Available online at www.sciencedirect.com
SCIENCE DIRECT
Computer Communications 27 (2004) 1868–1880
computer communications
www.elsevier.com/locate/comcom

SLA-based QoS pricing in DiffServ networks

Christos Bouras[a,b,*], Afrodite Sevasti[b,c]

[a]RA Computer Technology Institute—RACTI, 61 Riga Feraiou Str., Kolokotroni 3, 26221 Patras, Greece
[b]Department of Computer Engineering and Informatics, University of Patras, 26500 Rion, Patras, Greece
[c]Greek Research and Technology Network—GRNET, 56 Mesogion Avenue, 11574, Athens, Greece

Received 23 September 2003; revised 4 June 2004; accepted 23 June 2004
Available online 25 July 2004

Abstract

The availability of high-speed transmission media and networking equipment in contemporary networks, as well as the evolution of quality-demanding applications has focused research interest on the provision of advanced qualitative services in addition to the traditional best-effort model of the Internet. A number of alternatives for service differentiation and QoS provisioning have been proposed and standardized, but in the case of backbone, transport networks the DiffServ architecture has prevailed, due to its scalability and deployment feasibility. The provisioning of services according to the DiffServ framework has in turn raised the requirements for interdependent, controlled resource allocation and service pricing, with particular needs for pricing mechanisms that preserve the potential and flexibility of DiffServ. At the same time, such mechanisms should reflect resource usage, allocate resources efficiently, reimburse costs or maximize service provision profits and lead customers to requesting services that will maximize their revenue. In this work, after reviewing related research, the principles that a pricing scheme for DiffServ-based services should follow are presented, stressing the differences form traditional Internet pricing. Based on these principles, an analytical approach to pricing a particular class of DiffServ-based services and a methodology for applying this approach in a real network are proposed and evaluated.

Keywords: QoS charging and pricing; DiffServ; Traffic profile; Service level agreement

1. Introduction

An important issue in designing pricing policies for today's networks, is to balance the trade-off between engineering and economic efficiency. This trade-off, which is more or less constrained by the underlying network technology and the network services provided, has many dimensions. Some of them are how much measurement is required for the pricing policy to be enforced, the granularity of differently priced services, the level of resource aggregation at which pricing is done—both in time and in space—and the information required by the network for billing. In Ref. [1], it is emphasized that pricing schemes that determine prices over short intervals in order to maximize economic efficiency may be unrealistic. Instead, schemes where the utility and cost functions are known and valid for a duration longer than a connection's duration are recommended. It is also recommended to keep the costs' calculation simple and the monetary amounts that the customers will be asked to pay predictable. Results from Ref. [2], based on strong evidence of the history of all communication technologies and users' reactions claim that even the slightest attempt to impose complex, incomprehensible charging will have a substantial negative impact on usage.

All these principles for keeping pricing schemes simple and predictable seem to contradict the complexity introduced to the traditional best-effort service model of the Internet by the prevalence of the DiffServ model. DiffServ has been accepted as a means to provide service differentiation with credible QoS guarantees to individual flows crossing large transport networks without per-flow state maintenance and reservations, demonstrating thus a

* Corresponding author. Tel.: +30-2610-960375; fax: +30-2610-960358.
E-mail addresses: bouras@cti.gr (C. Bouras), sevasti@grnet.gr (A. Sevasti).

0140-3664/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2004.06.010

remarkable scalability. As such, DiffServ seems to be a promising solution for efficiently supporting the QoS demanding applications of the future. For more details on the DiffServ framework principles, the reader is directed to Ref. [3].

DiffServ anticipates for classification of individual flows in a small number of service classes at network edges as well as 'soft' reservation of resources and special handling of packets per service class, in the core. Allocation of a different amount of resources to each service class, differential treatment for packets and variety in the QoS guarantees provided are obvious reasons why the universal pricing schemes of the traditional best-effort Internet are no longer adequate. Differentiation of service must also be reflected in the pricing schemes used and this comprises a major challenge for the research community. The DiffServ principles apply mainly to transport backbone networks that serve thousands of flows simultaneously. Providers of such networks require efficient means to charge for the service differentiation and QoS they provide to their customers.

Until today, many proposals for pricing of DiffServ-based services have followed the 'usage-based per service class' model. Most approaches suggest a flat per-packet price within a certain service class and charge all traffic belonging to this service class according to this price. We claim that, for DiffServ-based services, a flat per packet or per transmitted-volume-unit price within a service class is not efficient from an economical and engineering point of view. There has to be some kind of differentiation in charging within the packets belonging to the same traffic class, to anticipate for over-subscriptions, congestion within a certain service class, etc. We propose a pricing scheme that applies to a significant portion of DiffServ-based services, demonstrates engineering and economic efficiency, preserves simplicity in calculation of customers' charges and effectively reveals the details of service differentiation and QoS provisioning. Our approach is innovative because it anticipates for externalities hidden in the costs involved and caused by the nature of such DiffServ services and also because it goes all the way up to the determination of actual prices. In economic theory, externalities are referred to as costs (for negative externalities) or benefits (for positive ones) that do not accrue to the consumer of the good ([4]).

In Section 2 of this paper a conceptual overview of the issues involved in QoS pricing is given and a set of principles are defined. Section 3 presents related research and Section 4 provides the architectural framework to which our QoS pricing approach applies. In Section 5, the analytical model of our proposal is thoroughly defined while Section 6 outlines a methodology for applying our approach in a real network and Section 7 presents an evaluation of the methodology. Our future work and conclusions are provided in the last section.

2. QoS and pricing

The introduction of QoS and differentiation in contemporary networks has advanced the role of pricing. Prior to this, pricing approaches were rather simplistic, focusing on a fair distribution of the costs for the provider to a population of customers. Theoretical models that were rarely adopted in practise, due to their complexity, would go one step further and use pricing as a measure for controlling congestion and discouraging customers from overloading the network.

Enhancing the network with a number of service classes differing in the qualitative guarantees provided, requires enhanced pricing models that, in addition, drive customers to an appropriate selection of a service class that maximizes their perceived utility. Using flat pricing in a network with multiple levels of QoS would not discourage all customers from selecting the highest, in terms of QoS guarantees provided, service class to carry their traffic. Congestion in this service class would then be inevitable and the quality offered would be compromised.

Thus, new roles have been appointed to pricing with the advent of QoS and service differentiation:

- Pricing should effectively reflect the utility of choosing a particular service class for each customer, co-estimating the quality guarantees that each service class provides. In this way, customers will refrain from using the service class with the highest quality in cases where the utility they perceive is not equally high because this will entail unprofitable costs for them
- Pricing must ensure incentive compatibility, or in other words the motivation for customers to express their demands for network resources within a particular service class in a reasonable manner. In this way, customers will not impose excessive requests for resources, unless they are prepared to spend in an unprofitable way. With respect to this dimension, it has to be emphasized that it is critical to provide QoS guarantees in high-speed networks in a controlled manner in terms of use of network resources. Indicatively, for traffic that can tolerate a certain delay due to packets' accumulation in buffers, buffer capacity is a scarce resource that should be carefully managed in resource allocation.

DiffServ-enabled networks are based on open loop congestion control mechanisms. For every flow or aggregate of flows being transmitted, there exists a traffic contract (most of the times referred to as service level agreement—SLA), which contains the agreed QoS parameters and a traffic descriptor or profile that the flow must obey. The traffic profile is usually such that it denotes the resources (e.g. in terms of bandwidth and buffer space) that a flow will occupy during transmission.

As long as its traffic descriptor is not violated, a flow is transmitted unaltered over the network equipment. However, since obeying to a traffic descriptor involves shaping and/or policing of traffic (by the network or the customer himself) according to the traffic contract, the traffic contract parameters are a means to effectively control the amount of resources that each flow is using. Therefore, a pricing scheme has to co-estimate these parameters in order to charge for transmission (see Fig. 1). Still, the traffic contract cannot be the only coefficient of pricing, since it is always possible that a flow contracted to conform to a traffic profile is actually using less resources. Such a flow would then be unfortunately charged for using more network resources than it actually would.

In a network that offers service differentiation and QoS, the utility function of customers is no longer solely dependent upon the volume of traffic being transmitted and the congestion experienced. It also depends upon the quality metrics guaranteed (such as end-to-end delay, jitter and packet loss) to the customer's traffic as well as the amount of resources within a particular service class that the customer's traffic occupies.

If we depict by $p_{c_k}(S_i)$ the costs that a customer has to pay for purchasing an SLA with the S_i traffic profile under service class c_k, then the objective of a pricing mechanism should be that of maximizing

$$U_{c_k}(S_i) + U(Q_{c_k}) - p_{c_k}(S_i) \quad (1)$$

for each customer K_i, where

- $U_{c_k}(S_i)$ the utility perceived by K_i through an SLA with traffic profile S_i for service by c_k
- $U(Q_{c_k})$ the utility (either positive or negative) of K_i from a set of quality guarantees (Q_{c_k}) offered by c_k
- $p_{c_k}(S_i)$ the price to be paid by K_i signed with the S_i SLA and receiving the treatment of c_k

It is obvious that the maximization of Eq. (1) over the sets of all customers and service classes of a network is not a trivial task, especially keeping in mind that $U_{c_k}(L_i)$ and $U(Q_{c_k})$ might differ among customers, especially when they demonstrate differing traffic patterns. It is highly likely that a number of relaxing assumptions will have to be made at this point for a pricing scheme to be comprehensible and deployable.

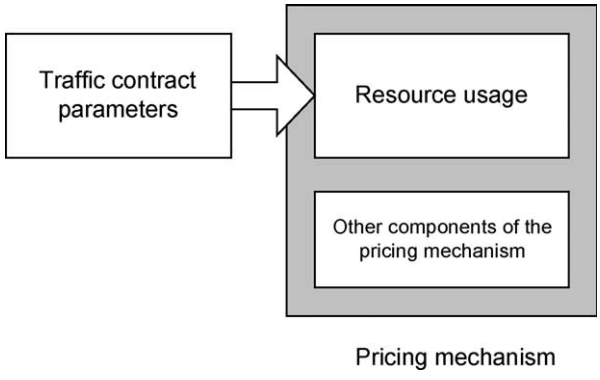

Fig. 1. How do traffic contract parameters affect pricing.

From the provider's point of view, resource usage from a traffic flow or aggregated flow is a reasonable basis upon which pricing of this particular traffic flow can be based. In the case of a transport network, thus, it is desirable to perform resources' dimensioning for the provision of each service provided taking into consideration the traffic profiles of the traffic flows or aggregates of each customer. Thus, the provider has to provide to his customers the incentives to describe their traffic profiles in the most accurate way, so that he will get a realistic estimate of the resources to be devoted to all the traffic aggregates belonging to each service class. We claim that this interdependence of traffic profiles and resources' dimensioning, as depicted in Fig. 1, must be regulated by appropriate pricing schemes, reflected in the last term of Eq. (1), $p_{c_k}(L_i)$.

Based on this conceptual approach to the problem of DiffServ-pricing or QoS-pricing, we present in the following section how the research community has dealt with this issue so far.

3. Pricing models

The evolution in networking that has emerged from the introduction of service differentiation and QoS provision by the DiffServ framework and equivalent approaches has affected traditional network pricing methodologies and shifted the interest from fixed access and connection fees to usage-based fees. The latter are considered appropriate to account for congestion costs, service differentiation, QoS provision and other relevant costs for pricing today's connectionless IP networks ([1,5]).

The surveys of Refs. [1,5] emphasize the role that a pricing mechanism must have on traffic management (congestion controls, resource provisioning and call admission). In Ref. [5], the author mentions that for the determination of a network pricing scheme one must decide on both the pricing policy and the price values that will be valid within the policy. Customer objectives (through a utility function) and provider objectives (either social fairness or maximisation of revenue or another goal) have to be modelled and a thorough understanding of how the engineering issues relate to pricing decisions is needed before trying to adopt pricing schemes closely related to traffic management.

In Ref. [6], the authors prove that differential pricing in multi-class networks results in better utilisation (combined cost and perceived performance) for all customers regardless of the service class they belong to, when compared to flat-rate pricing. Their results prove that different prices '…spread the benefits of multiple service classes around to all customers, rather than just having these benefits remain exclusively with customers who are performance sensitive'.

By introducing different prices for different classes the customers are led to choose the class that better suits their needs so that they will be served with the quality characteristics they need at the lowest possible cost.

Usage-based charging was traditionally based on accounting for the traffic flowing within a network, even in packet granularity, and then determining charges by multiplying the pre-determined price per packet with the number of packets transmitted. Later, usage-based pricing was proposed to account for congestion prices in traditional, best-effort networks. The 'smart market' approach that was introduced in Ref. [7], is based on per-packet charging and requires customers to declare their willingness to pay by bidding for network resources for each packet sent. Despite its accounting overhead, the 'smart market' approach has been innovative in introducing the notion of congestion pricing, in other words, the allocation of a congested resource in an analogous manner to each customer's valuation of it. As already explained, the DiffServ framework was designed so as to avoid fine granularity, dealing with traffic aggregates and keeping complexity at the edges of network domains. Therefore, in the case of DiffServ, per-packet or per-flow accounting has to be avoided, in order for the pricing scheme to preserve the scalability property.

One of the earliest works on the direction of pricing services provided by a DiffServ architecture is that of Ref. [8]. The purpose of this work, which sets the initial principles of DiffServ, is to introduce the 'expected capacity' framework, as a set of mechanisms that 'allocates' different amounts of bandwidth to different customers in a predictable and quite assuring way. This assurance for the bandwidth provisioned (or in other words the 'expected capacity') makes the latter a reliable basis for cost allocation. However, the proposition made is for a flat rate-like pricing where the customer pays for a certain access rate.

The establishment of long-term contracts (or SLAs) between the customer and the service provider, instead of detailed accounting, was initially proposed in Ref. [9]. These contracts contain the traffic profile negotiated between the provider and the customer. The profiles are in turn a very good approximation of the 'expected capacity' that the customer purchases from the network services' provider and thus are recommended as indication of resource usage by the customer and the basis for charging. However, Ref. [9] does not provide a specific solution to the determination of prices for different customers' expected capacities over different service classes.

The 'edge pricing' paradigm, presented in Ref. [10], complements 'expected capacity' pricing by shifting pricing activities to the edges of a domain but still does not provide a detailed solution for pricing of DiffServ-based services. Part of the 'edge pricing' paradigm is the approximation of congestion costs as the costs for transmitting during expected congestion conditions (QoS sensitive or class-based and time-of-day) along an expected path. In this way, pricing can be performed locally at the traffic access point.

Effective bandwidth is considered by bibliography as a measure of resource usage, which adequately represents the trade-off between sources of different types, taking proper account of their statistical characteristics and QoS requirements. Thus, the effective bandwidth of a flow can be considered a quantity that represents the 'expected capacity' that a customer buys when signing an SLA. In Ref. [11] two compatible approaches for charging flows obeying to traffic contracts (or SLAs) according to their effective bandwidth are presented:

- Charging in a linear function of time and volume, based on expected mean rate
- Charging according to an (upper) estimation of the flows' actual effective bandwidth, based on expected peak rate calculated by shaping/policing parameters

In Ref. [12], a framework where customers respond to changes in price signaled by the network, by dynamically adjusting network resource usage, so as to maximize perceived utility subject to customer budget and QoS constraints, is presented. More specifically, the authors define a cost function with a number of components and are proposing that the customers define quantitatively through a utility function the perceived monetary value of their transmission with certain transmission parameters (sending rate and QoS). The goal is then to maximize the surplus between this utility function and the cost of obtaining a service (calculated according to the components of the cost function), without exceeding minimum and maximum QoS requirements and, of course, their budget.

In Ref. [13], the authors are using game theoretic concepts to approach the issue of pricing in networks offering different priorities. The main goal is to specify the ranges in which the price for each priority class of traffic can be set in order to obtain a so-called 'Nash equilibrium'. The Nash equilibrium is a desired situation in the sense that having reached it, no customer can further increase his surplus or utility by changing his choice of priority class (or classes) to serve his traffic (or his strategy, to be consistent with the game theoretic terminology). However, a 'Nash equilibrium' alone will not be desirable unless it is also efficient, or 'Pareto optimal' meaning that there is no other combination that one customer will prefer and other customers will be indifferent. It is proved that in a two-customer system with Poisson arrivals to the queue, there is a unique Nash equilibrium that is Pareto optimal and maximizes revenue provided that the difference between the high-priority traffic price and the low priority traffic price is between a lower and an upper bound.

In Ref. [14], the same authors are proceeding with a scheme that allocates bandwidth to customers so that it is available for them only if they use it. Based on game theory, they claim that a pricing model based on three factors

(amount of allocated bandwidth, amount of utilised bandwidth and fixed call set-up charge) can lead to Nash equilibrium. The calculation of the Nash equilibrium state (the calculation of the bandwidth allocation values for all customers in the NE state) is modelled as a set of constrained non-linear interdependent equations. The authors claim that by using the pricing model proposed, the customers will be encouraged to reveal their real needs for resources and prevented from resource misuse so that the equilibrium will be achieved.

In Ref. [15], the authors are suggesting a single capacity parameter representing the amount of resources allocated to a user for a specific period of time. This parameter is claimed to represent the guaranteed bandwidth provided in a virtual leased line service. It is proposed that users can dynamically change their capacity allocations depending upon their instantaneous requirements and a schema of discount rates, premium rates and penalties is proposed when a user relinquishes or exceeds his provisioned capacity. Differentiation is provided by means of differentiated trunk reservations upon which also pricing is based, however explicit tariff calculations are not provided.

In Ref. [16] it is proposed that in a best-effort network providing two classes of service, a high-priority and best-effort one, packets are blocked from entering the network in the event of congestion and only packets for which users are willing to pay a marginal congestion cost are allowed to enter. In the attempt to identify this marginal cost it is shown that its dominant component is the delay imposed by high-priority traffic to the best effort traffic. However, it occurs that the lower the utilization of the high-priority traffic queue, the higher the variance of the delay that the packets experience in the best effort queue and thus the higher the variance in the marginal cost that high-priority packets are required to pay for. Once again, it has to be pointed out that a major requirement for a pricing scheme, that of predictable charges, is difficult to achieve through a per-packet marginal cost approach.

In one of the most recent approaches, presented in Ref. [17], an architecture for market management of differentiated services in Internet environments is proposed. The related project has developed middleware over a set of proposed mechanisms that allows service providers to implement different service models for pricing differentiated services. The business models supported include:

- A dynamic, user-oriented, self-admission control model that allows the user to select the quality he wishes to receive based on announced by the provider prices per priority level and his utility
- A dynamic model that uses Explicit Congestion Notification (ECN) marks to signal the level of congestion in the network to an ‘elastic’ agent at the user endpoint, which adapts traffic flow taking into consideration the price set for each marked packet
- A long-time-scale pricing scenario, based on SLAs and thus correct estimation of customer requirements, with flat charges of SLAs and provisioning for a feedback mechanism and the liberality for deviations on short-time scales

Our approach incorporates some of the characteristics of the first and third model as outlined here, in an effort to include the end-user and the network provider in the price determination process, balance between dynamic and long-term pricing and come up with predictable charges.

In an effort to summarize the different approaches to QoS pricing, one alternative is to price based on a-priori estimation of resources’ usage according to the contract or SLA signed between a customer and a service provider. The traffic profile can be directly or indirectly used to provide some indication of the resources needed by the customer and the service provider is then able to charge service provision according to the anticipated usage of resources allocated to the specific service class. This approach determines charges on a coarse granularity (per traffic profile) and, therefore, is more likely to allow for gaps between what is paid for and what is actually used. Customers might under-utilize or over-utilize resources in comparison to the contracted traffic profiles and impose negative externalities to legitimate customers due to over-utilization of resources without being ‘punished’ for this.

Another set of approaches is based on dynamic or a-posteriori pricing of differentiated services provision. In these cases, a unit of consumption is determined (e.g. on a packet level, on a flow level, on an aggregate level, etc.) and a price per consumption unit per service class is announced according to quality guarantees provided by the service class. Such approaches require careful mapping of the value of a unit of consumption to a price and a dense monitoring infrastructure in order to adjust per service class prices depending on usage and quality obtained by each service class. It is obvious that such schemes are closer to the traditional usage-based approach and require the set-up of an appropriate infrastructure, processing overhead and storage of monitoring data.

We propose a pricing scheme that operates over reasonable service provisioning intervals. We believe that prices in the DiffServ framework should initially play the role of the mediator between the customer and the service provider. As such, they should initially drive the customer to a rational selection of a traffic profile to be included in the SLA signed between him and the service provider for service by a specific service class. This selection should be based on prices (per traffic contract within the service class) announced by the provider for each upcoming interval. The selection of the traffic profile will then result in an a-priori indication of the costs that the customer will later be required to pay.

4. The architectural framework

The case that will be further investigated in this work is that of pricing a high-priority, low latency QoS service for the customers of a transport network. Such services are provided under different names in DiffServ-enabled WANs worldwide and are built according to the Expedited Forwarding Per-Hop-Behavior (EF PHB) of the DiffServ framework ([3]). The service will be referred to as EF-based service from now on and the traffic served by the EF-based service will be assumed to belong to the EF class of traffic. For a detailed specification and analysis of the EF-based service, the reader can refer to Ref. [18]. Reliable transmission of data with the least possible end-to-end delay, almost zero packet loss and the minimum possible variation between the end-to-end delay experienced by different packets (jitter) are the most crucial factors from the customer's point of view.

In an EF-based service, the provisioning of bandwidth is taken for granted and the focus shifts to the transmission quality obtained. The provision of such a service by a transport network provider has an analogy to the best-effort service provision: instead of bandwidth, the resource under contention is buffer space. The negative externalities imposed by congestion in best-effort service provisioning have their analogy to the negative externalities imposed by delay due to buffer occupancy and packets' waiting time in an EF-based service.

The most widely used traffic descriptor or traffic profile included in SLAs for provisioning of EF-based services is that of a token bucket (r,b) that imposes conformance to an average rate (r) and a maximum burst (b) to the traffic flow or aggregate to which it applies. This type of traffic profile will be used for the analysis in our proposed pricing scheme. The scheme must lead each customer to select the amount of buffer space to purchase from the provider in such a way that the bursts of his traffic are accommodated and that the customer does not have to shape his traffic more than he can endure. At the same time, the negative externalities imposed to the community of users by that amount of buffer space have to be compensated for, thus included in the price for this buffer space. Contrary to the existent approaches, what we are proposing is a distinction between the costs imposed to customers for the rate of their token bucket traffic profiles and the costs imposed to customers for the depth of their token bucket traffic profiles. This approach provides to the customers the incentives to provide EF traffic aggregates as well shaped as possible to the network provider. It also provides them with the incentive to provide the most accurate description of their detailed traffic profile (in terms of average rate and burstiness), rather than just an accurate description of their expected mean rate as proposed in Ref. [11].

For the purposes of our detailed analysis, the downstream domains or customers are modelled as sources of EF traffic. Between each of the customers and the Transport Domain (TD) there exists an SLA that specifies the characteristics (traffic envelope) of the marked as EF traffic injected by each customer into the TD and the specific bounded end-to-end delay guarantee (D) provided by the TD itself. EF aggregates are considered legitimate after being policed each one by its own token bucket (r,b) policer.

5. Pricing the SLAs

Over-provisioning and careful dimensioning can be intuitively assumed to guarantee the required transmission rate and low end-to-end delay for the EF traffic aggregates traversing a TD. In such a situation, the utility function of customers is no longer solely dependent upon the volume of traffic being transmitted and the congestion experienced. It depends upon the equivalent capacity that each aggregate perceives and the quality metrics guaranteed (end-to-end delay, jitter and packet loss). We assume that over-provisioning ensures that no EF packets are dropped due to overflow in router queues along the TD and that EF aggregates obtain a throughput, which is at least equal to their token bucket profiles' rate (r). Thus, the utility function of customers depends upon the rate (r) and burstiness allowance (b) purchased from the provider as well as the end-to-end delay (D) that the packets of each aggregate experience. We make the simplifying assumption that the utility perceived by the jitter guarantee is included in the delay factor. If we depict by $p_{\mathrm{EF}}(S_i)$ the costs that a customer has to pay for purchasing an SLA for EF-based service including the $S_i=(r_i,b_i)$ token bucket profile, then the objective of a pricing mechanism should be (apart from reimbursement of the provider's expenses for providing the EF-based service) that of maximizing

$$U_{\mathrm{EF}}(S_i) - c_{\mathrm{EF}}(D) - p_{\mathrm{EF}}(S_i) \tag{2}$$

for each customer K_i, where

- $U_{\mathrm{EF}}(S_i)$ the utility perceived by a customer K_i serviced by the TD according to SLA S_i
- $C_{\mathrm{EF}}(D)$ the cost (negative utility) of end-to-end delay D for customer K_i
- $p_{\mathrm{EF}}(S_i)$ the price to be paid by each customer signed with the SLA S_i and receiving EF-based treatment. For ensuring reimbursement of costs for provisioning of the EF-based class ($\mathrm{cost}_{\mathrm{EF}}$) the following should apply:

$$\sum_{i\in\{\text{set of SLAs offerred}\}} \mathrm{p}_{\mathrm{EF}}(S_i) \geq \mathrm{cost}_{\mathrm{EF}} \tag{3}$$

The pricing mechanism proposed should aim at restricting the customer's demands in such a way that, at the equilibrium, each customer's revenue calculated by Eq. (2) is maximized, without inequality Eq. (3) being violated. We claim that pricing of EF services should place costs on the traffic profiles on which resources' provision is made. This is mainly due to the fact that what really matters is how

many resources a customer has occupied in order to ensure a certain level of quality for his traffic, instead of how much of these resources the customer is actually using. The customer's actual needs for QoS should lead to the negotiation with the TD provider of an appropriate traffic contract, guaranteed D and pricing of the service provided. The provider should engineer its infrastructure so that once traffic contracts are signed with all customers, the provisioned transmission rates and the common to all customers end-to-end delay bound guarantee is ensured.

5.1. Provisioning and charging for transmission rate

According to the approach of Ref. [19], the TD provider can guarantee a worst-case end-to-end delay bound to all its customers, provided that the ratio ρ of the TD links' capacity to be devoted to the EF traffic injected to the TD is bounded as follows:

$$\rho < \min_l \frac{P_l}{(P_l - C_l)(h-1) + C_l} \tag{4}$$

where C_l is the capacity of each link of the TD, assumed constant $\forall l$, $l \in$ TD and equal to C and P_l is the maximum rate with which the EF traffic aggregate (emerging from the merging from EF aggregates upstream) is injected at each TD node, depending also on the fan factor for the specific node. Also h is the maximum number of hops within the TD that a customer's EF traffic can traverse. After the selection of the value for ρ based on Eq. (4), the router schedulers in the TD can be configured so as to ensure that only this desired ratio of the TD links' capacity is devoted to EF traffic. However, it is recommended by Ref. [19] that the TD provider chooses ρ such that it is much less than the quantity of the right part of Eq. (4). This is also referred to as the prerequisite of over-provisioning for supporting an EF class in related research. It requires that only a percentage of the capacity reserved for EF traffic ($C_{\mathrm{EF}} = \rho C$) can be subscribed for. Thus, if N is the set of customer aggregates routed through a node, for every node n of the TD it must hold that:

$$\sum_{i \in N} r_i < \rho C \Rightarrow \sum_{i \in N} r_i = a\rho C \Rightarrow a = \frac{\sum_{i \in N} r_i}{C_{\mathrm{EF}}}, \quad \alpha < 1 \tag{5}$$

Assuming that each customer will ask for the highest r_i possible, the network administrator has to turn up with a set of acceptable r_i values and corresponding prices for the customers so that Eqs. (4) and (5) are respected, thus distribute the available EF capacity in the most efficient way. In this process, ρ is constant for a certain topology and traffic engineering and the TD provider can only vary the selection of a value for a according to the total EF capacity he wishes to sell to his customers. Recommendations from related work ([20]) lead to the selection of small values for a, in the range of $\{0.05,\ldots,0.2\}$.

For the rest of the analysis, we will assume that the TD provider selects a value for ρ so that Eq. (4) holds and a value for a that is not negotiated with TD's customers whatsoever. Maintaining a constant value for ρ for the rest of our analysis helps in isolating the charging for EF traffic methodology from its side effects on the rest of the traffic that crosses TD. Non-EF traffic will thus be served by $(1-\rho)C$ capacity on each link of TD and will not be affected by any kind of distribution or re-balancing of the resources devoted to EF traffic due to the charging scheme proposed.

After the selection of a, the TD provider has to distribute a total of

$$r_{\mathrm{tot}} = \sum_{i \in N} r_i = aC_{\mathrm{EF}} \tag{6}$$

EF capacity among his customers.

Under the model that this work addresses, all of TD's customers are TDs themselves, the EF traffic aggregates of which have emerged as the result of aggregation of hundreds or thousands of micro-flows. The TD provider is suggested to distribute r_{tot} to his customers during the pricing mechanism's initialisation phase in a fair way according to

$$r_i = \frac{C^i_{\mathrm{access}}}{\sum_i C^i_{\mathrm{access}}} r_{\mathrm{tot}} \tag{7}$$

In this way, each customer K_i receives a share of the EF capacity available according to the capacity (C^i_{access}) of his access link to TD. In later, re-negotiation phases the TD provider might update the distribution of r_{tot} to each customer according to a ratio ρ_i that might differ from their access link ratios so that $r_i = \rho_i \times r_{\mathrm{tot}}$, while Eq. (6) is always respected. For the initialisation phase however we suggest $\rho_i = C^i_{\mathrm{access}} / \sum_i C^i_{\mathrm{access}}$.

After the pricing scheme's initialisation phase, we propose re-negotiation phases of all the contracted traffic profiles simultaneously over long-term intervals. During re-negotiations, each customer is able to base his new traffic profile's r value selection for the next period on statistical data for the utilization of the rate value allocated to him in the elapsed period. This data can directly be retrieved by the statistics of the token bucket policer of the customer's aggregate in the ingress of the TD, so that no per-packet accounting is required and overhead is thus avoided. More details on this are provided in the following sections. Long-term re-negotiation phases will allow customers to evaluate their needs for resource provisioning based on solid, single-dimensional measurements and request the corresponding resources from the TD provider. This model will be shown to demonstrate fluctuations in the beginning, leading to more stable distribution of resources after a number of re-negotiations. Fluctuations are also possible when a new customer requires EF services from the TD provider.

In terms of charging the provided EF rates for each phase, the TD provider is proposed to fairly spread the cost

of over-provisioning that EF traffic requires among the EF class customers. Thus, instead of charging each customer just for the EF contracted capacity r_i provided to him, the provider has to calculate EF capacity unit price according to

$$p_j^{\text{unit}} = \rho_j \times \{\text{cost of capacity } C_{\text{EF}} \text{ in the TD}\} \quad (8)$$

so that the unit price occurs as if the customer is using $\rho_j \times C_{\text{EF}}$ instead of the actual $\rho_j \times a \times C_{\text{EF}}$ capacity for his EF traffic. The total cost for providing an EF average rate of r_i to a customer is then

$$P_j = p_j^{\text{unit}} \times r_j = \rho_j \times \{\text{cost of capacity } C_{\text{EF}} \text{ in the TD}\} \times r_j \quad (9)$$

5.2. Provisioning and charging for burstiness

After the selection of ρ and a, the provisioning of resources for servicing EF traffic throughout TD is possible, by configuring all nodes' schedulers to provide a service rate of $C_{\text{EF}} = \rho C$ to the EF traffic on all TD links. It can be then shown ([19]) that the end-to-end delay D is bounded by a function of the same factors as ρ, thus topology and capacity configuration related factors, as well as the total buffering space b_{tot} reserved at each TD router for EF traffic and the over-provisioning factor a itself. For more details, refer also to Ref. [21].

The TD provider can thus calculate his available b_{tot} for a certain D guaranteed to its customers. It is apparent that according to the current TD's topology and capacity there is a limited amount of total buffer space at each router that can be distributed to customers. The customers must thus be prompted by the bucket depth charging policy of the TD provider to restrain themselves from selecting large values for b_i by the fact that this will penalize themselves and others in terms of the delay perceived by their packets. Also the TD provider has to distribute b_{tot} among his customers so that

$$\sum_{i \in N} b_i \leq b_{\text{tot}} \quad (10)$$

where N is the set of all customers. The latter holds because in the worst-case scenario where all customers' aggregates at some point are routed through one node of the TD's core and all aggregates' bursts coincide inside the buffer space of this node, the node must have enough buffer space to place packets, so that no packets are dropped.

It is at this point that the 'smart market' approach already presented earlier applies. As already mentioned, in the case of EF-based services, resources (i.e. buffer space) must be distributed to those who value them most and distribution has a direct impact on all customers (the end-to-end delay guaranteed by the TD). The TD provider announces the end-to-end delay that can be guaranteed to customers and the customers place bids on the available buffer space (b_{tot}) in order to obtain a share. The clearing price for a buffer position (P_b) is set at the point where the sum of demands for buffer space, starting to add from the higher-bids' demands, reaches the amount of available buffer space b_{tot} (see Fig. 2).

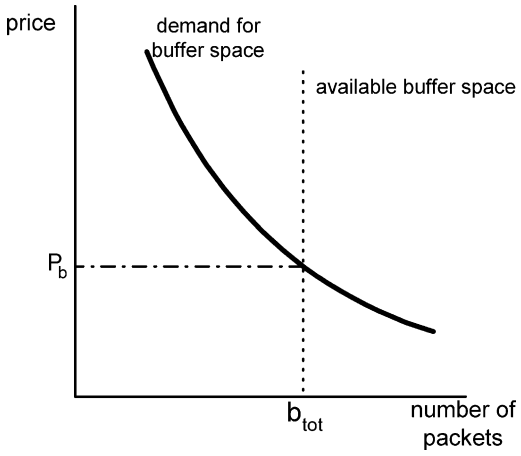

Fig. 2. The 'smart market' for a buffer space market.

In this way, customers who have valuated more a buffer position taking into consideration the end-to-end delay guarantee D, receive a larger portion of the available buffer space, or in other words obtain SLAs with larger b_i values than customers who placed lower bids. Of course each customer will be notified of the cost he will have to pay for buffer space when signing a token bucket (r_i,b_i) SLA as equal to

$$P_{b_i} = \frac{b_i}{b_{\text{tot}}} P_b \quad (11)$$

where P_b is the cost of buffering EF traffic in the TD.

In a real-life scenario, it is envisaged that the TD provider will distribute the available buffer space b_{tot} during the initialisation phase according to intuitive bids placed by customers, since no real-use data will be available. At the moment of re-negotiations, instead of speculating for the future, the customers are able to place bids on the available buffer space based on the statistics of the token bucket policer (r_i,b_i) of their aggregates for the elapsed period. Again, fluctuations will be observed in the first phases or when a new customer will require EF services from the TD provider. However, since the 'smart market' and bidding are proven to successfully integrate externalities in goods' provision costs, it is envisaged that in equilibrium, the buffer space will be distributed to those who value it most and are willing to compensate for the delay their bursts might cause to others.

6. Proposed pricing mechanism

Based on the theoretical analysis already made, it is proposed that the following algorithm is used for the provision and pricing of an EF-based service over a TD:

Step 1. Each customer agrees that his EF aggregate will be policed by a (r_i,b_i) token bucket policer as the traffic enters the TD.

Step 2. Based on his local policy for EF provisioning, the provider determines a (the provisioning factor) for EF traffic on the TD topology, so that Eqs. (4) and (5) hold.

For a TD topology composed of links with 2.5 Gbps capacity, a maximum fan-factor equal to 3 and a diameter h (maximum number of hops for a packet) as shown in the first column, the ratio bound ρ for providing an end-to-end delay bound to 20 customers attached with 155 Mbps links is provided in Table 1.

The value of C_{EF} that can be supported by the TD is shown in the third column of Table 1. According to the over-provisioning requirement, the TD provider has to also select a in the framework of Eq. (5)

Step 3. Initialisation phase for EF rate provisioning. The TD provider calculates

$$r_{tot} = \min_{l} a \times C_{EF} \tag{12}$$

over all the links l of the TD topology and then distributes SLA token bucket rates to all customers according to Eq. (7). From Eq. (9), the cost for providing an EF average rate of r_i to each customer is calculated and the customers are then informed in advance about one part of the cost they will be asked to pay for the upcoming operation phase.

Step 4. Initialisation phase for EF burstiness provisioning. According to the end-to-end delay demands of the applications supported (e.g. VoIP requires up to 150 ms of one way delay, in case VoIP packets traverse two or three TDs in a raw, each TD cannot contribute more than 50 or 75 ms to the delay perceived by packets) and the advertised quality that the TD provider wishes to sell to all EF customers, the TD provider determines the end-to-end delay guarantee provided (D) and then calculates the buffer space b_{tot} that can be distributed among all EF customers.

In the case of a TD with $a=0.2$, maximum number of hops equal to 6, a topology fan factor of 4, MTU = 4700 bytes and $C=622$ Mbps the bound on end-to-end delay provided to all customers for different b_{tot} values is provided in Table 2.

After b_{tot} is determined, SLA token bucket depths to all customers can be distributed, for example equally, according to

Table 1
Provisioning factor and allowed total of EF capacity for a series of h values

h	ρ	C_{EF}
3	0.6	1.5 Gbps
4	0.43	1.075 Gbps
5	0.33	825 Mbps
6	0.27	675 Mbps
7	0.23	575 Gbps

Table 2
Bounds on end-to-end guaranteed delay in a transport domain with a maximum EF space of b_{tot} for any node

b_{tot} (packets)	10	50	100	150	200	300
D (ms)	3.77	18.85	37.7	56.55	75.4	113

$$b_i = \left\lfloor \frac{b_{tot}}{k} \right\rfloor \tag{13}$$

where k is the total number of EF customers. As a result in the initialisation phase, each customer is asked to pay for allowed burstiness an amount of

$$P_{b_i} = \frac{b_i}{b_{tot}} P_b \tag{14}$$

Step 5. Operation phase. The service is initialised and provided for a number of days n_d. During the operation phase, at the interface of the edge router where each customer's EF traffic aggregate is policed according to the token bucket (r_i,b_i), the following statistics are maintained at regular intervals Δt:

$$r^i_{average} = r_i + \frac{\text{number of packets dropped by the } (r_i,b_i) \text{ token bucket}}{\Delta t} \tag{15}$$

$$b^i_{current} = \text{current burst size} \tag{16}$$

It is important to note at this point that these statistics can be collected without computational complexity and scalability problems, since only dropped packets are counted in the case of Eq. (15) and the value of a counter is recorded in the case of Eq. (16).

Step 6. SLA re-negotiation and prices' adjustment phase. After an operation phase is terminated, the statistics collected must be evaluated and the SLAs preserved or adjusted. Each customer is presented with the vectors $\{r_{average}\}$, $\{b_{current}\}$ for the previous operating period and, ideally, a graphic representation of the values of the collected statistics (see Fig. 3).

Based on the data collected from the previous operating period, each customer is applying for a new token bucket policer (r'_i, b'_i). The values of r'_i, b'_i can emerge from the $\{r_{average}\}$, $\{b_{current}\}$ vectors in a number of ways, e.g. the mean or median or upper values of the measured statistics can be used. A negotiation phase is here required and the TD provider can apply different policies in order to reach agreements with all its clients, e.g. first-come-first-serve, or normalizing demand according to available capacity determined in Step 3, providing each customer with a token bucket rate equal to

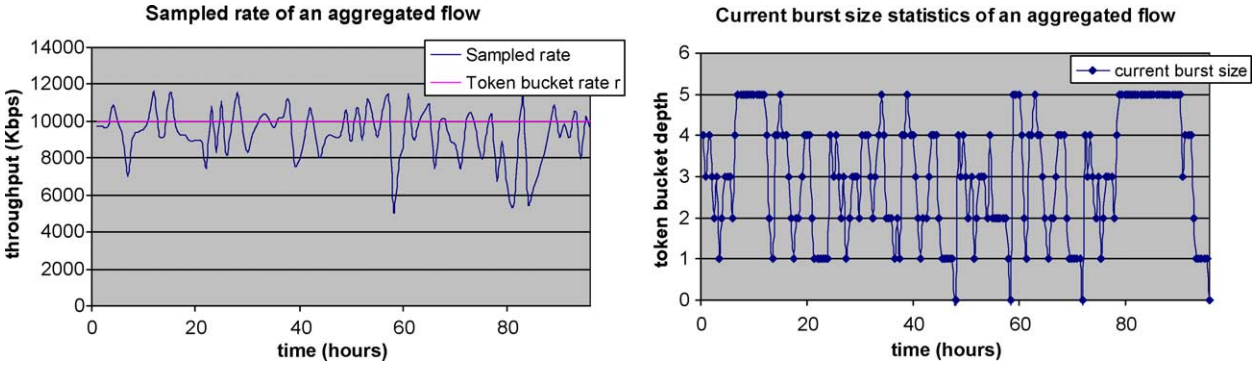

Fig. 3. Sampled data for the traffic aggregate of a customer during an operation period.

$$r_i^{t+1} = \frac{r_i'}{\sum_i r_i'} r_{\text{tot}} \tag{17}$$

for the upcoming operation phase.

Customers are also placing bids (bid_i) for the available buffer space b_{tot} in the upcoming operation phase, taking into consideration the sampled data of the previous operating period and the delay guarantee D provided by the TD. Each bid_i is in the form of a vector

$$\hat{v} = (b_j, p_j^b) \tag{18}$$

where b_j is the number of buffer spaces requested at price p_j^b per buffer space. Thus, each customer may request a series of (b,p^b) tuples. The TD provider is evaluating all bids in the order of p^b offers, starting from the highest offer and provides all token bucket positions for which the following holds.

$$\sum_j b_j \leq b_{\text{tot}} \tag{19}$$

In this way the token bucket b_i^{t+1} values for the next operating period are determined for all customers. The next operation phase can be initiated.

Steps 5 and 6 are iterated continuously during the service's operation.

7. Experimental evaluation

For the evaluation of the proposed methodology and algorithm, an experimental set-up investigating the convergence of the iterative procedure of SLAs negotiation and pricing was implemented. The approach followed is rather simplistic, however it demonstrates the effectiveness of the pricing methodology proposed and how it provides to the customers the incentives to better approximate their true traffic profiles and charged in a fair and exact manner.

The simple case of a TD composing of single backbone link was adopted. Three main customers inject aggregated EF traffic to the same PoP of the TD. Each customer's EF aggregate is composed of 4,2 and 3 MPEG video flows for customers C_1,C_2 and C_3 correspondingly. Each video flow is rather bursty with an average rate of 1.3 Mbps, packet size 200 bytes and an average burst size of 1700 bytes. Fifty-five Mbps are provisioned for EF traffic on the TD backbone link and an end-to-end delay of 19 ms is promised by the TD provider for a value of a equal to 20.5% and $b_{\text{tot}} = 30$. Background traffic was also used to load the TD backbone link. For the case presented below, we assume that the cost (negative utility) of end-to-end delay D for all three customers ($C_{\text{EF}}(D)$ in Eq. (2)) is represented by the same convex function.

In Table 3 the SLA traffic descriptors that occurred during the re-negotiation phases of the experiment based on the statistical data of Eqs. (15) and (16) in the form of $\{r_i(\text{Mbps}),b_i(\text{packets})\}$ are presented. Due to a relatively high P_b value in Eq. (11) set by the TD, the customers were led to reduce the burstiness metric b_i in their traffic contracts during the re-negotiation phases.

It is quite important to notice how, with small fluctuations, each customer updated his traffic contract throughout the iterations so as to describe more tightly his EF aggregate and shifted requested resources from the burstiness parameter b_i to the average rate r_i. Of course, the end-to-end delay bound of 19 ms was never violated during

Table 3
Traffic descriptors for all three customers during the re-negotiation phases

	Initialisation phase	Re-negotiation periods						
		1st	2nd	3rd	4th	5th	6th	7th
C_1	(4, 30)	(4.2, 20)	(4.3, 17)	(4.4, 14)	(4.45, 12)	(4.48, 11)	(4.5, 9)	(4.51, 8)
C_2	(4, 5)	(2.1, 10)	(2.15, 9)	(2.17, 8)	(2.18, 9)	(2.19, 8)	(2.2, 6)	(2.2, 6)
C_3	(5, 30)	(3.3, 20)	(3.4, 22)	(3.4, 19)	(3.45, 13)	(3.5, 11)	(3.52, 9)	(3.53, 8)

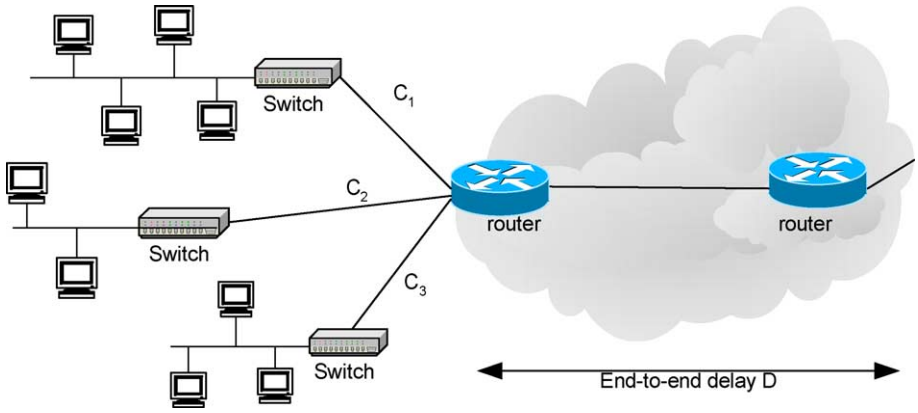

Fig. 4. Experimental topology.

all phases, since it consisted an upper bound for our case (Fig. 4).

Finally in Fig. 5, the normalized charges imposed to C_1 throughout the consecutive re-negotiation periods are depicted in a graph. One can observe how the statistical data provided to the EF service customer and the incentive-based pricing scheme proposed leads to a tighter traffic descriptor, which is also economically beneficial for the customer. From the TD provider's point of view a more efficient allocation of resources is achieved. The decrease in revenue for the TD provider is compensated by new customers that can be accommodated. By providing incentives to existent customers to reveal their true traffic profiles through some iterations, the provider can become aware of the true utilization of resources in his backbone and is then able to accommodate new customers without compromising quality.

8. Future work

The work presented here is a first step towards the direction of incorporating incentive-compatible, DiffServ-compliant and predictable pricing models in the provisioning of differentiated services over a transport network. The methodology presented concerns the case of reimbursement of costs for the provider. Our future work will focus on the case of the provider's profit optimisation and on further investigating the customers' utility function in Eq. (1), while adhering to the recommendation of Ref. [1], that the utility functions are specified so as to be valid for a duration longer than a connection's duration.

An interesting issue for investigation within the proposed pricing model is that of fluctuation of prices per unit of EF bandwidth and buffer space during consequent re-negotiation periods. This alternative will be provided as a tool to the transport network provider in cases where more control over the EF traffic injected to the network is required. Indicatively, a framework will have to be defined and tested in a way that for example prices are increased after an operation phase during which the majority of customers' traffic statistics supersede the registered traffic profiles. Such an enhancement would reinforce the incentive compatibility of the proposed mechanism. The details of the price adjustment mechanism will have to be studied and evaluated.

We also aim at dealing with the case of pricing services based on the Assured Forwarding PHB (AF PHB), as defined within the DiffServ framework. According to the AF PHB specification, there are no quantifiable timing requirements (delay or jitter) associated with the forwarding of AF packets. Thus, the AF PHB allows for short-term congestion (queuing) and minimizes long-term congestion (dropping). Priority is of relative nature and the packets of an AF class that are not within the specified rate are marked as belonging to the lower AF class. Therefore, auctioning mechanisms could be of particular use for the pricing of services built upon the DiffServ AF PHB, assuming the simple case where an AF PHB-based class is not allowed to use spare resources of best effort traffic or other PHBs. Bidding can be done either among flows of the same AF class but different precedence class, or among members of different AF classes. These issues will be further investigated in our future work, in order to determine whether the research work on auctioning mechanisms for allocating

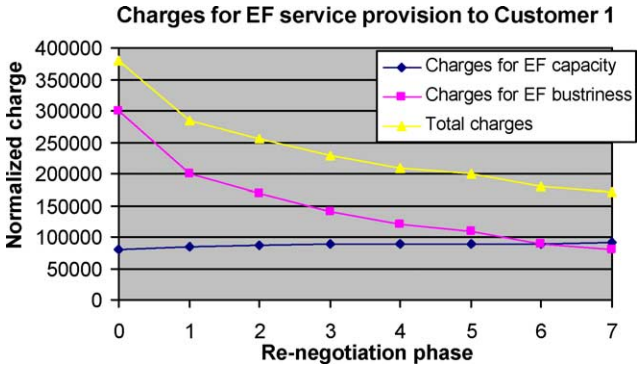

Fig. 5. Evolution of the charges paid by Customer 1 during the re-negotiation phases.

prioritised resources could be exploited for pricing AF PHB-based service classes.

9. Conclusions

The challenge of DiffServ-based services' pricing is to adhere to the DiffServ framework dominant characteristics: simplicity, operation over the existing IP-based infrastructure, shifting of processing load to network edges, separation of pricing mechanisms from the pricing strategy (maximisation of the social welfare, fairness, maximisation of supplier's revenue, etc.) a provider might choose to apply. The pricing mechanism proposed in this work is based on traffic profiles that the customers negotiate with a TD provider and concludes on prices announced to customers prior to the service provision interval. In this way, it is more likely to be accepted by customers, who are only faced with predictable costs.

In the case of EF-based services which are under consideration here, the traffic profiles of the customers comprise a representation of the utility that each customer finds on the service. At the same time, traffic profiles are used by the TD provider in order to dimension the EF-based service and allocate the resources used by it. However, the utility for the user is not only represented in the pricing mechanism through the quantity of resources allocated. In order to reflect different utilities for the same amount of resources allocated but different quality guarantees provided, quality guarantees are included in the utility function. More specifically, in the proposed pricing mechanism, the customers are invited to negotiate their traffic profiles and charging on the basis of the quality guarantees announced by the TD provider.

In an initials effort to apply in practice and evaluate the proposed pricing mechanism, it has been shown how adjusting the pricing coefficients based on customers' service usage statistics over long-term intervals leads to a tighter description of the traffic profiles. This is not only economically beneficial for the customer but also for the service provider, since it allows for more customers being provided with EF-based services without compromising the qualitative guarantees offered.

Thus, the proposed pricing mechanism uses the traffic profiles of customers as the intermediate between each customer and the provider. In this way, it reflects both the customers' revenue from the EF-based service provided and the costs for the service provisioning that the TD provider undertakes. Moreover, the proposed pricing mechanism takes into consideration the in-elasticity in demand for transmission rate that applies in the case of the customers of a backbone TD and efficiently allocates the available buffer space to those customers for which accommodation of their bursts is more valuable. Finally, the proposed mechanism provides indications of the quality that will be provided to customers (in terms of end-to-end delay), in order to assist them in the qualitative valuation of the service they will receive and express accurately their needs for resources.

References

[1] M. Falkner, M. Devetsikiotis, I. Lambadaris, An overview of pricing concepts for broadband IP networks, IEEE Communications Surveys and Tutorials 3 (2) (2000).
[2] A. Odlyzko, Internet Pricing and the History of Communications Computer Networks 36, Elsevier, Amsterdam, 2001, pp. 493–517.
[3] S. Blake, et al., An Architecture for Differentiated Services, RFC2475, 1998.
[4] T. Henderson, J. Crowcroft, S. Bhatti, Congestion pricing: paying your way in communication networks, IEEE Internet Computing September–October (2001) 85–89.
[5] L.A. DaSilva, Pricing for QoS-enabled networks: a survey, IEEE Communications Surveys and Tutorials 3 (2) (2000).
[6] R. Cocchi, D. Estrin, S. Shenker, L. Zhang, A study of priority pricing in multiple service class networks, in: Proceeding of SIGCOMM'91, 1991.
[7] J. MacKie-Mason, H. Varian, Pricing the internet in: B. Kahin, J. Keller (Eds.), Public Access to the Internet, Prentice Hall, New Jersey, 1995.
[8] D.D. Clark, W. Fang, Explicit Allocation of Best Effort Packet Delivery Service, Technical report, MIT, Lab for Computer Science, 1997.
[9] D.D. Clark, A model for cost allocation and pricing in the Internet, in: L.W. McKnight, J.P. Bailey (Eds.), Internet Economics, MIT Press, Cambridge, MA, 1996.
[10] S. Shenker, D. Clark, D. Estrin, S. Herzog, Pricing in computer networks: reshaping the research agenda, ACM Computer Communication Review 26 (2) (1996) 19–43.
[11] C. Courcoubetis, F.P. Kelly, R. Weber, Measurement-based usage charges in communication networks, Operations Research 48 (4) (2000) 535–548.
[12] X. Wang, H. Schulzrinne, Performance study of congestion price based adaptive service, in: Proceeding of NOSSDAV 2000, Network and Operating Systems Support for Digital Audio and Video, Chapel Hill, North Carolina 2000.
[13] L. DaSilva, D. Petr, N. Akar, Equilibrium pricing in multi-service priority based networks, in: Proceeding of IEEE GLOBECOM'97 1997.
[14] L.A. DaSilva, D.W. Petr, N. Akar, Static pricing and quality of service in multiple service networks, in: Proceeding of Fifth International Conference on Computer Science and Informatics, Atlantic City, NJ, vol. 1 2000 pp. 355–358.
[15] M. Singh Dang, R. Garg, R. Singh Randhawa, H. Saran, A SLA framework for QoS provisioning and dynamic capacity allocation in: Proceeding of Tenth International Workshop on Quality of Service (IWQoS Miami Beach) (2002).
[16] L. Jean Camp, C. Gideon, Certainty in bandwidth or price, in: Proceeding of the 29th Research Conference on Communication, Information and Internet Policy, Washington, D.C., 2000.
[17] B. Briscoe, V. Darlagiannis, O. Heckman, H. Oliver, V. Siris, D. Songhurst, B. Stiller, A market managed multi-service internet (M3I), Computer Communications 26 (4) (2003) 404–414.
[18] C. Bouras, A. Sevasti, Analytical approach and verification of a DiffServ-based priority service, in: Proceeding of Sixth IEEE International Conference on High Speed Networks and Multimedia Communications-HSNMC, Estoril, Portugal, 2003 pp. 11–20.
[19] A. Charny, J.-Y. Le Boudec, Delay bounds in a network with aggregate scheduling, in: Proceeding of First International Workshop on Quality of future Internet Services (QofIS), Germany, 2000.

[20] Y. Le Boudec, P. Thiran, Network Calculus: A Theory of Deterministic Queuing Systems for the Internet, LNCS 2050, Springer, Berlin, 2001.
[21] C. Bouras, A. Sevasti, Pricing priority services over DiffServ-enabled transport networks, in: IFIP INTERWORKING 2002 Conference, Sixth International Symposium on Communications Interworking, Fremantle, Perth, Australia, 2002 pp. 25–37.

Christos Bouras obtained his Diploma and PhD from the Computer Science and Engineering Department of Patras University (Greece). He is currently an Associate Professor in the above department. Also he is a scientific advisor of Research Unit 6 in Research Academic Computer Technology Institute (CTI), Patras, Greece. His research interests include Analysis of Performance of Networking and Computer Systems, Computer Networks and Protocols, Telematics and New Services, QoS and Pricing for Networks and Services, e-learning, Networked Virtual Environments and WWW Issues. He has extended professional experience in Design and Analysis of Networks, Protocols, Telematics and New Services. He has published 150 papers in various well-known refereed conferences and journals. He is a co-author of five books in Greek. He has been a PC member and referee in various international journals and conferences. He has participated in R&D projects such as RACE, ESPRIT, TELEMATICS, EDUCATIONAL MULTIMEDIA, ISPO, EMPLOYMENT, ADAPT, STRIDE, EUROFORM, IST, GROWTH and others. Also he is member of, experts in the Greek Research and Technology Network (GRNET), Advisory Committee Member to the World Wide Web Consortium (W3C), Task Force for Broadband Access in Greece, ACM, IEEE, EDEN, AACE and New York Academy of Sciences.

Afrodite Sevasti obtained her Diploma from the Computer Engineering and Informatics Department of Patras University in Greece. She holds a Master of Science in Computer Science and Engineering from the same Department, where she is also a PhD candidate. She also holds a Master of Science in Information Networking from the Information Networking Institute of Carnegie Mellon University. She has worked as an R&D Computer Engineer at the RA Computer Technology Institute (Greece) and she is currently with the Greek Research and Technology Network (GRNET) S.A. Her main interests and expertise lie in the fields of Computer Networks, Telematics, Distributed Systems and especially in technologies and architectures of high performance networks, in traffic and network resources' management, in Managed Bandwidth Services, provisioning of Quality of Service (QoS), SLAs and pricing/billing of next generation networks. She has published 17 papers in well-known refereed conferences and journals. She has participated in several R&D projects.