

Charging Multi-dimensional QoS with the Cumulus Pricing Scheme

Peter Reichl^{*1}, Burkhard Stiller⁺, Thomas Ziegler^{*}

^{*}Telecommunications Research Center Vienna (Forschungszentrum Telekommunikation Wien FTW),
Maderstraße 1, A – 1040 Vienna, Austria

⁺Institut für Technische Informatik und Kommunikationsnetze TIK, ETH Zurich
Gloriastrasse 35, CH – 8092 Zurich, Switzerland

ABSTRACT

The recently established Cumulus Pricing Scheme (CPS) has turned out to be a novel approach for efficiently charging differentiated Internet services, based on integrating different time-scales into one edge-pricing mechanism. Depending on an initial specification of expected resource requirements, customer and provider negotiate a contract fixing a flat rate charge for QoS delivery. As soon as the scheme has started, the customer receives a continuous coarse-grained feedback about her real resource consumption. To this end, over- or underutilization are expressed in terms of Cumulus Points CP, whose accumulation may indicate an imbalance between specified and actually monitored traffic and eventually requires to adapt the contract accordingly. This paper extends the original CPS for services that are characterized not only by their bandwidth or volume requirements, but by general QoS parameters. Starting with a discussion on CPS for different one-dimensional QoS parameters, consequences for the basic CPS mechanism are investigated, covering especially the determination of relevant thresholds for CPs. These investigations deliver crucial input for the specification of multi-dimensional QoS vectors within the initial contract. Suitable metrics are introduced and applied in order to reduce the complexity of the contract as well as of the different monitoring methods. Finally, the implementation of the extended scheme within an Internet Charging System is discussed.

Keywords: Internet Pricing, QoS, NUT Trilemma, Cumulus Pricing Scheme, Internet Charging System

1. INTRODUCTION

Providing Quality-of-Service (QoS) in IP-based networks has been subject of intensive research work over the last couple of years. With the rising commercialization of the Internet and the introduction of new types of services and applications like Voice-over-IP or Video-on-Demand with their stringent requirements on transmission quality, this area has gained even more importance. Whereas the main body of the related work focusses on technical mechanisms for providing QoS in a given network, like, e.g., allocation, control, measurements, and policing mechanisms, the increasing differentiation of services has led to the insight that besides the investigation of technical parameters, additionally also economic aspects need to be considered. As a consequence, research projects like M3I (Market-Managed Multi-service Internet) [13] or CATI (Charging and Accounting Technology for the Internet) [20] have been established to investigate the intimate relationship between the technical design and the economic consequences of providing QoS for differentiated Internet services.

In principle, market-managed mechanisms fulfill two different tasks. On one hand, they may allow for cost recovery, if the pricing scheme applied to the services offered is cost-based. This type of pricing schemes show a quite static behavior, since the price per unit service will be fixed on a longer time-scale. On the other hand, market-managed mechanisms may provide a congestion-control approach by deploying usage-based, dynamic pricing schemes. This duality is a simple consequence of the basic functionality of pricing as a means of information exchange between customer and Internet Service Provider (ISP). Prices and the subsequent flow of money are always an expression for the customer's interest in using a certain service, based on her willingness-to-pay, but at the same time they are an efficient tool for communicating the degree of scarcity of a required resource. It is the second aspect that often needs to be stressed, specifically in the context of QoS provision. As the flat rate schemes widely deployed nowadays do not provide any incentive for an efficient usage of the network and thus is an apparent obstacle for guaranteeing QoS, over the last few years a couple of pricing schemes for differentiated Internet services have been proposed that claim to provide correct economic incentives. For a comprehensive survey about the state of the art in this field we refer to [7] and [25].

While the M3I project deals with many of these pricing schemes, two important problems so far have remained unsolved:

1. Corresponding author, phone +43 1 505 28 30, fax +43 1 505 28 30-99, or email reichl@ftw.at

- The vast majority of currently relevant approaches focus on pricing either bandwidth or volume of a service, and in doing so neglect further QoS parameters like delay or throughput. Hence, the question of how to deal with pricing these parameters has hardly been tackled so far at all.
- If a service is characterized by a single QoS parameter only, this parameter can be used to be mapped onto a price, but often the QoS of a service are characterized by two or even more QoS requirements that have to be satisfied simultaneously. This leads directly to the question of how to map a multi-dimensional QoS characterization of a service onto a price and which technology is required to perform this mapping. As a multi-dimensional QoS approach has not been followed yet in the related literature, the specification of a QoS parameter mapping is essential for the definition of an appropriate pricing scheme for services with multi-dimensional QoS specifications.

The rest of the paper is structured as follows: After a short review on QoS basics, we deal with services characterized by one or more general QoS parameter (i.e. multi-dimensional QoS) and the respective metrics. After introducing the NUT principle and generalizing the “Feasibility Problem” for Internet tariffing [15] correspondingly, we review the traditional version of the Cumulus Pricing Scheme CPS. After extending CPS for different one-dimensional QoS parameters, consequences for the basic CPS mechanism are investigated, focussing especially on the form of the relevant thresholds. These investigations deliver crucial input for the specification of multi-dimensional QoS vectors within the initial contract. Suitable metrics are introduced and applied in order to reduce the complexity of the contract as well as of the different monitoring methods. Finally, the implementation of the extended scheme within an Internet Charging System is discussed, before a summary and outlook finish the paper.

2. PRICING-RELEVANT QUALITY-OF-SERVICE PARAMETERS

The main aim of this section is to introduce the basic framework for charging Quality of Service. The section starts with reviewing some QoS basics, especially the relevant parameters and their end-to-end characteristics, as well as examples for QoS requirements of a couple of important applications. Multidimensional Quality-of-Service is introduced, using QoS vectors and respective metrics¹. This allows to derive some general conclusions with respect to requirements for pricing services with guaranteed QoS.

Today's applications exhibit distinct performance requirements on packet networks. These requirements are generally expressed as end-to-end QoS parameters like throughput, packet delay, delay jitter, and packet loss probability in a Service Level Specification (SLS). In the context of this paper, end-to-end QoS parameters are defined as follows:

- **Throughput** is defined as the arrival rate [bit/s] at the receiver's application layer (subsequently denoted as "the receiver"). Note that this rate may be different from throughput measured at the link layer due to the fact that retransmissions and duplicated transmissions (common with transport protocols like TCP) are not visible from the application layer's point-of-view.
- **One-way delay (latency)** is defined as the difference in time between the arrival of a packet's last bit at the receiver and the transmission of the same packet's first bit by the sending application (subsequently denoted as “the sender”). Delay consists of four components: propagation delay, queuing delay at congested output ports, transmission delay (packet size divided by link capacity) and the processing (or forwarding) delay in routers and hosts.
- **Delay jitter** is defined as the variance of the inter-packet arrival time at the receiver.
- **Packet loss probability** is defined as the ratio of packets lost in the network to all packets transmitted by the sender.

Any QoS enabled network has to perform admission control and traffic control on a per-hop basis in order to be able to meet a flow's end-to-end guarantees. Thus, end-to-end QoS parameters are a composition of their corresponding hop-by-hop QoS parameters. As defined in [26] and [5], the most important composition-rules for end-to-end QoS parameters are:

- **Additive composition:** the end-to-end metric equals the sum of the per-hop metrics. As examples for additive QoS parameters, delay and delay jitter can be mentioned.
- **Concave composition:** the end-to-end metric equals the minimum of all per-hop metrics. Intuitively obvious, throughput is an example for a concave metric.
- **Multiplicative composition:** the end-to-end metric equals the product of the corresponding per-hop metrics. The packet loss probability e.g. can be easily defined as the inverse of a multiplicative metric - the probability a packet gets through. Let $L_{i,j}$ denote the drop probability for link (i,j) . For any path $\pi = (i,j,k,\dots,m,n)$ the end-to-end packet loss probability $L(\pi)$ can be defined as

1. Note for the rest of the paper that the notion of a “metric” is not used in a strictly mathematical sense, but rather as an intuitive concept for characterizing the “absolute” size of a variable parameter value or of the distance between two of them.

$$L(\pi) = 1 - ((1 - L(i, j)) \cdot (1 - L(j, k)) \cdot \dots \cdot (1 - L(m, n))) \quad (1)$$

As an input parameter to pricing schemes it may be necessary to combine the end-to-end QoS parameters which a network is able to guarantee into a single value. Publications on QoS routing [26] discuss possible ways to merge several QoS parameters into a single metric. Let e.g. $T(\pi)$ denote the throughput, $d(\pi)$ the latency and $L(\pi)$ the loss probability a flow's packets experience along a path π . Then [26] proposes to define the combined metric $C(\pi)$ as

$$C(\pi) = \frac{T(\pi)}{L(\pi) \cdot d(\pi)} \quad (2)$$

Assume now according to [4] that the QoS parameters, together with the bandwidth β and price p , form a triad

$$\tau_G = (\beta, G(\vec{q}), p) \quad (3)$$

that is sufficient for characterizing the traffic contract between customer and provider. Here, the so-called QoS vector $\vec{q} = (d, \Delta d, L, T)$ is determined by delay d , delay jitter Δd , packet loss probability L and throughput T , where the mapping $G = G(\vec{q})$ denotes a set of constraints (quality requirements) on the QoS vector. The four dimensions of this QoS vector are strictly conforming with the notion of a "flow specification" as provided by the Resource Reservation Protocol (RSVP). Here, the flowspec characterizes the desired QoS, which is complemented by the "filter specification" as well as a "session specification", both of which define the sequence of packets (the flow), which receives the specified QoS. The parameters of the QoS vector are applied in the router to configure the node's packet scheduler or further link layer mechanisms appropriately. More specifically, the reservation request performed by RSVP includes a reservation specification Rspec defining the desired QoS, a traffic specification Tspec describing the data flow, and a service class. The use of the flowspec within in RSVP and their defined services may be obtained from [27].

The problem of defining and measuring QoS parameters is subject to discussion in the IETF IPPM (IP Performance Metrics) WG, see [14], [1], [2]. For instance, measuring delay requires a tight time synchronization between the sending and the receiving host. For any kind of QoS parameter measurements sampling intervals have to be chosen carefully, the influence of the measurement on the system (e.g. with respect to the extent the measurement tool needs to inject extra packets into the net) has to be minimized and measurement errors have to be estimated. It is, however, worthwhile noting that (as it is shown in [4]) basically all end-to-end quality parameters may be expressed as simple linear or exponential functions of n , the number of hops between sender and receiver.

On the other hand, Quality-of-Service has an important user aspect, too. The customers are the one willing to pay for quality, and one could go as far as [12] to say that the payment itself is the decisive factor, rather than the reason why the charge is paid, i.e. QoS. The key concept for describing this aspect is the notion of utility, commonly defined as "the quality or condition of being useful" [12] and expressed in terms of a utility function mapping QoS parameters towards their (e.g. monetary) utility. Figure 1 sketches qualitatively some typical utility functions as reviewed in [7] or [25].

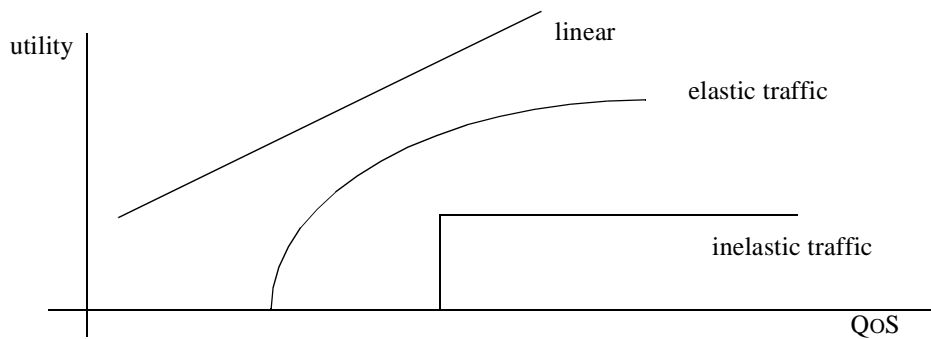


Figure 1. Typical Forms of Utility Functions

For further illustration, according to [18] we may classify applications according to their timeliness requirements:

- Interactive audio and video applications exhibit the most stringent timeliness requirements. For instance, good audio quality requires upper delay bounds of 150ms and loss probabilities smaller than 10%.
- Transaction applications like telnet or stock-market queries require low queuing delay and loss probability at the network-layer to give users a "realtime-feeling". Note that these kind of applications need to employ a reliable transport protocols like TCP as loss of information at the application layer would be unacceptable. Thus strict upper bounds for end-to-end delay cannot be given.
- Adaptive playback applications (e.g. video on demand, real player) generally require less strict QoS guarantees. Delay and delay jitter are not an issue for adaptive playback applications. However, a minimum throughput guarantee over long time-averages is an essential requirement.
- Bulk-data applications like FTP, WWW have no delay requirements and may be satisfied with best-effort service. However, a minimum-throughput guarantee is desired by many users generating bulk-data flows in order to ensure an upper bound on the flow transfer time.

[19] gives an example how to map these applications into Diffserv PHBs (Per Hop Behaviours). Interactive audio and video applications are mapped into an EF PHB [9]. Transaction applications are mapped into an overprovisioned AF PHB [8] with relatively small queue management drop thresholds to ensure low loss probabilities and small queuing delays. As their QoS requirements are quite similar, adaptive playback applications and "better than best-effort service" bulk data applications are merged into a single AF PHB with relatively high queue management drop thresholds to maximize throughput. Of course, best-effort service is still of major importance also in a QoS architecture as defined in [19].

3. THE NUT TRILEMMA AND THE CHARGING OF QUALITY-OF-SERVICE

As shown in [15], proposing a pricing scheme for differentiated Internet services is closely related with balancing the trade-off between three main requirement types: *Network efficiency* (i.e. the ISP's goal to fill its resources), *User acceptance* (i.e. the consumer's desire to have an understandable tariff which is transparent, stable and predictable) and *Technical feasibility* (i.e. the marginal conditions posed by the technical limitations of processing accounting data). Figure 2 represents this so-called "NUT Trilemma" and sketches a rough grading of some popular pricing schemes.

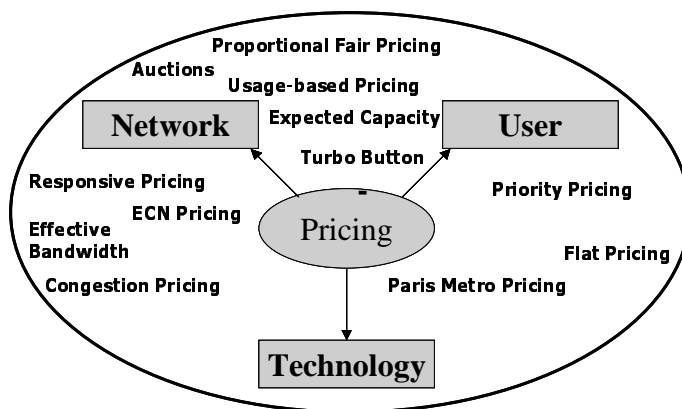


Figure 2. The NUT Trilemma

E.g. Flat Rate Pricing is highly estimated by the users (due to its simplicity and stability) and perfect in terms of technology (not requiring any measurement or accounting process at all), but on the other hand does not provide any incentive for an efficient network usage (as is sufficiently demonstrated by the recent breakdowns of a number of ISPs offering this scheme, see also [3] for a more detailed view), thus in Figure 2 we find this scheme (as well as its derivatives, e.g. Paris Metro Pricing) close to "U" and "T" and strictly opposite to "N". Usage-based pricing is much better in terms of efficiency as economic incentives are given, and still understandable for the user, but may require sophisticated accounting mechanisms, depending on the detail of measured data. Responsive Pricing and Smart Market Auctions are attempts to optimize the network aspect, but are not feasible in terms of complexity, etc. In this way, any Internet tariff can find its place in this schematic evaluation. As a general result of this evaluation, two things are interesting to note: (1) basically all relevant proposals have their starting point either at the network or the user aspect, and (2) there appears to be no scheme so far that succeeds in balancing the tradeoff sufficiently. On the other hand, the discussion in [15] has shown that the technical requirements are more stringent than both of the others (this fact has been termed the "Feasibility Problem of Internet Tariffing" [15]), an aspect that becomes even more important as soon as further QoS parameters like delay, jitter, loss probability etc. are to be integrated into a pricing scheme besides bandwidth or volume.

As a consequence, [15] has proposed a paradigm change: instead of developing economically efficient schemes whose complexity is reduced afterwards to make their implementation possible, we strongly argue for using general principles to construct a *feasible* scheme and extend it in a second step such that eventually it contains the correct economic incentives. In [16] it is demonstrated that one such possibility focusses explicitly on time-scale aspects of Internet pricing and views tariff schemes primarily as an expression of mappings between these time-scales. The resulting framework, the so-called Cumulus Pricing Scheme, will be reviewed in Section 4, before it is extended for the case of multi-dimensional QoS parameters.

In this context, there is another important issue to be discussed: Using bandwidth or volume (i.e. total bandwidth over a certain period in time) as the basic input parameter for a pricing scheme (as almost any scheme does, except for the flat rate-type schemes) is possible due to a symmetry property: the amount of bandwidth or volume in usage can be reduced both by the customer (i.e. through reducing her sending rate) and the network (e.g. through traffic policing or packet loss due to buffer overflow in the worst case), and can be increased by both of them, too (the user may increase her sending rate, the network can assign priority to the respective flows and assign them larger bandwidths). This is not true for all other parameters contained in a QoS vector: the direct influence of the user on those parameters is reduced to one direction only, as there is apparently no possibility for the user to decrease delay, jitter, packet loss probability etc. Decreasing these parameters is up to the network, with respect to certain natural marginal restrictions.

4. GENERALIZING THE CUMULUS PRICING SCHEME

The Cumulus Pricing Scheme has been recently established as a first proposal for a new class of Internet tariffs emerging from the so-called “Methodology of Time-Scales” MTS, a formal matrix-based framework for describing tariff schemes as multi-dimensional mappings between different time-scales as introduced in [16]. In this section, we shortly review the basic discussion leading to that framework, before we summarize the main features of the scheme itself.

4.1. The Way to a New Pricing Scheme

Starting from the NUT trilemma as described in Section 3, each current proposal for Internet Pricing may be characterized by the way the tradeoff between these three conflicting goals is solved, as sketched in Figure 2. As the “Feasibility Problem” mentioned above leads to the conclusion that among the three identified requirement types, “T” is the most basic one, we may note as a first conclusion that our search for a suitable pricing structure will start from flat rate-type schemes (as they are optimal with respect to “T”) and try to provide them with sufficient flexibility for introducing also aspects of economic efficiency.

From a second point of view, the suspense between “N”, “U” and “T” may also be traced back to the different time-scales involved. It is pretty obvious that Flat Rate Pricing operates on an extremely long time-scale, whereas Smart Market auctions are supposed to take place quite rapidly. This observation has led to the more general question of which time-scales one should consider as being involved in Internet pricing, and how we should force them to interact in order to create a good tariff scheme.

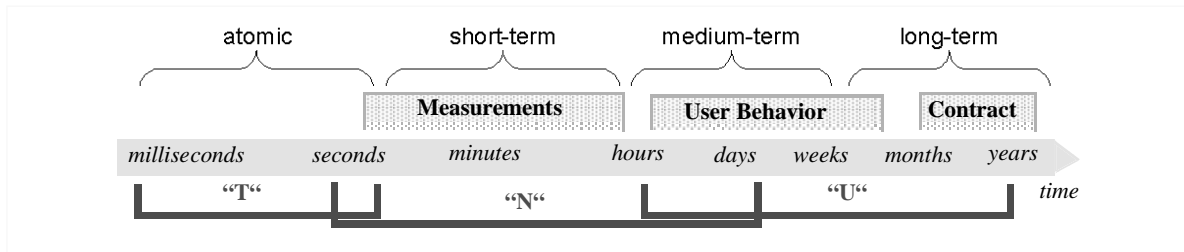


Figure 3. Time-Scales, Requirement Types and Pricing Activities

Standard literature on Network Management distinguishes three overall management time scales: short-term in minutes, medium-term in hours and long-term in weeks or months. For the context of Internet Pricing, it has turned out to be necessary to introduce a fourth time-scale describing the basic communication processes, the so-called “atomic” time-scale [23]. The resulting scheme, differentiating communication-based (atomic), application-based (short-term), billing-based (medium-term) and contract-based (long-term) activities, is depicted in Figure 3. In a second step, we have placed also the three requirement types originating in the NUT trilemma within this picture. In doing so, we notice that the requirements apparently have to be placed at the transition points *between* the different scales: Transparency and predictability (i.e. the user aspect) refers to relatively long time-scales (medium- and long-term), due to time requirements for human everyday life behaviour. The network aspect is correlated with short- and medium-term actions, due to the reaction time of applications and users, whereas the accounting aspect deals both with atomic and short-term events, processing data originating in packets and flows. As a consequence, the trade-off between the requirement types can be reformulated as a trade-off between time-scales.

Finally, CPS has not only been designed as a compromise between the basic requirement types and as a interaction mechanism for relevant time-scales, but also aims at integrating a couple of further useful properties coming from existing tariff proposals (see references in [7] and [25]). In our context, the following approaches are especially relevant:

- **Edge Pricing:** The idea of shifting charge computation to the (spatial) edges of the network is widely accepted as useful general principle as it allows the concentration of computing resources. Moreover, it may be easily transferred into the time dimension, thus substituting continuous charging activities by discrete ones that take place at "edges in time", e.g. at the begin and at the end of a given resource consumption process.
- **Expected Capacity Pricing:** Whereas already the edge pricing proposal is operating with the concepts of expected resource consumption and expected routing, the Expected Capacity Pricing approach explicitly focuses on requiring a user specification about her expected capacity and charging this specification accordingly, based on a long-term contract.
- **Effective Bandwidth Pricing:** Without going into details, two major general ideas deserve attention here. Characterizing complex traffic behaviour by a single number greatly simplifies relevant traffic contracts, a concept useful also for QoS parameters. Moreover, shifting the responsibility of traffic specification (through a choice between tariffs) within the traffic contract to the user introduces a type of user liability that is helpful both in terms of legal as well as economic aspects.
- **Usage-based Charging:** Finally, taking somehow the actual resource consumption into account directly yields straightforward improvements as far as economic efficiency is concerned.

4.2. The Cumulus Pricing Scheme

Based on the previous sections, we may characterize this new pricing scheme which has been established in [15] and [17] (see also [24] for its application within a generic Internet Charging and Accounting System) in different ways, i.e. (1) as a well-balanced "mixture" of established principles derived from a number of different proposals in the related literature, (2) in terms of multi-dimensional mappings between the various time-scales identified for network operation, network management, requirement types and pricing activities, or (3) in terms of its basic steps. To start with the third aspect, the whole scheme is based on an initial contract (the so-called "Cumulus Pricing Contract" CPC) between customer and ISP. The CPC is related closely to the concept of a Service Level Agreement (SLA) as deployed in DiffServ environments (see [24]) and contains a traffic specification delivered by the customer, a flat rate related to this specification which is offered by the provider, and a couple of thresholds that are necessary to watch later whether the CPC is still valid. Note that the flat rate explicitly depends on the traffic specification, e.g. through a tariff function. In [17], this function has been analytically derived for the case of bandwidth pricing, based on the general assumption that the form of the function has to provide the customer with the incentive to be honest about her real requirements. For the example of bandwidth, this "incentive compatibility" condition means that the charge finally paid by the customer is minimal in the case of a correct initial specification.

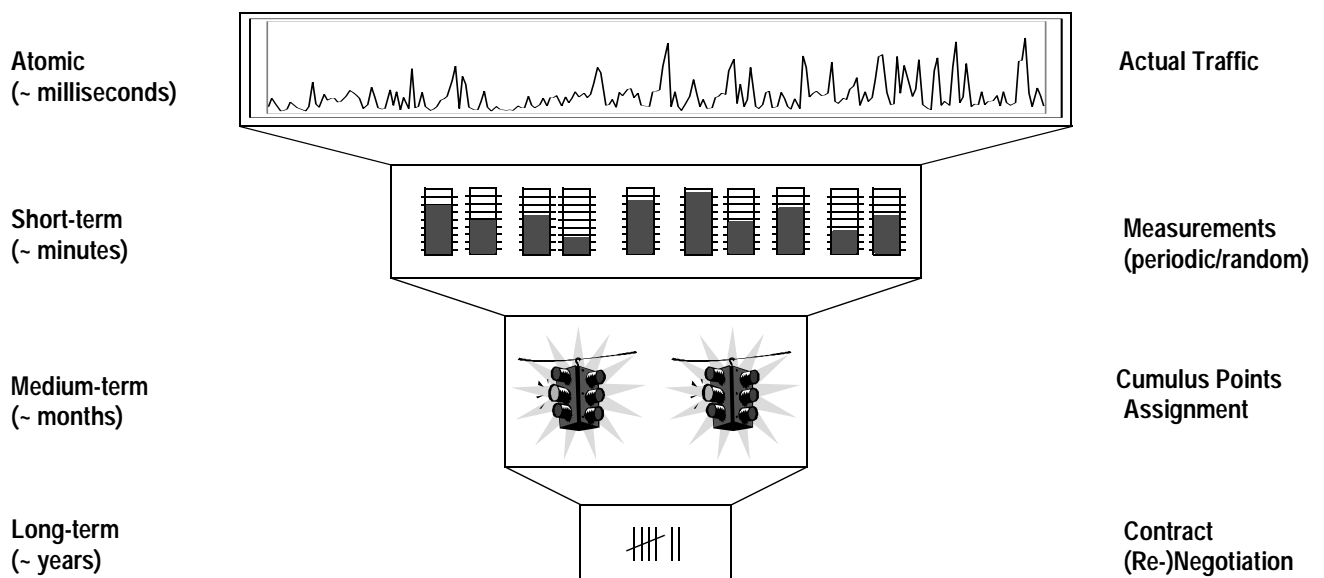


Figure 4. The Four Levels of Cumulus Pricing: Time-scales and Activities

Having negotiated the CPC, the actual traffic has to be monitored, where the way of performing measurements is left rather open to the provider. As a result of the measurements, from time to time the customer receives a feedback about whether she is over- or underusing her specification as presented in the CPC. This feedback is rather rough, i.e. in terms of a small number of flags which are called “Cumulus Points” CP and which appear in two flavours, i.e. as red and green ones, depending on the direction in which the CPC is violated. Note that the CPs have no immediate influence on the charge to be paid or the contract between customer and ISP, but are accumulated over time, thus indicating that the contract may be out of balance, but leaving still possibilities for appropriate reactions. Only if the accumulation of CPs exceeds a certain threshold, renegotiating the CPC becomes mandatory.

Figure 4 sketches the four levels of this traditional version of the Cumulus Pricing Scheme. From this figure, we can directly derive the second mentioned viewpoint, i.e. CPS as mapping of different time-scales. Designing a flat-rate type scheme that respects market forces requires to introduce flexibility into the long-term flat rate, i.e. the charges as agreed upon within the CPC can no longer be fixed forever. The market forces reveal themselves primarily in the form of a feedback to the customer. One might argue that such a medium-term feedback mechanism could also be placed on the level of network management instead of being a part of the tariff scheme. In our view, this alternative is not very useful, because prices are the most simple and universal communication interface between provider and customer, so if it is possible to express the required feedback in a money-related way, then the ISP should do so. Note that it is not an immediate consequence of this argument that feedback requires a change in the actual charge (in fact, CPS does avoid exactly this consequence), but at least a feedback mechanism telling the user whether she strongly over- or underuses her specified resource requirements (in the form of a warning or bonus system, respectively) turns out to be an integral part of the tariff, thus introducing the medium-term time-scale aspect into the scheme. Obviously, any useful feedback mechanism must be based on some sort of measurement and monitoring activity. This in fact is one of the crucial performance bottlenecks of any charging system. Therefore, we argue strongly for leaving the amount of measurement complexity as far as possible up to the provider. In this sense, our basic assumption concerns the mere availability of some monitoring facility, without posing any requirements to their granularity and the details of the records. Therefore, the ISP may (to view but the extreme cases) measure each packet if he likes to do so, but taking only rare snapshots in a periodic or aperiodic way should also be possible and sufficient. Even if the monitoring facilities have to be temporarily shut down for some reasons, this should not have any consequences for the provider. Note that we are well aware of the fact that these “low-level” requirements on the monitoring have also enormous social and legal aspects, e.g. public operators will have to enforce stricter rules concerning the indisputability of their measurements than private operators, but generally we argue for the reduced assumption of just any measurement facility, depending on the ISP’s policies.

However, all activities concerning the basic communication-related processes (e.g. sending packets, routing and switching etc.) are still restricted to the atomic time-scale. On the other hand, it is the complexity of these activities which is the ultimate cause for the existence of the feasibility problem. Put it in other words, in order to design feasible pricing schemes, we have strictly to avoid the appearance of the atomic time-scale within the tariff mechanism. A more formal description of this situation yields to a new view on Internet pricing schemes: tariffs may in this sense be formulated as mappings of the atomic time-scale to the higher ones, i.e. the short-term, medium-term and long-term time-scale or any combination of them. This is the essence of the “Methodology of Time-Scales” [16] as mentioned in the beginning of Section 4. For a further clarification we refer again to Figure 3 for a sketch of the relationship between time-scales, requirement types and CPS activities.

Finally, CPS integrates apparently a couple of well-established principles as formulated in the context of Edge Pricing, Expected Capacity Pricing, Effective Bandwidth Pricing and Usage-based Charging (see Section 4.1). CPS is an almost classical example for edge pricing, in fact not only in space (i.e. at the edges of the network) but also in time (i.e. at the edges of basic periods on the various time-scales), see [17]. It extends the concept of Expected Capacity Pricing to all relevant QoS parameters whose specification within the CPC forms the basis for the offered flat rate as well as the subsequent feedback mechanism. The influence of the Effective Bandwidth Pricing (also known as “Kelly’s abc scheme”[10]) is visible in putting the responsibility for specifying the traffic correctly (via a suitable choice of tariffs) into the hands of the customer. And finally, CPS can also be viewed as a sub-species of usage-based charging schemes, due to the measurements which are responsible for giving feedback to the customer but, in contrast to original usage-based schemes, do not form an immediate source of charges and therefore give the provider a large degree of freedom as far as their accuracy is concerned.

4.3. Cumulus Points and CP Thresholds

Let us now take a closer look to the feedback mechanism, i.e. the assignment of “Cumulus Points” CP. Following the CPC, the factual usage may not match the prediction given by the user (for whatever reason, be it e.g. an incorrect statement, changing habits, or new applications). As soon as these discrepancies exceed some threshold, the user receives regular (e.g. weekly or monthly) feedback in terms of the mentioned CPs. They exist as red and green flags: a red CP indicates that the user has been

overusing her capacities, a green one indicates the opposite, i.e. that the user might have been allowed to use more resources than she actually did. The larger the discrepancy between contract and reality, the more CPs may be assigned. CPs remain valid for a dedicated number of consecutive billing periods, and it is their accumulation that finally triggers certain consequences. Hence, receiving CPs requires no immediate reaction. However their successive accumulation over consecutive billing periods eventually may exceed a CP threshold and have consequences for the user, depending on ISP policies.

CP thresholds are defined formally in [15] and determine the assignment of Cumulus Points due to over- or underusing the resource with respect to the initial CPC. They may be derived in several ways:

- In the traditional version of [15], the violation of a CP threshold is interpreted as signal that the CPC is out of balance. In this sense, assigning a CP can be based on some sort of statistical test whether a current deviation of the actual resource usage from its agreed mean value is statistically significant or still lies within the borders of random fluctuation. A very straightforward approach [17] assumes e.g. the parameter considered (like bandwidth or volume) to be normally distributed with mean μ and standard deviation σ , and proposes relative σ -based CP thresholds $\vartheta_{(\sigma)} = 1 \pm \gamma \cdot \frac{\sigma}{x}$ with $\gamma \in \{1.3, 2.4, 3.1\}$, with a deviation from the mean by 1.3σ yielding one CP, by 2.4σ yielding two CPs etc. As 90% of random fluctuations of a normally distributed random variable lie within the interval $[\mu-1.3\sigma; \mu+1.3\sigma]$, 99% within $[\mu-2.4\sigma; \mu+2.4\sigma]$ and 99.9% within $[\mu-3.1\sigma; \mu+3.1\sigma]$, this is a very intuitive and uncomplicated approach to determine whether the contract still is in balance (i.e. larger fluctuations are still within the bounds determined by a normal distribution) or not.
- Whereas the first approach works well for the case of bandwidth or volume, qualitative QoS parameters like delay or jitter may require a different approach, as in this case the rough assumption of a normal distribution may no longer be justified (e.g. there is always a minimal delay between any two points in a network). Therefore, we now propose a new strategy for CP threshold determination based on user utility functions as introduced in Section 2. If we know (or at least are able to estimate) the utility function of a certain QoS parameter and its desired value as specified in the CPC, we may use any observation of the parameter to determine the utility of the observed value for the user and describe CP thresholds in term of this observed utility. One could e.g. agree that in case the utility of the actual QoS parameter value deviates by more than 10% from the specified value, one CP is assigned. For a deviation by 25% we can assign two CPs and for 50% three CPs, e.g. This approach respects far better the users' real needs and moreover is capable of dealing with hard QoS constraints as well.

Formalizing the second approach leads to the new concept of a “utility-induced QoS metric” as follows: assume $U(x)$ to be the utility function of QoS parameter x . Then, multiplying the utility $U(x_0)$ of the parameter value x_0 by an arbitrary factor α corresponds to a parameter value x_1 whose distance $\delta = |x_0 - x_1|$ equals

$$\delta = \left| x_0 - U^{-1}(\alpha U(x_0)) \right| \tag{4}$$

under very general conditions on the invertibility of the utility function. In terms of relative CP thresholds, this yields immediately the generalization of ϑ towards a non-constant function $\vartheta_{(U)}(x)$ depending on the QoS parameter x . Thus, (4) can be reformulated as

$$\vartheta_{(U)}(x) = \frac{|U^{-1}(\alpha U(x))|}{x} \tag{5}$$

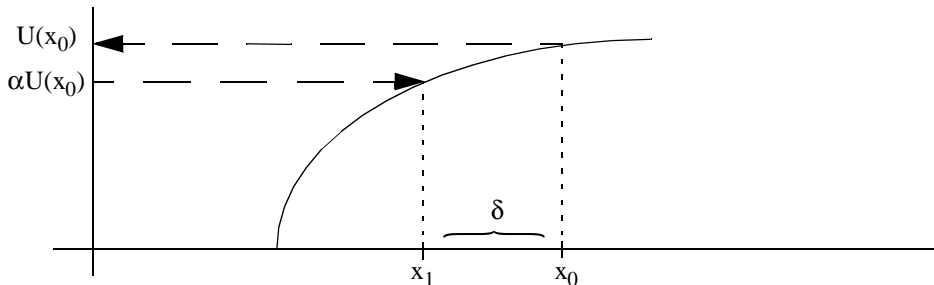


Figure 5. Utility-induced QoS Metrics

Figure 5 illustrates the concept of a utility-induced QoS metric. Note that for a linear utility function, (5) reduces to $\vartheta_{(U)}(x) = \alpha$, whereas e.g. for the common assumption of a logarithmic utility function $U(x) = \log x$ as characteristic for elastic traffic [11], we get

$$\vartheta_{(U)}(x) = \frac{1}{x}(e^{\alpha \log x}) = x^{\alpha-1}. \quad (6)$$

Summarizing this section, we have introduced the traditional Cumulus Pricing Scheme before examining it with respect to general QoS parameters. As a result, the concept of constant σ -dependent CP thresholds has been refined towards CP threshold functions whose form through (5) depends explicitly on the user's utility function. Based on these extensions, the following section investigates the application of CPS to scenarios with traffic characterized by one- or multi-dimensional QoS vectors.

5. CPS FOR ONE- AND MULTI-DIMENSIONAL QOS

Using the notation introduced by (3), traditional CPS can be formulated as triad

$$\tau_0 = (\beta, \dots, p) \quad (7)$$

where β denotes the bandwidth and $p = p(\beta)$ corresponds to the tariff function p as investigated in [17], i.e. a function that determines the flat rate to be paid by the user as well as excess charges (eventually to be paid as a result of the CPC renegotiation). Note that no further QoS parameters are considered, i.e. the mapping G is left open or equivalently set to zero.

There are three possible ways to include the QoS vector \vec{q} into the scheme. All of them depend on applying a suitable metric to (3), but differ with respect to where this metric is located:

- **Case A:** We use an approach as suggested by (2) and apply a direct transformation to the QoS vector. Imagine for instance that the CPC specifies traffic characterized by bandwidth β , throughput T , latency d and loss probability L . Then we could summarize these four parameters as

$$V(\pi) = \frac{T(\pi)}{L(\pi) \cdot d(\pi)} \cdot \beta(\pi). \quad (8)$$

This equation mirrors exactly the simultaneous requirement for high bandwidth and throughput and low latency and loss. Therefore, V may be interpreted as a single (real) number that describes the overall QoS of the service without going into details for the individual QoS parameters. In this case, every measurement within the CPS includes the full number of parameters, calculates the resulting actual V and compares it to the CPC specification, using e.g. the difference

$$|V - V(\pi)| \quad (9)$$

as metric. If the difference exceeds the CP thresholds, Cumulus Points are assigned. Note that of course there is a couple of possibilities for merging all QoS parameters into a single value, (8) representing but one example.

- **Case B:** Here, bandwidth/volume measurements are separated and independent from the measurement of the other (“qualitative”) QoS parameters. In this case, an excessive deviation from the agreed bandwidth may nevertheless yield no CPs, if the QoS as offered by the ISP and measured independently is worse than agreed upon. Therefore, for the triad τ a two-dimensional metric

$$\|(\beta, \vec{q})\| = \|\beta\| - \|\vec{q}\| \quad (10)$$

has to be designed that treats the two parameter types individually, where the metric $\|\cdot\|$ represents the result of applying CPS to the respective parameter in terms of assigned Cumulus Points. Figure 6 presents an example where excessive bandwidth consumption on the user side yields one (red) CP, whereas the bad QoS offered by the ISP forces assigning an additional Cumulus Point (which has also to be red from the ISP perspective and therefore has to be subtracted from the “bandwidth CP”; alternatively one could formulate (10) as a sum and denote as usually green CPs as negative numbers).

- **Case C:** Of course, one can go still into more detail and treat each individual parameter separately. In this case, for bandwidth, delay, jitter etc. there is run one individual CPS each, determining a weighted sum of CPs at the end. This approach allows assigning individual weights to specific QoS parameters in a very simple way due to the metric

$$\|(\beta, \vec{q})\| = \|\beta\|_{(0)} - \sum \xi_i \|q_i\|_{(i)} = \|\beta\| - \xi_1 \|d\|_{(1)} - \xi_2 \|\Delta d\|_{(2)} - \xi_3 \|L\|_{(3)} - \xi_4 \|T\|_{(4)}, \quad (11)$$

where the index for each metric indicates that different CPS parameters are used for different QoS parameters.

Summarizing shortly, these cases are linked to the following structural difference: Case A is the straightforward approach that treats bandwidth and QoS parameters equally, merges their information into a single parameter $V(\pi)$ (respecting the mutual relationships between all input parameters) and compares this common number to the CPC specification. Case B, however, takes regard to the asymmetry between bandwidth and the rest of the QoS parameters as discussed at the end of Section 3. Traditional CPS is used to determine whether the bandwidth thresholds bite and Cumulus Points are assigned. At the same time, also the other QoS parameters are compared to their individual specifications within the CPC in order to determine whether Cumulus Points must be assigned for them, too. In this way, the user may exceed her specified bandwidth and get e.g. a red CP, but at the same time delay and loss as delivered by the network are far too high and justifies the independent assignment of a green CP, with a resulting net assignment of zero CPs. Had the user stuck to her specification, she would be assigned a green CP for the bad quality delivered by the network. Finally, case C represents the finest level of detail, applying an individual CPS for every parameter and weighting the result according to the constraints of the respective applications, like the ones introduced at the end of Section 2. As we have seen there, e.g. for interactive audio and video applications, delay and loss are the most important QoS parameters and should be have considerable weight therefore, yielding $\xi_1 = 0.6$, $\xi_3 = 0.4$ and $\xi_2 = \xi_4 = 0$ as an example. Ftp traffic, on the other hand, has only small requirements in terms of throughput, hence $\xi_1 = \xi_2 = \xi_3 = 0$ and $\xi_4 = 0.2$ might be an appropriate choice for that case.

Apparently, the treatment of a significant number of single CPSs has to be regarded as a major drawback of case C. Therefore, in the presence of users desiring a simple and transparent scheme, case B appears to be a good compromise between the “U” and “N” requirements in the NUT trilemma. Figure 6 illustrates the difference between the three cases.

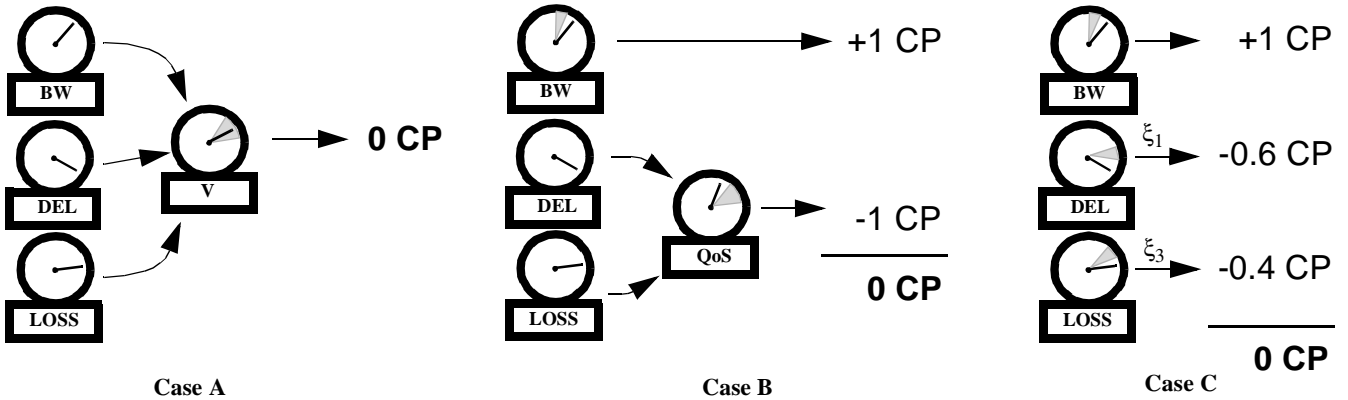


Figure 6. Three Metrics for Multi-dimensional CPS

6. MULTIDIMENSIONAL QOS IN AN INTERNET CHARGING SYSTEM

Existing charging systems perform the setting of prices, the function of charge calculation, and billing, which are integrated often in a monolithic manner. Additionally the maintenance of service classes, user profiles, customer data, identities, and banking account data may be included. Therefore, future charging systems need to be able to integrate different charging and accounting records, e.g., Internet Protocol Detail Records (IPDR) [6], since customer’s demand is determined by the so-called “one-stop billing” approach [21]. In addition, various pricing schemes for different services and their multi-dimensional QoS characterization have to be supported efficiently. Based on current investigations, in general, charging systems require to support the following tasks: (1) perform transport, service, and content charging; (2) perform accounting tasks according to transport and multi-service definitions; (3) support different levels of charging security; and (4) support auditing. The Internet Charging System (ICS), as described in [22], [23], and [24], provides a suitable technology to perform charging tasks for Internet services, while security requirements are being worked on. The major advantage of the ICS in support of multi-dimensional

QoS charging is based on the fact that the ICS interfaces may carry any number of QoS parameters measured from an underlying metering component, such as NeTraMet or NetFlow. Therefore, the data format to be defined for an accounting record is flexible. In addition, depending on the tariff to be applied, the set of required measurements, periodically or statistically taken, will be stored within ICS-internal databases.

7. SUMMARY AND OUTLOOK

This paper has dealt with the question of how to design a pricing scheme for differentiated Internet services that can deal with general QoS specifications, especially for the case of a multi-dimensional QoS vector. Besides discussing a couple of fundamental aspects generally related to this emerging new type of pricing schemes, it has been demonstrated that the Cumulus Pricing concept has been designed in a way that allows a rather uncomplicated extension to this new situation. Besides investigating the concept of CP thresholds that depend on the QoS parameter and introducing the notion of a utility-induced metric, a couple of ways have been described for merging the information of the QoS into the framework of SLA-based Cumulus Pricing Contracts CPC. Finally, a discussion of important aspects concerning the implementation of this extended scheme into an Internet Charging Systems has been provided.

As a general result, CPS offers eventually an Internet pricing approach that appears to be closer to the optimal tradeoff between the requirement types described in the NUT trilemma than any other of the established proposals. Moreover, in the form presented here, CPS is one of the very rare pricing mechanisms that are able to deal with QoS specified not only in terms of bandwidth or volume. Current and future work is heading towards two directions: on the one hand, a couple of theoretical issues of CPS needs to be further developed. An example for these issues is the form of the tariff function that is used to derive the flat rate to be charged from general QoS parameters. In [17], it has been shown that for the case of bandwidth, this function has to meet a one-over-square-root law in order to cause the user to specify her expectations for the required resources correctly. The rationale provided there needs to be refined for the general and multi-dimensional case. On the other hand, the implementation of CPS into the M3I platform as currently performed will allow a detailed technical and economic evaluation of this scheme including the reaction of real users to this new pricing mechanism for differentiated Internet services.

ACKNOWLEDGEMENTS

This work has been performed partially in the framework of the EU IST project Market Managed Multi-service Internet (M3I, IST-1999-11429), where ETH Zürich has been funded by the Swiss Bundesministerium für Bildung und Wissenschaft, Bern (No. 99.0536) and has been partially funded within the framework of the Austrian Kplus Competence Center Programme. The authors would like to express many thanks to their EU project partners in M3I.

REFERENCES

- [1] G. Almes, S. Kalidindi, M. Zekauskas: *A One-way Delay Metric for IPPM*. RFC 2679, September 1999.
- [2] G. Almes, S. Kalidindi, M. Zekauskas: *A One-way Packet Loss Metric for IPPM*. RFC 2680, September 1999
- [3] J. Altmann, B. Rupp, P. P. Varaiya: *Effects of Pricing on Internet User Behavior*; Netnomics, Vol. 3, No. 1, pp 67-84, June 2001.
- [4] G. Cheliotis: *A Market-Based Model of Bandwidth and a New Approach to End-to-End Path Computation with Quality Guarantees*. Workshop on Internet Service Quality Economics. Cambridge (MA), Dec. 1999.
- [5] L. Costa, S. Fdida, and O. Duarte: *Distance-vector QoS-based Routing with Three Metrics*. IFIP Networking 2000 / HPN - High Performance Networking, May 2000.
- [6] S.A. Cotton (edt.): *Network Data Management – Usage (NDM-U) for IP-Based Services*; IPDR Specification V 1.1, June 2000.
- [7] M. Falkner, M. Devetsikiotis, I. Lambadaris: *An Overview of Pricing Concepts for Broadband IP Networks*; IEEE Communications Surveys, 2nd Quarter 2000, pp 2-13.
- [8] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski: *Assured Forwarding PHB Group*. RFC 2597, June 1999.
- [9] V. Jacobson, K. Nichols, K. Poduri: *An Expedited Forwarding PHB*. RFC 2598, June 1999.
- [10] F. P. Kelly: *Charging and Accounting for Bursty Connections*. In: McKnight/Bailey (eds.): *Internet Economics*. MIT Press 1997, 253-278.
- [11] F. P. Kelly: *Charging and rate control for elastic traffic*. European Transactions on Telecommunications, vol. 8 (1997), 33-37.

- [12] K. Kilkki: *Service Differentiation in the Mobile Internet*. Tutorial 14, IEEE International Conference on Communications ICC01, Helsinki, June 2001.
- [13] M3I: *Market Managed Multi-service Internet*; 5th Framework EU Project, IST Program, No. 11429, <http://www.m3i.org>, May 2001.
- [14] V. Paxson et al.: *Framework for IP Performance Metrics*. RFC 2330, May 1998.
- [15] P. Reichl, P. Flury, J. Gerke, B. Stiller: *How to Overcome the Feasibility Problem for Tariffing Internet Services: The Cumulus Pricing Scheme*; IEEE International Conference on Communications, Helsinki, Finland, June 11-14, 2001.
- [16] P. Reichl, B. Stiller: *Nil nove sub sole? Why Internet Tariff Schemes Look Like as They do*; 4th Internet Economics Workshop Berlin (IEW'01), Berlin, Germany, May 25-26, 2001.
- [17] P. Reichl, B. Stiller: *Edge Pricing in Space and Time: Theoretical and Practical Aspects of the Cumulus Pricing Scheme*. Proceedings of the 17th International Teletraffic Conference, ITC-17, Salvador da Bahia, Brazil, September 2001.
- [18] J. Roberts, U. Mocci, J. Virtamo: *Broadband Network Traffic*. Final Report of Cost 242, Section 1.1, Springer Publishing, 1996.
- [19] S. Salsano et al.: *Specification of Traffic Handling for the first Trial*. Aquila Deliverable 1301, <http://www-st.inf.tu-dresden.de/aquila>, 2000.
- [20] B. Stiller, T. Braun, M. Günter, B. Plattner: *The CATI Project: Charging and Accounting Technology for the Internet*; 5th European Conference on Multimedia Applications, Services, and Techniques (ECMAST'99), Madrid, Spain, May 26-28, 1999, Lecture Note on Computer Science, Springer Verlag, Heidelberg, Vol. 1629, pp 281-296.
- [21] B. Stiller, G. Fankhauser, N. Weiler, B. Plattner: *Charging and Accounting for Integrated Internet Services - State of the Art, Problems, and Trends*; The Internet Summit (INET'98), Geneva, Switzerland, July 21-24, 1998, Session Commerce and Finance, Track 3.
- [22] B. Stiller, J. Gerke, Hasan, P. Reichl, P. Flury: *Charging for Differentiated Internet Services*; SPIE's International Symposium on the Convergence of Information Technologies and Communications (ITCOM 2001), Vol. 4526, Denver, Colorado, U.S.A., August 19-24, 2001.
- [23] B. Stiller, J. Gerke, P. Reichl, P. Flury: *Management of Differentiated Services Usage by the Cumulus Pricing Scheme and a Generic Internet Charging System*; 7th IEEE/IFIP Symposium on Integrated Network Management (IM'2001), Seattle, Washington, U.S.A., May 14-17, 2001, pp 93-106.
- [24] B. Stiller, P. Reichl, J. Gerke, P. Flury: *A Generic and Modular Internet Charging System for the Cumulus Pricing Scheme*; to appear: Journal of Network and Systems Management, Vol. 9, No. 3, September 2001.
- [25] B. Stiller, P. Reichl, S. Leinen: *Pricing and Cost Recovery for Internet Services: Practical Review, Classification and Application of Relevant Models*. NETNOMICS - Economic Research and Electronic Networking, vol. 3 No. 1, March 2001.
- [26] Z. Whang, J. Crowcroft: *Quality of Service Routing for Supporting Multimedia Applications*. Journal on Selected Areas in Communications, Vol. 14, No. 7, Sept. 1996.
- [27] J. Wroclawski: *The Use of RSVP with IETF Integrated Services*; Internet Engineering Task Force, RFC 2210, September 1997.