pricing strategy fails as regards the following points:

1. it is strongly dependent on the allocation policy and management capability of the operator, whereas (i) operators would like to charge not only on the basis of the amount of allocated resources, but also mainly on the basis of the cost to provide customers with QoS guarantees (e.g., buffer space may be cheaper than bandwidth, so that packet loss may be cheaper than delay jitter); (ii) end users would like to be charged with respect to the perceived QoS;
2. the effective bandwidth does not address all QoS parameters (e.g., propagation delay, transmission time, processing time), but only the queuing delay and packet loss probability. Note that in some cases the missing parameters affect the QoS level more than delay jitter and packet loss (e.g., in satellite communications);
3. the effective bandwidth is not flexible in the sense that the relative weights of the two QoS parameters (queuing delay and packet loss) cannot be regulated. A tuning operation is necessary in order to: (i) reflect the actual relative cost of the two parameters (operators' side), as explained in point 1 above; (ii) give an objective QoS index, the computation of which may depend on the type of application to be supported (users' side).

To overcome these drawbacks, we define the concept of *virtual delay*, that is to say a novel QoS index which includes all negotiable QoS parameters (maximum transfer delay, maximum delay jitter, and packet loss probability) in a flexible way. Since we are dealing with flows supporting inelastic applications, it is worth noting that the virtual delay is computed with respect to a specific traffic profile described by dual leaky bucket parameters (peak rate,  $P_S$ ,

sustainable rate,  $r_s$ , and burst tolerance,  $B_{TS}$ ) [30,31]. For this reason we do not explicitly consider the throughput as a QoS parameter.

The choice of defining the virtual delay, which is expressed in the time domain, is due to the following reasons:

- all QoS parameters to be taken into account in our model, with the exception of packet loss, are delays. Thus, in the computation of the virtual delay our effort focused mainly on the definition of a model for mapping the packet loss probability into a delay component. Such a model is based on consistent, technical considerations at IP level. As a result, if all contributions are mapped in the time domain, they can be managed easily and flexibly through a linear combination. In fact, the different contributions can be weighted to provide the network operator with a degree of freedom to take into account its needs, as described in points 1 and 3 above;
- any network service characterized by a given quality level can be mapped into a "virtual delay" value. This quantity, expressed in seconds, is likely to give an easy and pertinent understanding of the quality of the corresponding network service.

### **3. NETWORK COMMODITY AND PRICING LAW**

In this section, we briefly recall the concept of the "commoditization" of the network service and the relevant pricing law used to charge for IP guaranteed services [16,17] provided within a single domain.

#### **3.1. Basic assumptions of the model**

We start from the following assumptions for a single administrative domain:

- flows entering the domain are regulated by dual leaky bucket devices [30,31]. The traffic descriptors are: peak rate,  $P_s$ , sustainable rate,  $r_s$ , and burst tolerance,  $B_{TS}$ , according to the fluid traffic model. Remember that we are dealing with traffic associated with inelastic applications which cannot tolerate variations of the transmission rate/profile. These traffic descriptors have been standardized for IP networks supporting QoS guarantees [32]. A flow is a set of packets traversing an administrative domain, all of which belong to the same application session (also referred to as a *call* in the remainder of the paper) running between two hosts and receive the same QoS treatment [33]. The packetized model used in IETF documents includes a set of parameters named  $T_{spec}$  [32]; beyond  $B_{TS}$ ,  $r_s$ , and  $P_s$ , there are other two parameters:  $m$ , the minimum-policed unit, and  $M$ , the maximum datagram size. The value of these parameters depends on the packetization of the information emitted by the information sources. In this respect, we have not made any hypothesis. Nevertheless, we make use of the traffic

fluid model, since the effect of discretization of information into bits (the information unit) is totally negligible on our pricing model;

- an admission control function is implemented, as suggested in [12]. 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 and resource reservation strategies implemented within its network. We only assume that a single domain is able to control traffic congestion (i.e., the number of admitted calls within the network) so as to guarantee the QoS level of the service, which is described by a set of QoS parameters (see next point) that cannot be violated. Our approach is compliant with both per-flow and per-aggregate traffic management within the core network, and therefore with both the Integrated Services (IntServ) architecture [33] and the Differentiated Services (DiffServ) model [34], respectively. The set of network services offered is strictly dependent on the policies of the domain;
- for the reasons illustrated in the previous section, we assume that the QoS of the port-to-port IP network service provided to the specific flow is described by the following *service parameters*: the maximum transfer delay,  $D_{\max}$ , the maximum delay jitter,  $D_{\text{jitter}}$ , and the loss probability,  $P_{\text{loss}}$ , due to buffer overflow. We assume that these parameters may be negotiated between network service providers and network users. Other QoS parameters, such as channel reliability, resilience and call set-up time (*network parameters*), characterize the intrinsic quality of the network, do not depend on the specific flow, and cannot be negotiated.

### 3.2. Network commoditization

The administrative domain offers a guaranteed service and it is clearly interested in getting revenue from the market. 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 negotiated by users), which can be summarized by the virtual delay,  $d \geq 0$ . This quantity is a comprehensive and all-inclusive appraisal of the transfer delay, delay jitter and loss probability, and characterizes an edge-to-edge service offered by a network service provider. A network service is modeled by the virtual delay, which gives a measure of the perceived QoS level (the computation of the virtual delay is reported in section 4): the higher the level of 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 (i.e., a bit) with a measure expressed in commodity units. We define the price of the transfer of an information unit as the quantity  $P(d) = \gamma \alpha f(d) = \beta f(d)$ , where  $\alpha$  is the price per commodity unit

and  $\gamma$  is a price variation factor that accounts for market fluctuations (in section 6 we make it depend on service demand). Thus,  $\beta = \alpha\gamma$  is the market commodity price (i.e., the price per commodity unit).

The choice of  $f(d)$  is really strategic, since its dependence on  $d$  indicates how, and how much, the operator wants to differentiate the tariff according to the QoS level. To do this, the network operator should consider both the benefits perceived by users and their willingness to pay.

In this respect, an important concept, widely used in literature, is the *utility function*. We assume that the utility function gives a measure of the user's preferences in terms of sensitivity to the perceived QoS level. This understanding of utility function is quite common in the literature (e.g., see [3] and references therein). It is strictly related to the users' willingness to pay (*demand curve*) [3],  $\bar{U}(d)$ , expressed in terms of currency units per time-unit or per volume-unit.  $\bar{U}(d)$  may be assumed as an average value known from a specific market analysis which is beyond the scope of this work. In our proposal, the function  $f(d)$  ranges between 0 and 1 and is proportional to  $\bar{U}(d)$ .

In other words, given two QoS levels,  $d_1$  and  $d_2$ ,  $\bar{U}(d_1)/\bar{U}(d_2) = f(d_1)/f(d_2)$ . Thus, the basic assumption of our scheme is that, in principle, the higher the benefit a user perceives, the higher the amount of money he/she is willing to pay, and the higher the price charged by the operator.

Thus, we assume that  $f(d)$  is the utility function associated with the application service (e.g., voice over IP); the application service is associated with a specific  $d$  value, which means that it must be supported by a specific network delivery service to be charged. Thus, once the  $d$  value is established, the relevant value of  $f(d)$  is related to the willingness of users to pay for receiving that network service, that is the QoS level represented by the virtual delay  $d$ . Such values are then converted into prices by the factor  $\beta$  ( $P(d) = \beta f(d)$ ). It is worth noting that the choice of the market commodity price  $\beta$  has to take into account the dynamics of the market (see the relevant analysis in section 7).

Since we are dealing with inelastic applications, utility rapidly decreases when the amount of network resources allocated (and thus the QoS level) falls below a given threshold. For this reason, the utility function associated with inelastic applications is commonly assumed to be a sigmoidal-like function [5,35,36]. An increasing function  $g(x)$  is sigmoidal-like if it has one inflection point  $x_0$ , and  $\partial^2 g(x)/\partial x^2 > 0$  for  $x < x_0$  and  $\partial^2 g(x)/\partial x^2 < 0$  for  $x > x_0$ . A frequently used sigmoidal-like function is  $g(x) = 1/(1 + e^{-b(x-a)})$  ( $b > 0$ ).

Since the virtual delay value decreases when the QoS level improves, we assume  $f(d) = 1/(1 + e^{b(d-a)})$  (with  $a, b \geq 0$ ), which is a sigmoidal-like function reflected in the line  $d=a$ , with values in the range from 0 to 1 (Fig. 2).

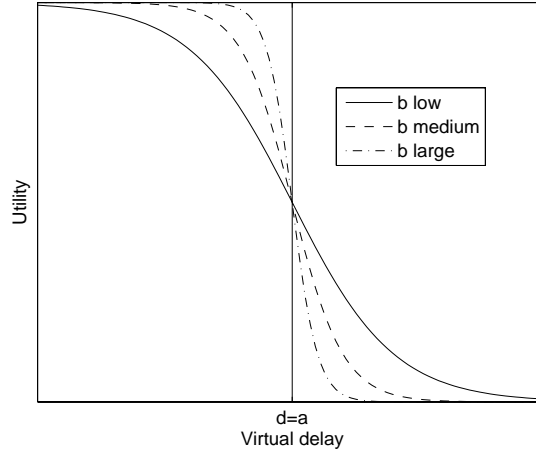


Fig. 2 – An example of function  $f(d)$ .

### 3.3. Tariff model

Below, we assume that both the value of  $\beta$  and the QoS level are constant throughout the duration of a call.

Let  $T$  be the duration of a session and  $t_0$  its starting time. The tariff applied to charge for the service offered to a flow entering the domain with an instantaneous bandwidth equal to  $B_{ist}(t)$  is [16] (see Fig. 3(a))

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

where  $B_{res}$  is the bandwidth value that a domain charges on a per-time basis; thus  $B_{res} T$  is the traffic volume charged independently of the resources actually used by the flow. Note that extra-usage of bandwidth (beyond  $B_{res}$ ) is charged on a per-volume basis; that is to say, when  $B_{ist}(t) > B_{res}$ , the tariff charges according to the resources actually used. Thus, as highlighted in Fig. 3(a), the tariff consists of a component depending on the duration time of the call (*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, from this point of view, may be regarded as a tunable knob. In principle, the value of  $B_{res}$  may range from zero to the peak rate of flow. Note that if the value of  $B_{res}$  increases, the price charged to users also increases, unless  $B_{ist}(t) \geq B_{res}, \forall t$  (in this case the tariff would remain unchanged).

Each network operator is obviously free to choose the value of  $B_{res}$ , according to its own pricing policy from an economic and market analysis. From the technical point of view, the quantity  $B_{res}$  may be accurately selected as being

equal to the effective bandwidth associated with the flow (e.g., see [30,31]), which ranges within  $[r_S, P_S]$ . In packet networks, the effective bandwidth is the value of the bandwidth typically used for admission control purposes; it is dependent on traffic characteristics, amount of network resources, and requested performance. In our opinion, the choice to set  $B_{res}$  as equal to the effective bandwidth is reasonable, since, in this way, operators charge users in proportion to the reserved resources, and protect themselves against unfair behavior by users, who are encouraged to correctly describe the service they really need. The concept of effective bandwidth will be further discussed for the computation of the virtual delay in the following section.

From Eq. (1), it is possible to express the tariff  $Q$  as

$$Q = a_V(d)\tilde{V} + a_T(d)T, \quad (2)$$

where  $\tilde{V}$  is the amount of traffic volume charged on a per volume-unit basis, and  $a_V(d)$  and  $a_T(d)$  are the per volume-unit charge and the per time-unit charge, respectively. It is clear that the tariff strongly depends on the instantaneous bandwidth 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, we can identify two extreme cases compliant with the dual leaky bucket operation: constant rate and “extremal” ON/OFF [37]. Correspondingly, it is possible to identify the extreme tariffs (note that  $a_T = a_V B_{res}$ ):

- the *minimum* tariff (associated with the constant rate case) is

$$Q_{\min} = a_V B_{res} T + a_V \max(0, r_S - B_{res}) T = \begin{cases} a_V r_S T & \text{if } B_{res} < r_S; \\ a_V B_{res} T & \text{if } B_{res} \geq r_S \end{cases} \quad (3)$$

- the *maximum* tariff (associated with the extremal ON/OFF case) is

$$Q_{\max} = a_V B_{res} T + a_V (P_S - B_{res}) T_{on} \frac{T}{T_p} = a_V \frac{B_{res} P_S + P_S r_S - B_{res} r_S}{P_S} T. \quad (4)$$

The latter is relevant to an extremal ON/OFF flow (see Fig. 3(b)), so that the amount of volume  $\tilde{V}$  is maximized.  $T_{on}$  is the duration of the ON state in which the source transmits at the peak rate and it is equal to  $T_{on} = B_{TS} / (P_S - r_S)$ ,  $T_{off}$  is the duration of the OFF state and it is equal to  $T_{off} = B_{TS} / r_S$ , and  $T_p$  is the duration of the period, equal to the sum of  $T_{on}$  and  $T_{off}$ . It is correct to say that the highest tariff corresponds to the maximum burstiness of the transmission rate; in fact, bursty flows (in particular “extremal” ON/OFF flows) stress network resources more than flows with a smoothed rate [30,31].

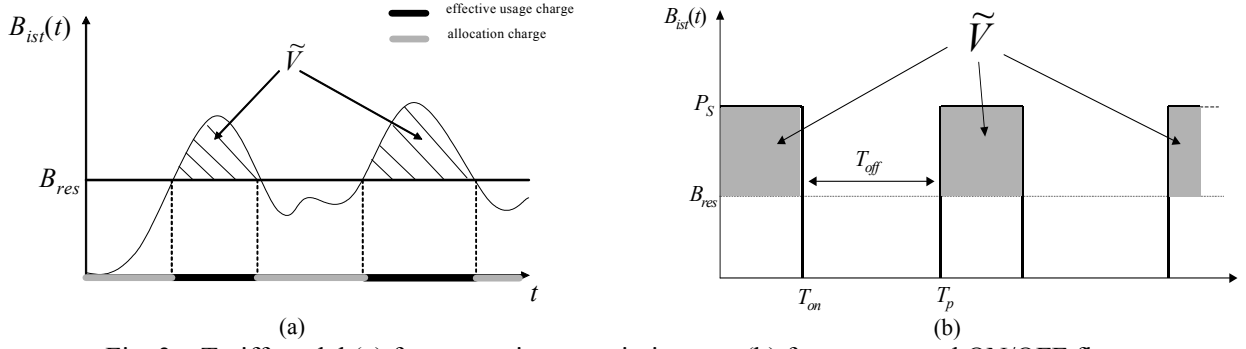


Fig. 3 – Tariff model (a) for a generic transmission rate (b) for an extremal ON/OFF flow.

Finally, under the realistic assumption that  $B_{res} \geq r_S$ , we can rewrite Eqs. (3) and (4) as follows:

$$Q_{\max} = T[\beta(B_{res}(1 - r_S/P_S) + r_S)]f(d); \quad (5)$$

$$Q_{\min} = T[\beta B_{res}]f(d). \quad (6)$$

The quantities in square brackets in Eqs. (5) and (6) represent the price per-commodity unit per-time unit.

#### 4. COMPUTATION OF THE VIRTUAL DELAY

The virtual delay must take into account in a simple, flexible way not only all the QoS parameters, but also the cost of deploying various QoS features and the actually perceived QoS level. Below, we show how to compute the value of the virtual delay that characterizes an edge-to-edge service, from technical considerations at the IP layer.

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,  $P_{loss}$ . We recall that the selection of this set of parameters is due to our view of the network service within the specific framework described in section 2. This does not prevent the introduction of other parameters if they are deemed necessary and are 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 queuing in nodes mainly causes delay jitter, 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, loss probability may be traded for queuing delay, and therefore it may be represented as a contribution to the virtual delay evaluation.

Now our goal is to find the component of the virtual delay related to the loss probability,  $D_P$ . Once this law has been

found, the value of virtual delay is given by

$$d(D_{\max}, P_{\text{loss}}) = D_{\max} + D_P(P_{\text{loss}}). \quad (7)$$

With reference to Fig. 4, we associate a port-to-port service described by  $(D_{\max}, D_{\text{jitter}}, P_{\text{loss}})$  with an equivalent node that introduces an (almost) constant delay  $D_C = D_{\max} - D_{\text{jitter}}$ , (which is a comprehensive, all-inclusive appraisal of all the constant components of delay along the network, e.g., processing time, transmission time, propagation delay), a maximum queuing delay equal to  $D_{\text{jitter}}$ , and a loss probability equal to  $P_{\text{loss}}$ . We call  $C$  the amount of capacity from the input port to the output port dedicated to the specific service, and consequently  $B = CD_{\text{jitter}}$  is the buffer space of the equivalent node.

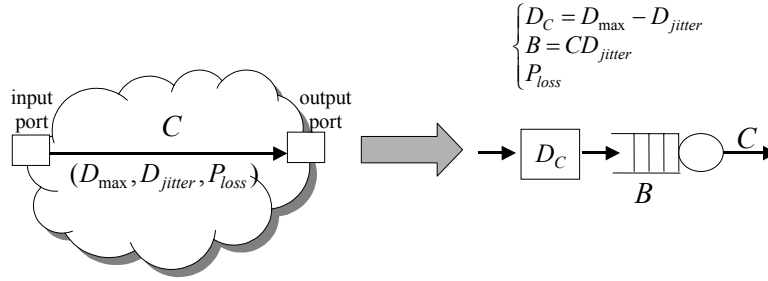


Fig. 4 - Definition of the equivalent node to represent a port-to-port network service.

The effective bandwidth associated with a given flow is a measure of its resource occupation within a network element to obtain a given performance level ([30,31]). Below, we evaluate the effective bandwidth on the equivalent node, which is an abstraction of the network domain. The final goal is exclusively to define a model to map a loss probability to the time domain, and not to evaluate the effective bandwidth used by the operator for admission control within the network.

In [30], the Authors have shown that the maximum buffer occupancy of a single flow, characterized by a set of dual leaky bucket traffic descriptors  $(P_{S0}, r_{S0}, B_{TS0})$ , which feeds a buffer of unlimited size, served at a transmission capacity  $c (\leq P_{S0})$ , is given by:

$$b = B_{TS0}(P_{S0} - c) / (P_{S0} - r_{S0}). \quad (8)$$

For given source parameters, Eq. (8) specifies the equation of a straight line, which is the lower envelope of the set of  $(b, c)$  pairs which guarantee zero loss.

In addition, if we set the maximum queuing delay of an information unit as  $D_{\text{jitter}} = B / C$ , then the following additional equation applies:

$$b/c = B/C. \quad (9)$$

This equation provides another infinite set of pairs  $(b,c)$ . The intersection of Eqs. (8) and (9) determines 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

$$c_0 = P_{S0} B_{TS0} / (D_{jitter} (P_{S0} - r_{S0}) + B_{TS0}), \quad (10)$$

$$b_0 = c_0 D_{jitter}. \quad (11)$$

Then, let  $c_{p0}$  be the value of effective bandwidth and let  $b_{p0}$  be the effective buffer corresponding to a service characterized by a value of queuing delay equal to  $D_{jitter}$  and by a value of loss probability equal to  $P_{loss}$ . We calculate  $(c_{p0}, b_{p0})$  by using the approach illustrated in [31] (see also the Appendix).

Our goal is to associate the network service with losses with a virtual service without losses, by (virtually) increasing the effective buffer space allocated to the flow. From Eq. (8), once the amount of bandwidth  $c_{p0}(P_{loss})$  assigned to the flow has been determined, the amount of buffer needed to avoid losses is equal to

$$\overline{b}_{p0}(P_{loss}) = B_{TS0} (P_{S0} - c_{p0}(P_{loss})) / (P_{S0} - r_{S0}). \quad (12)$$

This means that the buffer allocated to the flow should be increased by a value equal to  $\overline{b}_{p0} - b_{p0}$ . This would imply an additional virtual queuing delay associated with the loss probability equal to

$$D_P(P_{loss}) = (\overline{b}_{p0} - b_{p0}) / c_{p0}. \quad (13)$$

This leads to the final equation of the virtual delay:

$$d(D_{max}, P_{loss}) = D_C + D_{jitter} + D_P(P_{loss}) = D_C + \left( \frac{P_{S0} - c_{p0}(P_{loss})}{P_{S0} - r_{S0}} \right) \frac{B_{TS0}}{c_{p0}(P_{loss})}. \quad (14)$$

Therefore, this parameter can be considered as an index describing the QoS level of the service.

Note that it is possible to assign different weights to the different contributions  $(D_C, D_{jitter}, D_P)$ , according to the pricing policy of the domain, the type of application service to be supported by the network and the cost for the operator according to the various delay contributions. For simplicity's sake and so as not to make application any less general, in the following we assume that all weights are set at 1, as highlighted in Eq. (14).

## 5. ANALYSIS OF THE VIRTUAL DELAY BASED TARIFF MODEL

In this section, our goal is to provide numerical results from a theoretical analysis of the pricing approach which

focuses on the sensitivity of the model to system parameters. In our analysis, we make the following assumptions:

- the quantity  $B_{res}$  in the tariff expression is assumed to be equal to the effective bandwidth,  $c_p$ , associated with the flow described by the traffic parameters  $(P_S, r_S, B_{TS})$  and computed for pricing purposes on the equivalent node. As mentioned above, the choice to set  $B_{res}$  as equal to the effective bandwidth is reasonable, since, in this way, operators charge users in proportion to the reserved resources, and protect themselves against unfair behavior by users, who are encouraged to correctly describe the service they really need;
- the function  $f(d)$ , which maps virtual delay values in a number of commodity units, is assumed to be equal to  $f(d) = 1/(1 + e^{b(d-a)})$  with  $a, b \geq 0$ . The rationale of this choice is explained in section 3;
- as illustrated in detail in the previous section, the virtual delay, computed from Eq. (14), identifies the QoS level of the network service starting from (i) a set of nominal traffic descriptors  $(P_{S0}, r_{S0}, B_{TS0})$  representative of a given service class to support a specific application (e.g., voice over IP); (ii) a set of QoS parameters  $(D_{max}, D_{jitter}, P_{loss})$  known a priori, i.e., negotiated by the user and the operator at the connection set-up. In other words, it is not necessary to measure the perceived QoS on a per-call basis. It is worth noting that it is the task of the admission control procedure on the operator's side not to violate the negotiated level of QoS (i.e., not to admit too high a number of calls within the network). This clearly means that the virtual delay (and thus the tariff) associated with a call is computed before the beginning of data transmission.

From Eqs. (5) and (6), the minimum and maximum tariffs charged for a  $T$ -seconds call are

$$Q_{max} = T[\beta(c_p(1 - r_S/P_S) + r_S)]f(d), \quad (15)$$

$$Q_{min} = T[\beta c_p]f(d). \quad (16)$$

The value of parameter  $b$  of function  $f(d)$  must be set in order to tune the influence of the QoS level (i.e., virtual delay) on the tariff charged to network users. In other words,  $b$  has to be increased or decreased to either sharpen or smooth the price difference between different network services with values of  $d$  at approximately the value of  $a$  (i.e., the abscissa of the inflection point of  $f(d)$ ). Note that the dependence of the tariff on the QoS level is also taken into account by  $c_p$ ; a better service implies a higher value of  $c_p$  and therefore a higher tariff.

To sum up, the tariff may be considered to depend on two factors: the first can be tuned to the virtual delay, whereas the second depends on the effective bandwidth. This implies that the former is strictly related to the performance level and the latter takes into account the amount of network resources allocated. Consequently, two different flows

requesting the same amount of resources to be allocated (i.e.,  $c_p$ ), but with a different QoS level (i.e., different  $d$  values), may be charged different costs, and vice versa.

Now, let us proceed with the analysis of the tariff as regards the system parameters. To do so, we consider a voice over IP service. We assume the following nominal dual leaky bucket parameters, compliant with the G.726 codec with rate equal to 32 Kbps and silence suppression [38]:  $P_{s0}=32$  Kbps,  $r_{s0}=13.6$  Kbps, and  $B_{TS0}=5300$  bytes.

The target performance level of the guaranteed transfer service is specified as follows:  $D_{\max}=150\div 200$  ms, ( $D_C=140$  ms and  $D_{\text{jitter}}=10\div 60$  ms), and  $P_{\text{loss}}\leq 10^{-3}$ . These parameters describe the performance level to satisfactorily support a voice over IP service (see [38] and references therein).

The minimum system edge-to-edge capacity is set at 2.048 Mbps. In addition, the value of  $a$  and  $b$  in  $f(d)$  has been set at 1.85 and 10, respectively. These values are inputs to our pricing model and come from a market analysis, as discussed in section 3.2. Also, the price per-commodity unit is  $\beta=\alpha=2.8\cdot 10^{-5}$  \$ (here we assume  $\gamma$  constant and equal to 1), where \$ is a generic currency unit. Please note that at this stage the value of market commodity price is arbitrarily set (we only want to obtain realistic values of the tariff charged versus different quality levels of the service), since our first goal is to analyze the tariff model from a strictly technical viewpoint. The analysis to properly set the market commodity price is presented in section 7.

Fig. 5 illustrates a possible service classification based on the perceived service identified by  $d$  (on the abscissa). On the basis of the QoS parameters describing the network service to support a voice over IP service, we have divided services into three classes: premium service ( $P_{\text{loss}}=10^{-5}\div 10^{-4}$ ,  $D_{\max}=0.15$  s), good service ( $P_{\text{loss}}=10^{-4}$ ,  $D_{\max}=0.15\div 0.2$  s), basic service ( $P_{\text{loss}}=10^{-4}\div 10^{-3}$ ,  $D_{\max}=0.2$  s). The ordinate of Fig. 5(a) shows the function  $f(d)$  and the ordinate of Fig. 5(b) represents the per-second tariffs, in both the best and the worst case. As expected, a better service in terms of QoS is charged at a higher price.

We remark that the validity of the analysis illustrated below is quite general. The peculiarities of the pricing system remain unchanged for other values of traffic parameters, QoS parameters, and utility function (the estimation of which is beyond the scope of this paper).

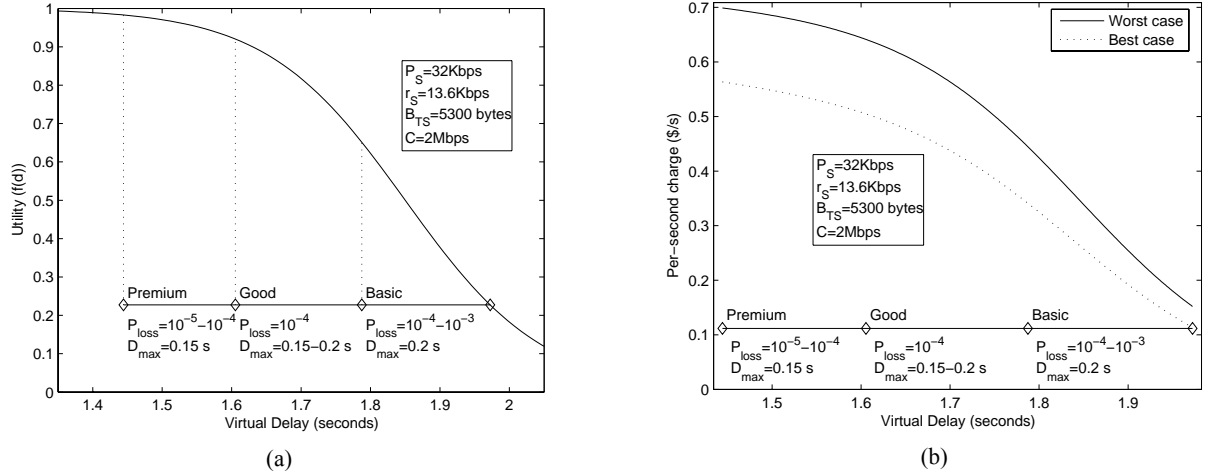


Fig. 5 – (a) Function  $f(d)$  and (b) Tariffs versus virtual delay.

### 5.1. Performance level variations

In Fig. 6(a), we show the virtual delay as a function of  $P_{loss}$ , with the maximum delay,  $D_{max}$ , set as a parameter. As expected, a better service in terms of lower values of packet loss probability is mapped into a smaller value of  $d$ . Since  $f(d)$  decreases with  $d$ , and the effective bandwidth clearly increases when the packet loss probability decreases, the result is that improved performance (i.e., low values of loss probability) has a higher price. In Fig. 6(b), the per-time charge in both the best and the worst case is plotted as a function of the packet loss probability with  $D_{max}$  set as a parameter. The tariff corresponding to the extremal ON/OFF profile (worst case) is higher than the one corresponding to the transmission rate remaining below the effective bandwidth (best case).

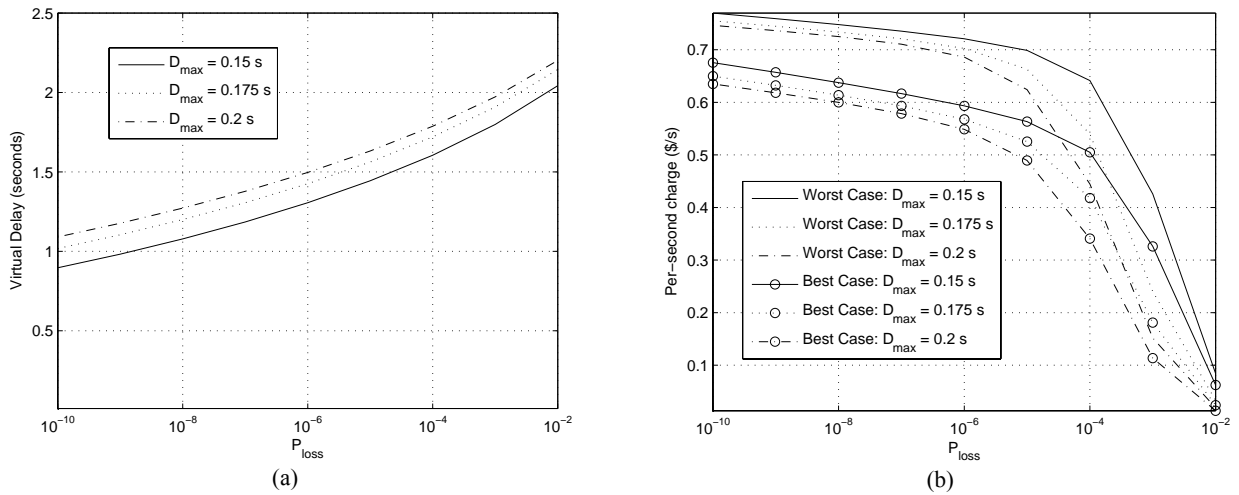


Fig. 6: (a) Virtual delay as a function of loss probability; (b) Per-second charge as a function of loss probability.

### 5.2. System resource variations

When the transmission capacity  $C$  between two edge nodes varies, we have to analyze the effects of system scaling on applied tariffs. In Fig. 7(a), from Eq. (14), we obtain the behavior of the virtual delay as a function of  $C$  ranging from

256 to 2048 Kbps, with maximum delay jitter fixed at 10, 35 and 60 ms, and packet loss set to  $10^{-3}$ . It is evident that the virtual delay is a function which increases with the value of the system capacity.

In Fig. 7(b), while maintaining the same settings, we obtain the equivalent per-second charge (\$/s) as a function of the capacity,  $C$ , in both the best and the worst case tariffs/emission patterns. The result is compliant with the previous figure, i.e., the tariffs decrease with the value of  $C$ .

To explain these results, let us consider that increasing the system capacity means increasing the statistical multiplexing gain. In other words, the effective bandwidth associated with a flow decreases with the system scale (e.g., see [30,39]). The effect of this phenomenon is evident in both the virtual delay, which increases when the effective bandwidth decreases (see Eq. (14)), and in the per-commodity unit per-time unit expressions (see Eqs. (15) and (16)), which decrease with  $C$ . Therefore, an operator which owns a high availability of transmission resources can lower the tariff. This is consistent with the concept of "economies of scale" (from an economic point of view), and with the fact that the statistical multiplexing gain increases with the system size (from a technical point of view) [39].

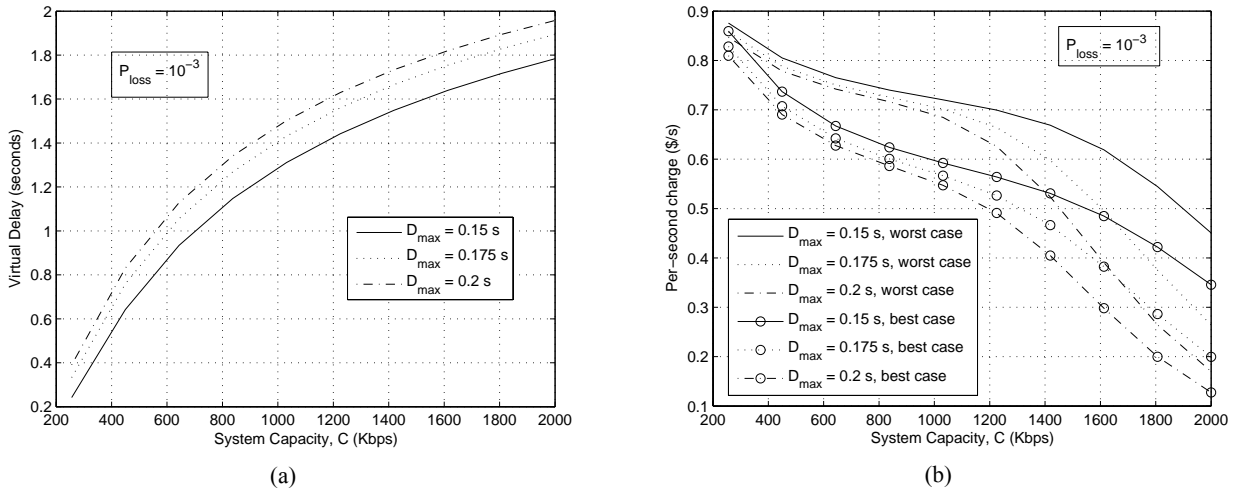


Fig. 7: (a) Virtual delay as a function of the system capacity; (b) Per-second charges as functions of the system capacity.

### 5.3. Traffic descriptor variations

It is worth remarking that the network operator might allow users to negotiate traffic descriptors at the beginning of the communication session. In this sub-section, we analyze the sensitivity of the tariff to variations in traffic descriptors ( $P_S, r_S, B_{TS}$ ) relative to the nominal values ( $P_{S0}, r_{S0}, B_{TS0}$ ). We stress that the nominal values are always used to evaluate the virtual delay value as a QoS index associated with a given traffic class. Thus, the sensitivity of the tariff is simply due to variations of the effective bandwidth in the expression of the price per-commodity unit per-

time unit ( $\beta f(d)B_{res} = \beta f(d)c_p$ ). Thus, the higher the values of  $r_s$ ,  $P_s$ , and  $B_{TS}$ , the higher the value of  $c_p$  [30,31], and therefore the higher the tariff (see Eq. (1)). Consequently, users are strongly encouraged to behave fairly (i.e., to give a correct description of their traffic), and resources can be efficiently allocated within the network.

Fig. 8 shows the per-second minimum and maximum tariffs (relevant to both the best and the worst case tariffs/emission patterns, respectively) as a function of the peak rate, sustainable rate, and burst tolerance, with maximum delay jitter set at 10, 35 and 60 ms, and the same packet loss equal to  $10^{-3}$ . The dependence of the tariff on the peak rate and the sustainable rate is heavy, whereas the burst tolerance slightly affects the price charged due to the effect of statistical multiplexing [30,31].

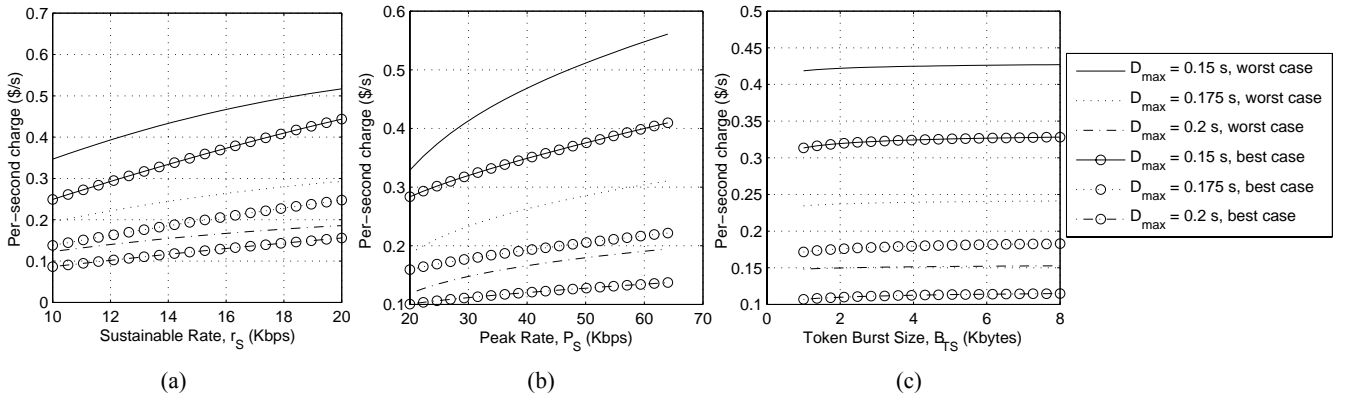


Fig. 8 - Per-second charge as a function of (a) the peak rate  $P_s$ , (b) the sustainable rate  $r_s$ , (c) the burst tolerance  $B_{TS}$ .

## 6. EXTENSION OF THE TARIFF MODEL TO ACCOUNT FOR SERVICE DEMAND

We are dealing with the subset of network services designed to offer hard QoS guarantees to satisfactorily support inelastic applications, which cannot adapt the transmission rate to either varying network conditions or users' willingness to pay. Then, from the tariff model described in the previous sections, we assume that the network administrator charges each call 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 also assume that the price charged is constant throughout the duration of the call and is established at call set-up. In our opinion, this model simplifies the tasks of 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 they will be charged with. We thus limit the price and QoS negotiation phase to the beginning of the communication session only. The assumption of constant price during call lifetime is due to its intrinsic transparency and simplicity.

We also want to make the price increase according to the amount of traffic present in the network, i.e., according to the current amount of service demand. Please note that this approach differs from "congestion pricing"<sup>2</sup>, since congestion in our model is avoided a priori by an admission control scheme.

The next step in the development of the model is the introduction in the tariff of a dynamic factor (i.e.,  $\gamma$ ), which takes into account current service demand.

For this purpose, let us consider the model used to compute the value of virtual delay. As shown in sub-section 5.2, the virtual delay increases with the amount of total capacity  $C$  of the equivalent node through the effective bandwidth (which decreases with  $C$ ). Consequently, if we compute the value of virtual delay each time using the amount of capacity currently available,  $C_{ava}(t)$ , instead of the total capacity  $C$ , the tariff charged would then increase together with the bandwidth demand. From now on we will refer to the effective bandwidth computed from  $C$  as  $c_{p0}$ , and to the effective bandwidth computed from  $C_{ava}(t)$  as  $c_{p0}^*$ . Therefore, the new equation to compute the value of the modified virtual delay,  $d^*$ , is

$$d^*(D_{max}, P_{loss}) = D_C + D_{jitter} + D_P = D_C + B_{TS0} (P_{S0} - c_{p0}^*) / ((P_{S0} - r_{S0}) c_{p0}^*) \quad (17)$$

We make the market factor  $\gamma$  (see section 3.2) depend on  $d^*$ , by means of an exponential negative function, that is  $\gamma(d^*) = e^{-md^*}$ , and consequently  $\beta = \alpha e^{-md^*}$ . Thus, the market price of a commodity unit depends on the amount of service demand. Recalling that the number of commodity units associated with the service unit (i.e., the transmission of an information unit) is given by  $f(d)$ , then, the market price of the service unit is equal to  $\alpha \gamma(d^*) f(d)$  and depends on the QoS level through the function  $f(d)$  and on the amount of service demand through the function  $\gamma(d^*)$ . We remark that our choice of the exponential function for  $\gamma(d^*)$  makes sense since:

- it is a monotonic, positive, non-increasing function with  $d$ , so that the lower the values of the virtual delay (i.e., the higher service demand), the higher the tariff;
- it is easily configurable, since it depends on a single parameter ( $m$ );
- its codomain is finite, so that any value of virtual delay is associated with a finite value (i.e., a finite price);

---

<sup>2</sup> Congestion pricing is commonly adopted in literature to charge for elastic applications (e.g., see [40,41]), and it is a very effective tool to control network congestion [42]. In response to price variations, users adapt their sending rates with the aim of maximizing their utility, so that a efficient, fair sharing of network resources is reached. In addition, congestion pricing can be applied on both a per-packet and per-call basis, and therefore the tariff applied to deliver a packet or to support a call may be strongly dynamic, even throughout the call itself. Thus, users have to be continuously informed about the price charged by the network according to the degree of congestion encountered by the packets along the path, in order to be able to react to price changes and to possibly renegotiate the QoS level and the tariff. A simpler application of congestion pricing is the well-known *time of day pricing*. Studies of its effectiveness can be found in [43,44].

- it has the remarkable property that  $\partial\gamma(d^*)/\partial d^* = -m\gamma(d^*)$ . This means that small variations of  $d^*$  from a value  $d_x^*$  imply price variations which are proportional (through  $m$ ) to the value  $\gamma(d_x^*)$ . In other words, the higher the service demand (i.e., the lower the available resources), the higher the incremental price to add to the tariff to charge for an incoming call. In other words, service demand pricing becomes more aggressive when the amount of available resources decreases.

Below, we refer to the pricing scheme which accounts for service demand as “dynamic pricing”.

To obtain quantitative results, we consider an administrative domain which offers an edge-to-edge transfer service (voice over IP) characterized by the following QoS service parameters:  $D_{\max}=175$  ms, ( $D_C=140$  ms and  $D_{\text{jitter}}=35$  ms), and  $P_{\text{loss}}=10^{-3}$ . Flows are assumed to be homogeneous and described by the same traffic descriptors defined in the previous section ( $P_{S0}=32$  Kbps,  $r_{S0}=13.6$  Kbps, and  $B_{rS0}=5300$  bytes). The transfer capacity associated with the port-to-port service is equal to 2.048 Mbps. Note that the value of virtual delay associated with this setting is equal to  $d=1.91$  s. Thus, assuming the same price parameters adopted in the previous theoretical analysis (i.e.,  $a=1.85$  and  $b=10$  in the function  $f(d)$ , and  $\alpha=2.8 \cdot 10^{-5}$  \$), and introducing the price variation factor,  $\gamma(d^*)$ , then the market price of the service unit is now equal to  $10^{-5} e^{-md^*}$  \$, with  $m$  variable and  $d^*$  computed from Eq. (17). In addition, 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 the following, we carry out the analysis relevant only to  $Q_{\min}$ , the expression of which is shown in Eq. (16), corresponding to a per-second charge equal to  $10^{-5} e^{-md^*} B_{res}$  \$/s. It is straightforward to extend the work to the maximum tariff  $Q_{\max}$ .

In general, the number of flows supported by the port-to-port service depends on the (proprietary) admission control policies deployed by the network operator within its domain. Without loss of generality, we have chosen  $N_{\max}$  as equal to  $\lfloor C/B_{res} \rfloor=113$ , so that the maximum value of utilization factor is  $\rho_{\max} = N_{\max} r_{S0} / C = 0.75039$ .

The virtual delay  $d(t)$  associated with the service requested by a flow 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  $t$ .

Finally, 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  $N(t) \in \{0,1,2,\dots,N_{\max}\}$  of currently active flows. This is due to the fact that  $d^*(t)$  depends on  $C_{ava}(t)$ . Fig. 9(a) shows  $d^*$  as a function of the number of active flows. As expected, the virtual

delay decreases with the number of active flows up to the value of active flows 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, the statistical multiplexing gain then also decreases. In other words, the value of  $c_{p0}^*$  increases up to the maximum value  $c_0$ , which corresponds to the effective bandwidth in the case with no losses, given by Eq. (10). Since the virtual delay depends on the amount of network resources through  $c_{p0}^*$ , then, once the maximum value  $c_0$  is reached,  $d^*$  reaches its minimum value and remains constant, with  $N$  higher than the threshold value  $N_t$ . The constant value of the modified virtual delay is  $d_{\min}^* = D_C + D_{jitter}$  and, consequently,  $D_p = 0$ . Note also that the value of  $d^*(t)$  for  $N(t) = N_{\max}$  cannot 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 increases with  $N$  up to  $N_t$ , and then it remains constant, as shown in Fig. 9(b), where  $m$  is a variable parameter. The higher the parameter  $m$ , the lower the per-second tariff charged. Furthermore, the dependence of the tariff on the virtual delay increases with the parameter  $m$ . The problem of setting the price parameter  $m$  will be further discussed in section 7.

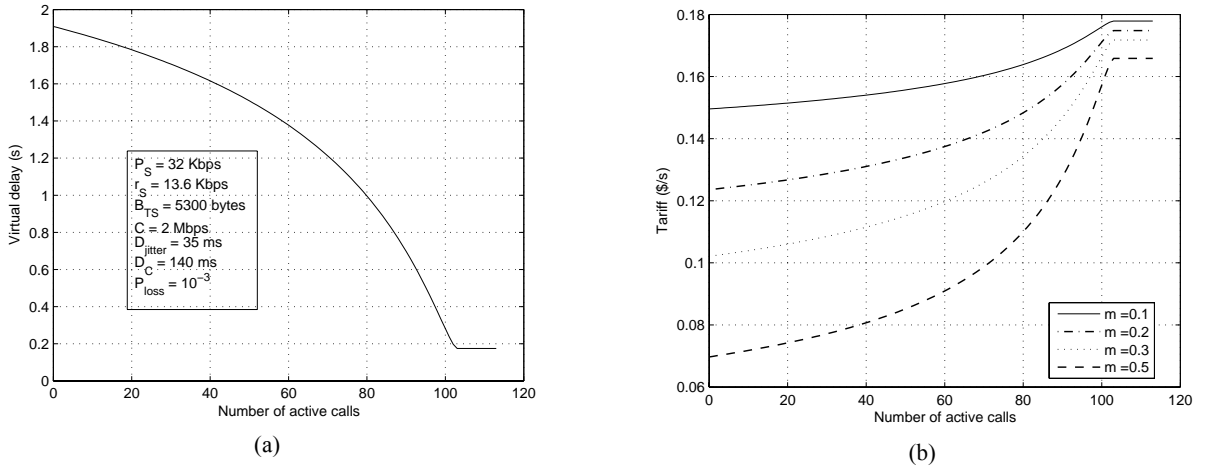


Fig. 9: (a) Virtual delay  $d^*$  and (b) Per-second tariff as a function of the number of active flows.

## 7. PRICE SETTING IN AN APPLICATION SCENARIO

So far, we have analyzed the tariff model from a strictly technical point of view. In this section, our goal is to give some hints for pricing purposes, and to indicate practical implications and tradeoffs for network operators. In more detail, we present an approach to define a model to identify the most appropriate QoS level (i.e.,  $d$  value) and the relevant price to be charged. Our objective is to maximize the total income, while taking into account some constraints, such as the social fairness and the service availability. The basic assumption is that the user service

demand is a-priori estimated by the network operator. We then investigate the effectiveness of the proposed pricing approach and provide a numerical analysis in a specific application scenario.

### 7.1. Performance metrics

As regards user service demand relevant to a specific service with a given QoS level, we make the following assumptions. Calls are generated according to a Poisson arrival process with parameter  $\lambda$ , and their duration is modeled as an exponentially distributed, random variable with mean value equal to  $1/\mu$ . As regards users' willingness to pay  $U$  (expressed in currency units per-second, \$/s), we model this as a random variable with a probability density function  $P_U(u)$ . Also, we are assuming that the operator charge a price depending on the current network status (i.e., the number of active calls,  $N(t)$ ).

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

Fig. 10.

Under the above assumptions, our system can then be modeled as an  $M/M/N_{\max}$  queue with discouraged arrivals, and therefore as a Markov chain characterized by a number of states equal to  $N_{\max}+1$ , Poisson arrivals, and exponentially distributed holding times. In turn, each state  $i$  is characterized by an arrival rate equal to

$$\lambda_i = \lambda \cdot g_i = \lambda \int_{Q_i}^{+\infty} P_U(u) du, \quad i=0,1,2,\dots,N_{\max} \quad (18)$$

and by a call departure rate equal to

$$\mu_i = i\mu, \quad i=1,2,\dots,N_{\max}. \quad (19)$$

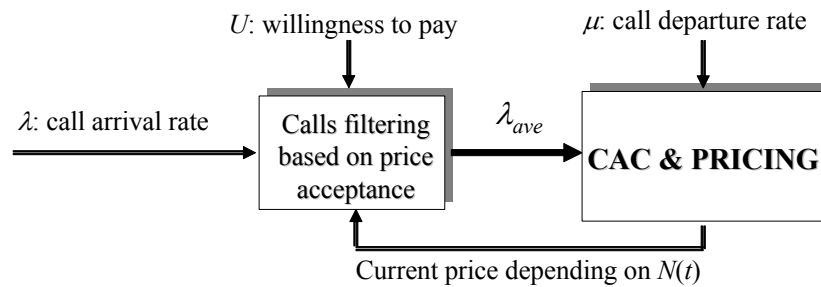


Fig. 10 - System model.

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 state  $i$  (i.e., when  $i$  flows are active). Such a model is similar to the discouraged arrivals queue system described in [45].

As is well known, the probability,  $P_i$ , of having a number  $i$  of active flows in the system is [45]:

$$\begin{cases} P_i = P_0 \prod_{j=0}^{i-1} \left( \frac{\lambda_j}{\mu_{j+1}} \right), & i = 1, \dots, N_{\max} \\ P_0 = 1 / \left( 1 + \sum_{k=1}^{\infty} \prod_{j=0}^{k-1} \frac{\lambda_j}{\mu_{j+1}} \right) \end{cases} \quad (20)$$

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 calls that overcome only price acceptance control). The average number,  $N_{ave}$ , of active flows is equal to

$$N_{ave} = \sum_{i=0}^{N_{\max}} i P_i = (\lambda_{ave} - \lambda_{N_{\max}} P_{N_{\max}}) \frac{1}{\mu} = \lambda_{ave}^* \frac{1}{\mu},$$

where  $\lambda_{ave}^*$  is the average rate of call arrivals that actually enter the

system (i.e., those that overcome both price acceptance control and subsequent admission control). Thus, the average utilization factor of capacity is given by:

$$\rho_{ave} = N_{ave} r_{S0} / C. \quad (21)$$

The call blocking probability due to the lack of resources,  $P_{block}^{res}$ , which is a measure of service availability, is equal to

$$P_{block}^{res} = P_{N_{\max}} \lambda_{N_{\max}} / \lambda_{ave}, \quad (22)$$

whereas the call blocking probability due to too high a price,  $P_{block}^{price}$ , which is a measure of social fairness<sup>3</sup> of the tariff, is equal to

$$P_{block}^{price} = \sum_{i=0}^{N_{\max}} P_i (1 - g_i). \quad (23)$$

In addition, the average total revenue obtained by the network operator in a given time period  $T_O$  is given by the following equation in the steady-state condition:

$$Q_{T_O} = N_{T_O} Q_{call,ave}, \quad (24)$$

where  $N_{T_O}$  is the number of flows that are served in the period  $T_O$  (equal to  $\lambda_{ave}^* T_O$ , if  $T_O$  is long enough), and  $Q_{call,ave}$  is the average revenue per call, which is equal to

<sup>3</sup> In accordance with [2], we consider tariffs to be socially fair if access to the network is not determined by users' wealth.

$$Q_{call,ave} = \frac{1}{\mu} \sum_{i=0}^{N_{max}-1} Q_i P_{Q_i}, \quad (25)$$

$P_{Q_i}$  being 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 [45]:

$$P_{Q_i} = \lim_{\Delta t \rightarrow 0} \Pr[N(t) = i | A(t, t + \Delta t)] = \lim_{\Delta t \rightarrow 0} \frac{\Pr[A(t, t + \Delta t) | N(t) = i] P_i}{\Pr[A(t, t + \Delta t)]} = \lim_{\Delta t \rightarrow 0} \frac{\Delta t \lambda_i P_i}{\Delta t \lambda_{ave}^*} = \frac{\lambda_i P_i}{\lambda_{ave}^*}, \quad (26)$$

where  $A(t, t + \Delta t)$  is the event that an arrival occurs in the interval  $(t, t + \Delta t)$ .

Also, from Eqs. (24), (25), and (26), the total revenue can be rewritten as

$$Q_{T_o} = \frac{T_o}{\mu} \sum_{i=0}^{N_{max}-1} Q_i P_i \lambda_i. \quad (27)$$

Consequently, the average per-second revenue is equal to

$$Q_s = \frac{1}{\mu} \sum_{i=0}^{N_{max}-1} Q_i P_i \lambda_i \quad (28)$$

Finally, the average customer surplus (i.e., the average difference between the willingness of a user to pay for a given service and the amount he/she is charged for),  $S_{ave}$ , which can be viewed as a measure of customer satisfaction from the negotiation, is equal to

$$S_{ave} = \sum_{i=0}^{N_{max}-1} P_{Q_i} \int_{Q_i}^{\infty} (u - Q_i) P_U(u) du. \quad (29)$$

Note that we have set the customer surplus to 0 when a call is not accepted due to either lack of resources or excessive price. A more meaningful parameter is the average customer surplus normalized to the average willingness of users to pay,  $S_{ave} / \bar{U}$ .

## 7.2. Network scenario

In order to show the application of our proposals in a realistic scenario, we consider a network operator offering the VoIP service, charged according to dynamic prices, depending on the current network status, as illustrated in section 6. The traffic parameters and the system capacity are those already used in previous sections. The range of QoS the operator is able to offer is the one depicted in the abscissa of Fig. 5 ( $d \in [1.4362 \ 1.94]$ ). We assume that  $N_{max}$  is equal to  $\lfloor C / B_{res} \rfloor$ , where  $B_{res}$  is equal to the effective bandwidth computed on the equivalent node, strictly dependent on the QoS level. Once the operator has defined the appropriate QoS level to offer, the maximum number of supported

flows  $N_{\max}$  follows. In this case  $N_{\max}$  can range between 100 (corresponding to  $d=1.436$ ) and 114 (corresponding to  $d=1.94$ ). Clearly, a lower quality level allows increasing the maximum number of supported flows.

The average call duration is set at 240 s ( $\mu=1/240$  s<sup>-1</sup>), whereas the call arrival rate  $\lambda$  is a variable parameter. This model is typically used for voice services [45].

We assume that the network operator can preliminarily estimate the willingness of users to pay, thus the distribution of the random variable  $U(d)$ , which depends on the QoS level of the service. This is beyond the scope of this work, and even though we are conscious of the complexity of modeling users' demand, for the sake of simplicity and without loss of generality of the analysis presented in the previous sub-section, we assume that:

- the willingness of a user to pay for a given value  $d$  is a random variable, uniformly distributed between  $U_{\min}(d)$  and  $U_{\max}(d)$ . We recall that the values  $a$  and  $b$  ( $a=1.85$  and  $b=10$ ) characterizing the function  $f(d)$  are obtained from the behavior of the average willingness to pay  $\bar{U}(d)$  (i.e.,  $\frac{\bar{U}(d_1)}{\bar{U}(d_2)} = \frac{f(d_1)}{f(d_2)}$ ), as discussed in section 3.2;
- $\frac{U_{\min}(d_1)}{U_{\min}(d_2)} = \frac{U_{\max}(d_1)}{U_{\max}(d_2)} = \frac{\bar{U}(d_1)}{\bar{U}(d_2)} = \frac{f(d_1)}{f(d_2)}$ . For instance, for  $d=1.91$  s,  $U_{\min}=0.06$  \$/s and  $U_{\max}=0.2$  \$/s.

### 7.3. Numerical results

In the previous two sub-sections, we have defined the performance metrics and illustrated the reference network scenario. Here, our goal is to set both the QoS level (i.e.,  $d$ ) and price parameters (i.e.,  $\alpha$  and  $m$ ) relevant to a VoIP service. Clearly, in principle, the objectives of the optimization problem are up to operator that is possibly willing to put the service into the marketplace. In this work, our choice is to maximize the operator's revenue while considering constraints in terms of price blocking probabilities.

We start solving the following optimization problem:

$$\max_{d \in [d_{\min}, d_{\max}], m \geq 0, \alpha \geq 0} Q_s \quad (30)$$

In other words, our goal is to maximize the total per-second income of the operator, without any special constraint. The problem has been solved numerically for different values of the offered load (in erlangs). The offered load varies in the range from 90 to 100 erlangs, since, with respect to the possible values of  $N_{\max}$  (depending on the QoS level,  $d$ ), it represents a normalized load in the range from 0.79 to 1, which is typical in operation.

Table I reports the optimal QoS level,  $d$ , and the optimal price parameters ( $m$  and  $\alpha$ ) values which maximize the per-

second revenue of the operator,  $Q_s$ . We also report the other performance metrics: the call blocking probability due to too high a price,  $P_{block}^{price}$ , call blocking probability due to the lack of resources,  $P_{block}^{res}$ , resource utilization,  $\rho_{ave}$ , and normalized customer surplus,  $S_{ave}/\bar{U}$ .

Offered Load	Optimization variables			Performance metrics				
	$d$ (s)	$\alpha$ (\$)	$m$	$Q_s$ (\$/s)	$P_{block}^{price}$	$P_{block}^{res}$	$\rho_{ave}$	$S_{ave}/\bar{U}$
90	1.4362	$1.34 \cdot 10^{-5}$	0.01	17.2146	0.2867	$7.05 \cdot 10^{-6}$	0.426303	0.27397
95	1.4362	$1.34 \cdot 10^{-5}$	0.01	18.1702	0.28728	$4.45 \cdot 10^{-5}$	0.45	0.273529
100	1.4362	$1.34 \cdot 10^{-5}$	0.01	19.1232	0.2878	$2.13 \cdot 10^{-4}$	0.472809	0.27309

Table I: solution of the optimization problem (Eq. (30)) with dynamic pricing scheme.

The main comment is that the variables which identify the optimum are constant with the offered load. Clearly, the higher the load, the higher the operator's revenue, the utilization coefficient, and the call blocking probabilities. It is worth noting that the coefficient  $m$  is very small; this means that the variability of the price is very low. Also, the QoS to be offered on the marketplace is the highest (i.e., the lowest  $d$  value).

However, from the results in Table I, the high value (about 28%) of the blocking probability, due to users' inability to pay, leaps to the eye. This results in a pricing scheme which is not socially fair, and the operator may like to take into account this aspect. In fact, the choices of the operator should be governed by the following considerations. If it is assumed that the user behavior is not influenced by previous experience (stateless user model), then the operator could maximize the revenue without accounting for the call blocking probability due to very high prices if compared to the willingness of users to pay. This could happen when the operator exercises a monopoly, and also when users (or a user agent [46] or a broker [16,17] on their behalf) mechanically scan all network offers and choose the one that currently maximizes their benefit, without considering previous failed/successful price negotiations. On the other hand, if it is assumed that the users' behavior is influenced by previous experience (stateful user model), then the network operator could also sacrifice part of current potential revenues in order to guarantee service provisioning, and consequently to safeguard future revenues.

For this reason, we modify the optimization problem by including a constraint on the maximum value of the  $P_{block}^{price}$ , thus obtaining the problem:

$$\begin{aligned} & \max_{d \in [d_{\min}, d_{\max}], m \geq 0, \alpha \geq 0} Q_s \\ & P_{block}^{price} \leq \overline{P_{block}^{price}} \end{aligned} \quad (31)$$

For instance, if we set this constraint to  $\overline{P_{block}^{price}} = 5\%$ , Table II reports the results of the optimization problem.

Offered Load	Optimization variables			Performance metrics				
	$d$ (s)	$\alpha$ (\$)	$m$	$Q_s$ (\$/s)	$P_{block}^{price}$	$P_{block}^{res}$	$\rho_{ave}$	$S_{ave}/\overline{U}$
90	1.4362	$9.18 \cdot 10^{-6}$	0.14	15.1421	0.0498	$1.25 \cdot 10^{-2}$	0.560796	0.4863
95	1.4362	$9.39 \cdot 10^{-6}$	0.29	15.747	0.0499	$2.68 \cdot 10^{-2}$	0.583	0.486447
100	1.4722	$9.8 \cdot 10^{-6}$	0.51	16.181	0.04997	$4.25 \cdot 10^{-2}$	0.604058	0.48696

Table II: solution of the optimization problem (Eq. (31)) with dynamic pricing scheme.

As expected, this constraint decreases the operator's revenue. In addition, we remark that the value of  $m$  is more meaningful than in the previous case, thus resulting in both a higher price dynamics and an increased normalized customer surplus. The increasing of  $S_{ave}/\overline{U}$  is also due to less rejected calls.

The operator may be also interested in offering a service characterized by a low call blocking probability value due to lack of resources (i.e., when the system is full). This implies that the service availability is high, which may be a desirable feature to obtain customer fidelity and to attract new ones, and thus to safeguard future revenues. For this reason, we introduce in the optimization problem another constraint on the maximum value of the  $P_{block}^{res}$ , thus obtaining the problem:

$$\begin{aligned} & \max_{d \in [d_{\min}, d_{\max}], m \geq 0, \alpha \geq 0} Q_s \\ & \begin{cases} P_{block}^{price} \leq \overline{P_{block}^{price}} \\ P_{block}^{res} \leq \overline{P_{block}^{res}} \end{cases} \end{aligned} \quad (32)$$

For obtaining numerical results we have set this constraint to  $\overline{P_{block}^{res}} = 0.5\%$ . Table III reports the outcome of the optimization problem.

Offered Load	Optimization variables			Performance metrics				
	$d$ (s)	$\alpha$ (\$)	$m$	$Q_s$ (\$/s)	$P_{block}^{price}$	$P_{block}^{res}$	$\rho_{ave}$	$S_{ave}/\overline{U}$
90	1.54421	$1.07 \cdot 10^{-5}$	0.57	14.6179	0.04975	$4.84 \cdot 10^{-3}$	0.565168	0.49341
95	1.68823	$1.26 \cdot 10^{-5}$	0.98	12.9247	0.0499	$4.76 \cdot 10^{-3}$	0.596	0.5148
100	1.86826	$1.23 \cdot 10^{-5}$	0.48	7.67512	0.04995	$4.87 \cdot 10^{-3}$	0.627823	0.49710

Table III: solution of the optimization problem (Eq. (32)) with dynamic pricing scheme.

As expected, the second constraint further decreases the optimal operator's income. Also, such an income decreases with the offered load. This effect is due to the presence of the two constraints on blocking probabilities. In fact, introducing an upper bound on  $P_{block}^{price}$  means decreasing the price and thus increasing  $\lambda_{ave}$  (i.e., the number of calls overcoming the price acceptance control and being subject to the CAC, see Fig. 10). If  $\lambda_{ave}$  increases and there is a bound on  $P_{block}^{res}$ , in order to satisfy the two constraints, the operator has to decrease the QoS level (i.e., to increase  $d$ ). In fact, lowering the QoS level means increasing the maximum number of admissible flows. Also, note that the values of  $m$  are large and, as expected,  $S_{ave}/\bar{U}$  slightly increases with respect to the previous case.

We have also solved the optimization problem assuming a static pricing scheme (i.e.,  $m=0$ ). In this case the operator charges a constant price independent of the current network status. Table IV reports the results of the optimization problem.

Offered Load	Optimization variables		Performance metrics				
	$d$ (s)	$\alpha$ (\$)	$Q_s$ (\$/s)	$P_{block}^{price}$	$P_{block}^{res}$	$\rho_{ave}$	$S_{ave}/\bar{U}$
90	1.65223	$9.42 \cdot 10^{-6}$	13.6232	0.0492	$4.08 \cdot 10^{-3}$	0.565932	0.48678
95	1.83226	$9.87 \cdot 10^{-6}$	8.90786	0.0492	$4.24 \cdot 10^{-3}$	0.597	0.4867
100	No solution found		-	-	-	-	-

Table IV: solution of the optimization problem (Eq. (32)) with static pricing scheme ( $m=0$ ).

The first comment is that the operator's revenue is reduced with respect to the cases of dynamic prices. In fact, a price depending on the current network status allows tracking the dynamics of the service demand. In other words, the price parameter  $m$  introduces an additional degree of freedom to adapt performance metrics, thus enlarging the space of solutions of the optimization problem. In fact, when the system is very stressed (i.e., a high value of offered load and hard constraints on blocking probabilities), the static pricing scheme does not offer a viable solution. The decrease of revenue is around 7%, 31%, and 100%, for the three offered loads, respectively. It is also important to note that the dynamic pricing scheme guarantees higher values of normalized customer surplus.

All the optimization problems above are non linear and can be solved by using a number of methods (e.g., see [47]). In order to verify the tolerance of the worst-case computational burden to find the optimum, we have used an exhaustive search with a very dense search grid. For all the optimization problems analyzed, we have found the solution by six minutes, using a standard PC. Since this computation must be done once at the beginning of each

market activity, the relevant time is not a problem.

## 8. IMPLEMENTATION ISSUES

Some considerations have to be made regarding the applicability of the proposed pricing model in real networks.

The first is that our model foresees a strong QoS support at IP level, and, at this stage, this proves to be an impediment, since neither IntServ nor DiffServ have not been widely deployed yet though implemented in many high level commercial routers and layer-3 switches. However, in principle, our pricing model is compliant with both approaches. All we need is for the administrative domain to be able to provide a number of classes of network service characterized by specific QoS parameters and tailored to support a specific application service. In this work, we have especially dealt with voice over IP service, but the analysis is general and could also be used for other services (i.e., video services). The QoS level of a class of service is guaranteed a priori, since the network operator is able to control the amount of traffic. The differentiation of the service level is mapped to different prices through the virtual-delay, which is a technical estimation of the QoS level. We stress that in the considered scenario, the virtual delay does not have to be measured for each call, but it is evaluated a priori on the basis of the service guarantees associated with the considered class of service.

Another important implementation issue concerns the accounting aspects to collect resource consumption data to support usage-based charging. To accomplish an accounting management architecture within each administrative domain, three main entities are identified in [13]: *network devices*, *accounting servers*, and *billing servers*. Network devices collect measurement data and send them by means of an accounting protocol (e.g., RADIUS, TACACS+ or SNMP) to accounting servers that generate the *session records*, in which accounting information is stored [14]. Network devices and accounting servers are logically separated but can be implemented in the same device. Session records are then sent to the billing server (which may or may not be a centralized entity), the aim of which is to compute the charge and generate invoices to users. The rules for generating, transporting and storing accounting data are known as *accounting policies* [15]. Our pricing model foresees traffic measurement on a per-call basis at the ingress nodes only. This means that accounting procedures are limited to the edge part of the network and that core nodes are not interested in them. Each edge router stores a record for each customer, and accounting data are stored in this record. Accounting devices at the ingress nodes are correctly configured during the call set-up phase, i.e., before data transmission commences.

As regards the price setting accomplished in the previous section for a specific network scenario, it is worth noting that assumptions on the service demand made to perform the analysis enabled us to obtain quantitative results from a theoretical analysis. Nevertheless, both the general applicability and the qualitative behavior of the system model are preserved under different assumptions, and the quantitative analysis can be performed nonetheless from a theoretical (when possible) or simulative analysis. As regards the estimation of user service demand, we remark that the results presented in this paper may be exploited as a potential building block of an all-encompassing vision of future, personalized and easy-to-use services, pursued within the framework of the Simplicity (<http://www.ist-simplicity.org>) and Primo (<http://primo.ismb.it>) projects. The objective is to provide each user with a personalized profile, stored in a portable device. Ideally, by plugging this device (e.g., a JAVA card or a USB stick) into the terminal, each user will personalize both terminal and services alike. The charging approach is clearly an important component from this point of view. Furthermore, the availability of users' profiles and preferences offers an interesting implementation venue for the concepts described in this paper, since it could help the network operator to retrieve user information for management purposes, e.g., to promptly estimate user service demand.

## **9. CONCLUSIONS AND FUTURE WORK**

In this paper, we have addressed the ever-increasing problem in the evolution of the Internet of defining a clear, objective and convincing pricing criterion to charge for value-added network services to support inelastic applications. We have described a novel criterion, based on the concept of the so-called virtual delay, to identify the quality of the edge-to-edge service provided by an administrative domain. We have illustrated how to determine this QoS index from a purely technical point of view. The virtual delay is able to take into account in a simple, flexible way not only QoS parameters and the cost of deploying various QoS features, but also the actually perceived QoS level. We have presented an analysis of the virtual delay, and consequent tariff, when system parameters vary. This analysis confirms the consistency of the approach.

Furthermore, we have extended the concept of virtual delay, by making it dependent on the status of resource availability. As regards QoS-enabled networks, admission control should avoid network congestion. However, this does not hinder a network operator from taking important decisions under competitive conditions. In fact, suitable pricing policies might be applied to control the traffic offered. The difficult decision is to determine both the QoS level of the service to put into the market and the relevant price settings. In this way, it is possible to maximize profits

while preserving social fairness and guaranteeing service availability. However, in principle, the choice of the performance objective is up to the operator, and the quantitative analysis can be repeated accordingly. In the proposed tariff model, the price can vary dynamically according to the amount of service demand and is an additional, effective network control tool, which allows an operator to finely tune performance metrics. Once the QoS level of a service has been fixed and the relevant demand has been estimated, the operator has only to configure a few initial pricing parameters on the basis of the performance desired. Our results offer a tool to simplify this task. After this initial configuration, the ongoing price value is automatically obtained according to the current network status.

Finally, we have discussed a number of challenges related to the implementation of our pricing model in real networks, and in particular, issues related to QoS, accounting, and pricing configuration. The main comment is that our approach is feasible in real networks, and that the real impediment to its deployment is the implementation of QoS mechanisms, which is a basic assumption for our work, since, at this stage, the objective of our tariff model is to charge for network services supporting inelastic application with strong QoS guarantees.

In this respect, future work will investigate the possibility of extending the proposed pricing model to charge for traffic supporting elastic applications, for which the throughput is a fundamental QoS requirement to be taken into account in the model. Moreover, another future objective is to investigate the potentiality of the Simplicity approach to retrieve useful user information to help the operator calculate a dynamic pricing configuration.

## ACKNOWLEDGEMENTS

The authors sincerely thank the anonymous referees for their constructive suggestions and comments that have largely contributed to the improvement of the quality of this paper.

This work has been co-funded by the European Union in the framework of the IST project SIMPLICITY and by the Italian Ministry of Education, University, and Research (MIUR) within the FIRB project PRIMO.

## 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, 26(18), December 2003.
2. M. Falkner, M. Devetsikiotis, I. Lambadaris, An overview of pricing concepts for broadband IP networks, *IEEE Comm. Surveys*, Second Quarter 2000.
3. L.A. DaSilva, Pricing for QoS-enabled networks: a survey, *IEEE Communications Surveys*, Second Quarter 2000.
4. C. Courcoubetis, R. Weber, *Pricing Communication Networks: Economics, Technology and Modelling*, Wiley Ed., March 2003.
5. S. Shenker, Fundamental design issues for the future Internet, *IEEE JSAC*, 13(7), September 1995.
6. J. Altmann, K. Chu, "A proposal for a flexible service plan that is attractive to users and Internet service providers," *IEEE INFOCOM 2001*, Anchorage, USA, April 2001.

7. C. Courcoubetis, S. Sartzetakis, V. A. Siris, G. D. Stamoulis, *Charging Communication Networks: from Theory to Practice*, D.J. Songhurst Ed., Elsevier, Amsterdam, 1999.
8. Charging and Accounting Technologies for the Internet (CATI), <<http://www.tik.ee.ethz.ch/~cati/>>.
9. Internet Next Generation project, <<http://ing.ctit.utwente.nl>>.
10. WHYLESS.COM – The Open Mobile Access Network, <<http://www.whyles.org>>.
11. MOBIVAS, <<http://mobivas.cnl.di.uoa.gr/>>.
12. G. Huston, Next steps for the IP QoS architecture, IETF RFC 2990, November 2000.
13. B. Aboba, J. Arkko, D. Harrington, Introduction to accounting management, IETF RFC 2975, October 2000.
14. N. Brownlee, A. Blount, Accounting attributes and record formats, IETF RFC 2924, September 2000.
15. G. Carle, S. Zander, T. Zseby, Policy-based accounting, IETF RFC 3334, October 2002.
16. D. Di Sorte, M. Femminella, G. Reali, S. Zeisberg, Network service provisioning in UWB open mobile access networks, IEEE JSAC, 20(9), Dec. 2002.
17. D. Di Sorte, G. Reali, Minimum price inter-domain routing algorithm, IEEE Communications Letters, 6(4), April 2002.
18. D. Di Sorte, M. Femminella, G. Reali, A QoS index for IP services to effectively support usage-based charging, IEEE Comm. Letters, 8(11), Nov. 2004.
19. S. Shenker, D. Clark, D. Estrin, S. Herzog, Pricing in computer networks: reshaping the research agenda, ACM Computer Comm. Review, 26(2), 1996.
20. X. Wang, H. Schulzrinne, RNAP: a resource negotiation and pricing protocol, NOSSDAV 1999, Basking Ridge, USA, June 1999.
21. D. Di Sorte, G. Reali, Pricing and brokering issues over interconnected IP networks, Journal of Network & Computer Applications, Elsevier, 28(4), November 2005.
22. J. Hwang, J. Altman, H. Oliver, A. Suarez, Enabling dynamic market-managed QoS interconnection in the next generation Internet by a modified BGP mechanism, IEEE ICC2002, New York, USA, April-May 2002.
23. R. Cocchi, D. Estrin, S. Shenker, L. Zhang, A study of priority pricing in multiple service class network, ACM SIGCOMM, Zurich, Switzerland, Sept. 1991.
24. A. Odlyzko, Paris metro pricing for the Internet, ACM Conference on Electronic Commerce (EC 1999), 1999.
25. D. Clark, A model for cost allocation and pricing in the Internet, MIT Workshop on Internet Economics, Cambridge, USA, 1995.
26. P. Reichl, B. Stiller, T. Ziegler, Charging multi-dimensional QoS with the Cumulus pricing scheme, SPIE ITCom, Denver, USA, August 2001.
27. S. Jordan, H. Jiang, Connection establishment in high speed networks, IEEE JSAC, 13(7), September 1995.
28. C. Corcoubetis, F. Kelly, V.A. Siris, R. Weber, A study of simple usage-based charging schemes for broadband networks, Telecommunication Systems, Kluwer, 15(3-4), 2000.
29. C. Corcoubetis, F. Kelly, R. Weber, Measurement-based usage charges in communication networks, Operation Research, 48(4), July-August 2000.
30. 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.
31. K. Kumaran, M. Mandjes, Multiplexing regulated traffic streams: design and performance, IEEE INFOCOM 2001, Anchorage, USA, April 2001.
32. S. Shenker, C. Partridge, R. Guerin, Specification of Guaranteed Quality of Service, IETF RFC 2212, Sept. 1997.
33. R. Braden, D. Clark, S. Shenker, Integrated services in the Internet architecture: an overview, IETF RFC 1633, June 1994.
34. S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An architecture for Differentiated Services, IETF RFC 2475, December 1998.
35. J.W. Lee, R.R. Mazumdar, and N.B. Shroff, Non-convex Optimization and Rate Control for Multi-class Services in the Internet, to appear in the IEEE/ACM Transactions on Networking.
36. M. Chiang, S. Zhang, and P. Hande, Distributed rate allocation for inelastic flows: Optimization frameworks, optimality conditions, and optimal algorithms, IEEE INFOCOM, Miami, USA, March 2005.
37. F. Lo Presti, Zhi-Li Zhang, J. Kurose, D. Towsley, Source Time Scale and Optimal Buffer/Bandwidth Trade-off for Regulated Traffic in an ATM Node, IEEE INFOCOM, Kobe, Japan, 1997.
38. B. Goode, Voice Over Internet Protocol (VoIP), Proceedings of the IEEE, 90(9), September 2002.

39. D.D. Botvich, N.G. Duffield, Large deviations, economies of scale, and the shape of the loss curve in large multiplexers, *Queueing Systems*, 20, 1995.
40. W. Wang, H. Schulzrinne, Pricing network resources for adaptive applications in a differentiated services network, *IEEE INFOCOM 2001*, Anchorage, USA, April 2001.
41. A. Ganesh, K. Laevens, R. Steinberg, Congestion pricing and user adaptation, *IEEE INFOCOM 2001*, Anchorage, USA, April 2001.
42. M. Caesar, D. Ghosal, R.H. Katz, Resource management for IP telephony networks, *IWQoS 2002*, Miami Beach, USA, May 2002.
43. I.C. Paschalidis, J.N. Tsitsiklis, Congestion-dependent pricing of network services, *IEEE/ACM Transactions on Networking*, 8(2), April 2000.
44. 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.
45. L. Kleinrock, *Queueing Systems, Volume I: Theory*, Wiley Interscience, New York, 1975.
46. J. Altmann, P. Varaiya, Managing usage-based pricing in a future telecommunication market, *PAAM 1999*, London, UK, April 1999.
47. S Hillier, Gerald J Lieberman, *Introduction to Operations Research*, 8<sup>th</sup> edition, McGraw-Hill, 2004.
48. M. Mandjes, J.-H. Kim, Large deviations for small buffers: an insensitivity result, *Queueing Systems*, 37 (2001).

## APPENDIX

In this Appendix, our goal is to provide readers with more details about the model used in this paper to compute the values of effective bandwidth and buffer when a small amount of packet losses may be tolerated within a network node with capacity  $C$  and buffer  $B$  ( $B/C = D_{jitter}$ ).

We undertake to calculate the effective bandwidth and buffer pair  $(c_{p0}, b_{p0})$  by using the approach illustrated in [31], in particular with the small buffer approximation ([31], section III). In more detail, having defined

$$\alpha(c_{p0}) = \frac{c_{p0}}{P_{S0}} \log\left(\frac{c_{p0}}{r_{S0}}\right) + \left(1 - \frac{c_{p0}}{P_{S0}}\right) \log\left(\frac{P_{S0} - c_{p0}}{P_{S0} - r_{S0}}\right), \quad (29)$$

$$\beta(c_{p0}) = \frac{2}{P_{S0}} \sqrt{\log\left(\frac{c_{p0}(P_{S0} - r_{S0})}{r_{S0}(P_{S0} - c_{p0})}\right) \left(\frac{c_{p0}P_{S0} + P_{S0}r_{S0} - 2r_{S0}c_{p0}}{B_{TS0}}\right) + \left(\frac{P_{S0}(c_{p0} - r_{S0})}{B_{TS0}}\right)}, \quad (30)$$

following the result of [48], the Authors show that the effective bandwidth is the solution of the equation:

$$\alpha(c_{p0}) \cdot C / c_{p0} + \beta(c_{p0}) \cdot \sqrt{BC / c_{p0}} = -\log(P_{loss}) \quad (31)$$

Thus,  $c_{p0} = c_{p0}(P_{loss})$  and  $b_{p0}(P_{loss}) = D_{jitter} \cdot c_{p0}(P_{loss})$ . The maximum number of flows  $N$  that can be allocated in the node  $(B, C)$  while guaranteeing the performance  $(D_{jitter}, P_{loss})$  is equal to  $N = C / c_{p0} = B / b_{p0}$ .

## BIOGRAPHIES

**Dario Di Sorte:** is a Post-Doc researcher with the Networking Group at the Dipartimento di Ingegneria Elettronica e dell'Informazione (DIEI) of the University of Perugia, Italy. He received the Ph.D. in "Telematics and Information Society" from the University of Florence in June 2003, and the "Laurea" degree in Electronic Engineering magna cum laude from the University of Perugia in January 2000. His research interests focus on IP networks, especially on Quality of Service, pricing, and mobility management.

**Mauro Femminella** is a researcher of the Department of Information and Electronic Engineering of the University of Perugia, Italy. He received the "Laurea" and the Ph.D. degree in Electronic Engineering on 1999 and 2003, respectively, both from the University of Perugia. His research interests focus on satellite networks, content delivery networks, wireless LANs, and IP networks, particularly on quality of service, pricing and mobility management.

**Gianluca Reali** is an Associate Professor of the Department of Information and Electronic Engineering of the University of Perugia since January 2005. He received the Ph.D. degree in Telecommunications from the University of Perugia in 1997 working on Spread Spectrum techniques. He was a researcher at the University of Perugia from 1997 to 2004. He has worked in IP networks, particularly in transport and resource management protocols. He was involved in many European ACTS and IST projects and FIRB and PRIN projects co-funded by the Italian government.