# Pricing QoS over transport networks

*Christos Bouras and*
*Afrodite Sevasti*

**The authors**

**Christos Bouras** is based both at the Computer Technology Institute-CTI, and in the Department of Computer Engineering and Informatics, University of Patras, Patras, Greece.
**Afrodite Sevasti** is based both in the Department of Computer Engineering and Informatics, University of Patras, Patras, and the Greek Research and Technology Network (GRNET), Athens, Greece.

**Keywords**

Pricing, Performance measurement (quality), Transport management, Traffic

**Abstract**

In recent years, a number of alternatives for service differentiation and QoS provision have been proposed and standardized in communication networks. In the case of backbone networks the DiffServ architecture has prevailed, due to its scalability and deployment feasibility. The provisioning of differentiated services has raised the requirements for interdependent controlled resource allocation and service pricing, with particular needs for pricing mechanisms that preserve the potential and flexibility of the DiffServ framework. 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. Presents the key issues involved in the area of pricing DiffServ-based services and the research work carried out in this field, while at the same time outlining the basic principles that such a pricing infrastructure should obey with respect to the particularities that apply to the case of DiffServ services provision.

**Electronic access**

The Emerald Research Register for this journal is available at
**www.emeraldinsight.com/researchregister**

The current issue and full text archive of this journal is available at
**www.emeraldinsight.com/1066-2243.htm**

Internet Research
Volume 14 · Number 2 · 2004 · pp. 167-174

DOI 10.1108/10662240410530871

## 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 Falkner *et al.* (2000), 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 Odlyzko (2001), 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 be incompatible with 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 remarkable scalability. As such, DiffServ seems to be a promising solution for efficiently supporting the QoS demanding applications of the future.

The DiffServ framework has been designed for the provision of QoS services in large-scale networks, where the extensive aggregation of packet flows does not allow solutions for QoS provisioning on a per-flow scale. It operates on the basis of marking the packets of individual flows that belong to a certain QoS class with a single differentiated services codepoint (DSCP) value (or a group of DSCPs). In the interior of each DiffServ-enabled domain, servicing of packets by network elements is performed according to their DSCP value and not to the flow to which they belong. In other words, in the interior of a DiffServ domain, all packets belonging to the same QoS class are indistinguishable and thus receiving the same treatment. For more details on the DiffServ framework principles, the reader is directed to Blake *et al.* (1998).

DiffServ anticipates the classification of individual flows in a limited set of service classes at network edges as well as "soft" reservation of resources and special handling of packets per service class, in the core of the network. DiffServ-based service class examples include a service class that serves eligible packets with the topmost priority, a service class that ensures a minimum service rate for packets in congestion conditions, a service class that serves packets only when excess capacity is available. Allocation 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.

This work focuses on pricing schemes and methodologies that have been proposed for the case of the DiffServ framework as well as on the basic principles to which a pricing scheme operating over a DiffServ-enabled network must obey. In section 2, the general principles of the "QoS-pricing" or "DiffServ-pricing" problem are outlined (we consider the two terms as interchangeable throughout this work, since QoS provision in backbone, transport networks has recently become synonymous to DiffServ provision). In section 3 relevant research approaches that analyse, estimate, quantify and/or attempt to solve the problem of DiffServ-pricing are listed. Section 4 summarizes existent results and proposes a universal framework on which DiffServ-pricing should be based. Our future work and conclusions on this issue 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, due to their complexity, were rarely adopted in practise, 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.

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 time referred to as service level agreement (SLA)), which contains the agreed QoS parameters and a traffic descriptor or profile that the flow must obey. Congestion control is only exercised at the traffic ingress point by enforcing compliance with the corresponding traffic profile (e.g. by dropping the packets exceeding the traffic profile parameters) without any feedback mechanism. Thus, regardless of the congestion conditions in the backbone, as soon as packets enter the network they are served according to the principles of the service class to which they belong.

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. One of the most common traffic profile descriptors is that of a token bucket $(r,b)$, that imposes conformance to an average rate and maximum burst size to a traffic flow or aggregate. It is common to use the notation of $r$ as the rate and the notation of $b$ as the depth parameters of the token bucket. Such a traffic profile denotes that for any time interval $t$, the total amount of traffic for the flow or aggregate is equal to $rt+b$. An SLA usually contains, apart from a traffic profile, a description of the type of service that a flow or aggregate belonging to a certain traffic class receives (e.g. end-to-end delay guarantees for individual packets) and a number of technical parameters entailed in provisioning that service by the network (availability, etc.).

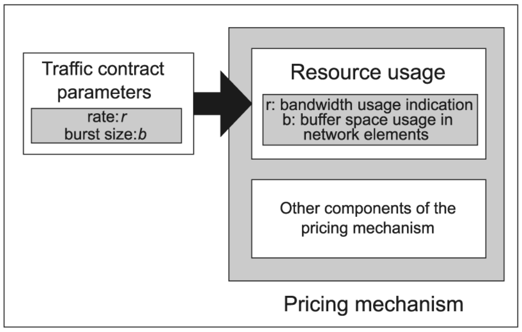
As long as its traffic descriptor is not violated, the flow is transmitted unaltered over the network equipment. However, since obeying to a traffic descriptor involves shaping and/or policing of traffic (by the network or the customer himself) according to the traffic contract, the traffic contract 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 Figure 1 for the case of the token bucket traffic descriptor). 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 consumes.

This brief analysis outlines the new roles that 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 excessive 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.

According to this analysis, in a network that offers service differentiation and QoS, the utility function of customers is no longer solely dependent on the amount of traffic being transmitted and the congestion experienced. It also depends on 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.

**Figure 1** How do traffic contract parameters affect pricing



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

$$U_{s_k}(L_i) + U(Q_{s_k}) - p_{s_k}(L_i) \qquad (1)$$

for each customer $C_j$, where:

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

It is obvious that the maximization of (1) over the sets of all customers and service classes of a network is not a trivial task, especially keeping in mind that $U_{s_k}(L_i)$ and $U(Q_{s_k})$ 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.

From the provider's point of view, resource usage from a traffic flow or aggregated flow is a reasonable basis on 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 service provider has to give 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 must be regulated by appropriate pricing schemes, reflected in the last term of (1), $p_{s_k}(L_i)$.

Based on this theoretical 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

Networking charging and accounting approaches can be divided into three categories:

(1) Best-effort charging and accounting (such as the one of MacKie-Mason and Varian (1995)), for pricing traditional Internet traffic

served without differentiation or QoS guarantees.

(2) Flow- or reservation-based charging and accounting of integrated services (such as the one proposed in Fankhauser *et al.* (1998)), for pricing individual traffic flows receiving differentiated service with certain quality guarantees.

(3) Charging and accounting of DiffServ-based services (for pricing aggregates of flows according to the DiffServ, open-loop congestion principles), proposed pricing models for which will be presented in this section

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 (see Falkner *et al.* (2000), DaSilva (2000)) appropriate to account for congestion costs, service differentiation, QoS provision and other relevant costs for pricing today's connectionless IP networks.

The surveys of Falkner *et al.* (2000) and DaSilva (2000) emphasize the role that a pricing mechanism must have in traffic management (congestion controls, resource provisioning and call admission). In DaSilva (2000), 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 or 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.

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 (MacKie-Mason and Varian, 1995), 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 Clark (1996). 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 Clark (1996). 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, Clark (1996) 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 Shenker *et al.* (1996), 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.

In Cocchi *et al.* (1991), 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 (Cocchi *et al.*, 1991).

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.

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 Courcoubetis *et al.* (1997) two compatible approaches for charging flows obeying to traffic contracts (or SLAs) according to their effective bandwidth are presented:

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

Semret *et al.* (1999) proposed a distributed auction mechanism called progressive second price (PSP), which does not assume any specific mapping of resource allocation to QoS. PSP is an auction mechanism for sharing of a single divisible resource among bidders. Each player submits a sealed bid and the object is sold to the highest bidder at the bid price of the second highest bidder. An equilibrium state where all players bid their true valuation is reached. Like all successful auctioning mechanisms, resources are always allocated to those who value them most.

In Wang and Schulzrinne (2000), 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 DaSilva *et al.* (1997), 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 DaSilva *et al.* (2000), 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 a 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.

## 4. Our proposed model

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.

Fixed charging according to, for example, access speeds has been mostly popular in the flat service model of the traditional best-effort Internet. Connection fees have been mostly appropriate for connection oriented networking architectures and service provisioning (e.g. resource reservation through the resource reservation protocol (RSVP)). DiffServ-based services provisioning, however, is not a flat service model and does not operate on a connection-oriented basis. Therefore, neither fixed access fees nor connection set-up fees are appropriate. As already mentioned, the majority of relevant research work has appointed usage-based charging per service class as the most appropriate alternative in order to capture the particularities of DiffServ service provisioning.

One alternative is for pricing 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.

During the service provisioning interval, a usage-based approach should be followed. However, contrary to the common belief of quantitative usage-based monitoring per unit of consumption, we propose charging according to deviations from the predetermined traffic profile. Such an approach requires less storage of monitoring data since it does not account for the whole amount of traffic transmitted, but only for the amount that exceeds or falls short of the traffic profile. This can be implemented in a straightforward manner at the points where traffic metering and policing is performed in a DiffServ architecture, i.e. in edge routers of a service provider's domain, obeying thus to the "edge pricing" idea. In this way, deviations from the predetermined traffic profiles are taken into consideration in the pricing procedure after a service provisioning interval and the gap between what is paid for and what was actually used diminishes.

As per-service class aggregates of each customer pass through metering and/or policing filters at network edges, two cases where deviations from final costs will differ from the a priori estimation for the traffic contract exist: when the customer either under-utilizes his predetermined traffic profile or over-utilizes it. In the latter case, most commercial IP networking equipment (routers) can provide a report of the "illegal" packets during an operation interval in a very straightforward manner (usually with a single operating system command). In the former case, minor modifications will be required in order for reports on how much more traffic a customer could have "squeezed" in his traffic profile to be feasible.

In either case, the provider can easily present to each customer an indirect but precise report on his usage of a certain service class during a predetermined operation period, through statistics on the deviation of the customer's agreed service profile. It is then up to the service provider's policies to drive customers to more accurate traffic profile estimations for the following operation period. This can be accomplished by pricing schemes that increase exponentially the price per out-of-profile packet as the deviation from the predetermined traffic profile increases.

As far as the role of pricing in service class selection is concerned, through the proposed scheme it is also possible for the network provider to fluctuate prices for traffic profiles within a service class from one operating interval to another when demand for resources within the service class fluctuates. In other words, the number of subscriptions for a certain service class will be co-estimated when determining prices for the specific class during the next operation interval. This trade-off between subscription level and prices within a service class should carefully balance around a point where the qualitative guarantees of the service class are not compromised. In this way, over-subscription of a certain service class and the corresponding degradation in offered quality is avoided or at least the correct incentives are provided to customers to restrain themselves in service class selection. For example, the highest quality service class will be the most expensive to request and its price will increase exponentially as the number of requests increases.

In order to summarize this section, the key elements of the proposed scheme are outlined:

- the fluctuation of prices for usage of each service class through operating intervals, in order to reflect the customers' demand for the guarantees offered by each class; and
- the calculation of costs for each customer of a service class according to an initial charge for the traffic profile purchased and the deviation from the profile during each operating interval.

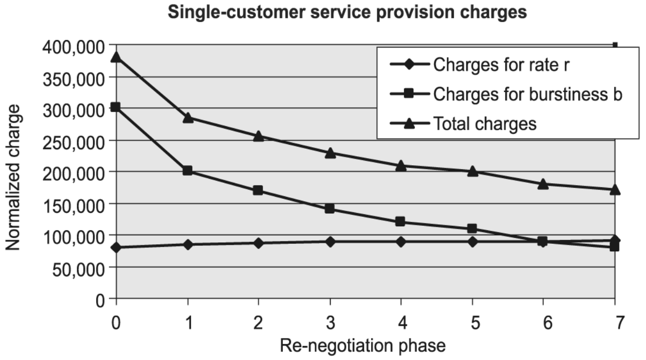
An analytical methodology for calculating the costs that have to be imposed to each customer of a specific service class during a service interval is proposed and evaluated in Bouras and Sevasti (2004). Based on this approach, a set of procedures is also proposed for the provision and pricing of a specific DiffServ-based service over a transport domain (TD). This proposal is summarized in the sequel.

Initially, each customer agrees on the traffic contract with which his traffic will be policed while entering the TD. Based on his local policy for provisioning of the specific service, an analytical methodology for the service configuration and for calculating the costs that have to be imposed to each customer during a service interval as presented in Bouras and Sevasti (2004), the TD provider makes the necessary configurations. According to them, the qualitative guarantees that can be offered to the service customers in the upcoming service interval can be effectively approximated. At the same time, the cost for providing the requested service to each customer is calculated and the customers are then informed in advance about the cost they will be asked to pay for the upcoming operation phase.

The service is initialized and provided for a number of days. During the operation phase, at the interface of the edge router where each customer's traffic aggregate is policed according to the corresponding traffic profile, traffic profile statistics are maintained at regular intervals. After an operation phase is terminated, the statistics collected are evaluated and the SLAs are either preserved or adjusted. Each customer is presented with the statistical data of the previous operating period. Based on these data, each customer is able to apply for a new traffic profile for his traffic in the upcoming service period, if the already existing profile did not effectively approximate the real traffic transmitted or if the customer considers the cost corresponding to the existing traffic profile unfair. 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 resources. Customers can also place bids for resources in the upcoming operation phase, taking into consideration the sampled data of the previous operating period and the qualitative guarantees provided by the TD. The TD provider is evaluating all bids in the order of offers, starting from the highest offer and provides resources up to its availability of resources for a certain level of offered quality guarantees. In this way the traffic profiles for the next operating period are determined for all customers and the next operation phase can be initiated.

The proposed methodology, the analytical approach of which is out of the scope of this work, has been evaluated in Bouras and Sevasti (2004) and the evaluation results are briefly provided here. We provide here some of the simulation results of applying the proposed mechanism to a real-world scenario of providing a DiffServ-based service to a set of customers. For our specific case study, customer traffic profiles were in the form of token bucket profiles. It was noted during the simulation that, with small fluctuations, each customer updated his traffic contract throughout the iterations so as to describe more tightly his traffic profile and thus requested resources were redistributed efficiently. In Figure 2, the normalized charges imposed to a single customer throughout the consecutive re-negotiation periods are depicted. One can observe how the statistical data provided to the service customer and the incentive-based pricing scheme proposed led to a tighter traffic descriptor, which was economically beneficial for the customer.

**Figure 2** Evolution of the charges paid by a single customer during the re-negotiation phases of a DiffServ-based service provisioning



As part of our evaluation, a more efficient allocation of resources was also achieved from the TD provider's point of view. The decrease in revenue for the TD provider was compensated by new customers that could now be accommodated. By providing incentives to existent customers to reveal their true traffic profiles through re-negotiation periods, the provider could become aware of the true utilization of resources in his backbone and was then able to accommodate new customers without compromising quality.

## 5. Conclusions – future work

In this work, after outlining the role of pricing in the new era of QoS-enabled networks, we have presented related research work with respect to how successfully this role is dealt with. It turns out that, although the theoretical principles of QoS-pricing have been thoroughly examined, not many practical and ready to implement approaches exist.

On this basis, we provide an overview of our proposal for pricing over a network that offers multiple levels of quality. The proposed approach applies to DiffServ-based services and demonstrates engineering efficiency since it preserves complexity at network edges and, thus, preserves scalability. It aspires to obtain simplicity in calculation of customers' charges and effectively reveal service differentiation and QoS provision. We also have to point out that this approach provides customers with incentives to restrain their demands for resources, and a realistic estimation of the costs they will be asked to pay for the use of a service prior to the period to which these costs apply.

Our work is based on a coarse categorization of DiffServ-based services according to the nature of quality guarantees they provide, since this coefficient is part of the optimization problem (factor $U(Q_{s_k})$ in (1)). An analytical approach for a specific sub-category of DiffServ-based services has already been devised and evaluated, however we aim to develop per service-class and per-traffic-contract-utilization (within a service class) theoretical models for each broad category of DiffServ-based services. We are also currently investigating the most efficient determination of granularity for the proposed pricing intervals.

## References

Blake, S. *et al.* (1998), An architecture for differentiated services, RFC2475.

Bouras, C. and Sevasti, A. (2004), "A new pricing mechanism for a high-priority DiffServ-based service", *Proceedings of the 6th International Conference on Advanced Communication Technology (ICACT2004)*, Korea.

Clark, D.D. (1996), "A model for cost allocation and pricing in the Internet", in McKnight, L.W. and Bailey, J.P. (Eds), *Internet Economics*, MIT Press, Cambridge, MA, pp. 155-68.

Cocchi, R., Estrin, D., Shenker, S. and Zhang, L. (1991), "A study of priority pricing in multiple service class networks", *Proceedings SIGCOMM '91*, Zurich, pp. 123-30.

Courcoubetis, C., Kelly, F.P. and Weber, R. (1997), "Measurement-based charging in communications networks", *Technical Report 1997-19*, University of Cambridge, Cambridge.

DaSilva, L.A. (2000), "Pricing for QoS-enabled networks: a survey", *IEEE Communications Surveys & Tutorials*, Vol. 3 No. 2, pp. 14-20.

DaSilva, L., Petr, D. and Akar, N. (1997), "Equilibrium pricing in multi-service priority-based networks", *Proceedings IEEE GLOBECOM'97*, pp. S38 6.1-S38 6.5.

DaSilva, L., Petr, D. and Akar, N. (2000), "Static pricing and quality of service in multiple service networks", *Proceedings 5th International Conference on Computer Science and Informatics*, Atlantic City, NJ, Vol. 1, pp. 355-8.

Falkner, M., Devetsikiotis, M. and Lambadaris, I. (2000), "An overview of pricing concepts for broadband IP networks", *IEEE Communications Surveys & Tutorials*, Vol. 3 No. 2, pp. 2-13.

Fankhauser, G., Stiller, B., Vögtli, C. and Plattner, B. (1998), "Reservation-based charging in an integrated services network", *Proceedings 4th Informs Telecommunications Conference*, FL.

MacKie-Mason, J. and Varian, H. (1995), "Pricing the Internet", in Kahin, B. and Keller, J. (Eds), *Public Access to the Internet*, Prentice-Hall, Englewood Cliffs, NJ, pp. 269-314.

Odlyzko, A. (2001), "Internet pricing and the history of communications", *Computer Networks*, Vol. 36 No. 5, pp. 493-517.

Semret, N., Liao, R.R.-F., Campbell, A.T. and Lazar, A.A. (1999), "Market pricing of differentiated Internet services", *Proceedings 7th IEEE/IFIP International Workshop on Quality of Service*, London, pp. 184-93.

Shenker, S., Clark, D., Estrin, D. and Herzog, S. (1996), "Pricing in computer networks: reshaping the research agenda", *ACM Computer Communication Review*, Vol. 26 No. 2, pp. 19-43.

Wang, X. and Schulzrinne, H. (2000), "Performance study of congestion price-based adaptive service", *Proceedings NOSSDAV 2000: Network and Operating Systems Support for Digital Audio and Video*, NC, pp. 1-10.

## Further reading

Clark, D.D. and Fang, W. (1997), "Explicit allocation of best effort packet delivery service", Technical report, Lab for Computer Science, MIT, Cambridge, MA.