Retrospective Pricing Models for Internet Services: Solving the Flat Rate Dilemma “à la Flensburg”

Peter Reichl[1], Burkhard Stiller[2], Hongguang Ma[2], Hasan[2], Jan Gerke[2], Placi Flury[2]
[1]FTW Forschungszentrum Telekommunikation Wien, Vienna, Austria
[2]Computer Engineering and Networks Laboratory TIK, Swiss Federal Institute of Technology, ETH Zurich, Switzerland
E-Mail: reichl@ftw.at, [stiller|ma|hasan|gerke|flury]@tik.ee.ethz.ch

Abstract

The question of how to provide Quality of Service (QoS) has turned out to be of central importance for designing the Internet services of the future and is intimately linked with the problem of how to charge for QoS. Technically, the Differentiated Services (DiffServ) architecture for Internet backbones offers a suitable technology support, whereas appropriate and technically feasible pricing schemes provide an incentive for customers to choose efficiently among these various service classes. This paper is devoted to recent developments in this area, focussing on retrospective aspects of some currently rather popular pricing schemes. Especially the Cumulus Pricing Scheme (CPS) is investigated in great detail, with a focus on deriving numerical parameter values for thresholds and prices of the various service classes as part of the Service Level Agreements (SLA) between providers and customers. Moreover, a CPS-specific simulation tool is introduced, and various simulation results supplement the evaluation of this promising new Internet pricing scheme.

Keywords: Internet Charging System, Differentiated Services, Pricing Models, Cumulus Pricing Scheme.

1 Introduction

Despite of numerous proposals for charging Interent services in an economic efficient way [30], the classical flat rate scheme is still the most popular one in terms of practical application. It is undisputable that flat rate pricing has two tremendous advantages: (i) In terms of technical realization, flat rate schemes do not require complicated monitoring, mediating and measuring mechanisms. Instead, after setting up an initial traffic contract, the whole charging procedure reduces to billing a constant charge per month. The absence of expensive requirements for measuring detailed data about resource consumption is especially helpful for solving scalability issues related to the Differentiated Services (DiffServ) architecture [1] for Internet backbones which currently is the most promising candidate for a suitable technology in support of multiple service classes and Quality-of-Service (QoS). (ii) In terms of user acceptance, flat rate schemes are very popular, too, as they are easy to handle and to communicate, minimize the incentive to legal and non-legal disputes about billing issues and have also proven to be well-accepted by all sorts of Internet users, as e.g. the extensive investigations performed by the INDEX [11] experiment have demonstrated.

On the other hand, it is equally indisputable that more and more Internet Service Providers (ISPs) offering flat rate products are getting into trouble due to excessive overusage of the offered services. Among recent names here are Breathe (a major British ISP), T-Online (of Deutsche Telecom) and many more. In all these cases, the basic problem always comes from the obvious fact that flat rates do not provide any incentive to use Internet services in an economically efficient way.

In [24], this situation has been generally described as the "Feasibility Problem of Internet Tariffing": There, it has been demonstrated that each Internet Tariff Scheme has to balance out three different types of requirements, i.e. technical feasibility, economic efficiency and user acceptance, where admittedly the criterion of technical feasibility outweighs the other two: any suitable pricing scheme first of all has to be technical feasible, afterwards a trade-off between user acceptance (i.e. transparency, predictability, simplicity etc.) and economic efficiency (i.e. maximizing throughput, revenues etc.) may take place.

This paper starts from a somewhat different point of view. Instead of prioritizing the technical aspect, we propose to start the design of a suitable pricing scheme for differentiated Internet services from the user. In this sense, we aim at schemes that incorporate concepts that are well-accepted by the user and modify them in order to fulfill the mentioned tradeoff.

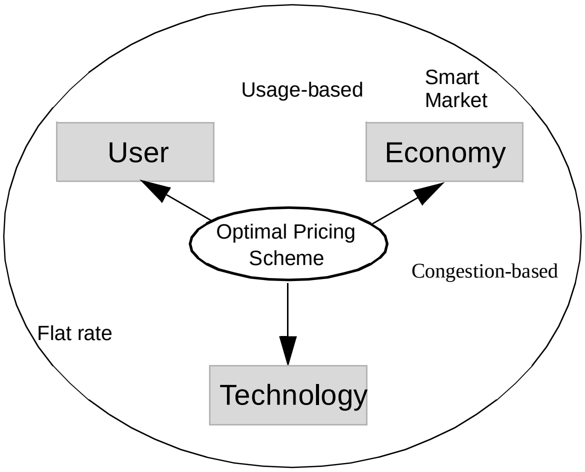


Figure 1: The "Space of Pricing Schemes"

Figure 1 illustrates this approach in terms of the three-dimensional "space of pricing schemes". Whereas a hypothetical "optimal Internet tariff" would be located in the centre of the competing forces of the three mentioned requirement types, the distance of the different existing proposals indicates how successful they perform in balancing out the tradeoff between technological feasibility, economic efficiency and user acceptance. Thus, auctions e.g. provide a very close feedback about the actual market situation, but are quite complicated for the user (as she has to estimate her utility function, design a bidding strategy, offer bids at various places etc.) as well as for the implementation. Usage-based approaches are more user-friendly (as it is widely accepted that the charge for any good has some relationship with delivered quantity) and provide also some economic incentives, but are complicated to realize. Congestion-based pricing (e.g. by using the ECN bit in the TCP/IP header) uses existing Internet technology and derives its economic efficiency from the fact that it provides incentives to avoid congestion, but is very complicated to communicate to the user etc. Investigating the related proposals in this way demonstrates that any proposed solution is far from being optimal, and indicates in the same moment the direction into which any amendment of existing proposals should aim at.

This paper is organized as follows. Based on a brief review of related work (Section 2) and an introductory survey of the DiffServ architecture (Section 3), the possibilities and limitations of amending pure flat rate schemes are discussed. In Section 4, it is demonstrated how the well-known central record of German traffic offenders located in the city of Flensburg has stimulated an example of restrospective tariffs that turns out to be a significant improvement of flat rate without losing any of its main advantages. As a further main contribution of this paper, a simulation tool designed for this so-called "Cumulus Pricing Scheme" together with simulation results is presented in Section 5. Finally, Section 6 provides a concluding discussion.

2 Related Work

Although it has been argued that pricing in general and usage-based pricing in particular can impose a high overhead on telecommunication systems [15], [25], any form of usage-based pricing for Internet services still is of significant relevance, since underlying resources (such as satellites, frequencies, cables, routers/switches, and most notably operating personnel) are scarce and costly, waste of resources is unacceptable, unfairness exists among users, and ISPs eventually may lose revenues. Flat-fee pricing models for multiple services show the economic draw-back of not being incentive-compatible [5], [7], [14], and [25]. Therefore, [10] proposes a per-flow billing system for TCP (Transmission Control Protocol) flows. Advanced per-flow charging and accounting approaches based on reservations have been tackled in [9], [12], and [17]. Multi-service packet networks are proposed by adding classification and scheduling to routers, but not policing [3]. As far as pricing for multi-service networks is concerned in general, [11] indicates that users are not disinclined to flexible pricing models. Finally, the project M3I (Market-Managed Multi-service Internet) [19] aims at designing and implementing a next-generation system enabling Internet resource management through market forces, specifically by enabling differential charging for multiple service levels.

3 Differentiated Services in the Internet

The traditional Internet has been designed with a best-effort service. With rising Quality-of-Service (QoS) requirements of applications, several proposals have been submitted to meliorate this heritage. Two approaches are the Integrated Services (IntServ) [2] and the Differentiated Services (DiffServ) Architecture [1]. IntServ and the Resource Reservation Protocol (RSVP) allow for highly tailored QoS specifications and precise resource reservations to support specified QoS. Consequently, every application flow needs to be managed individually, and as a result IntServ lacks of proper scalability for large-scale backbone networks [31]. The DiffServ approach [1] avoids these disadvantages by shifting complexity to network edges and by the discretization of QoS specifications into some few classes. Tasks at edges comprise admission control, traffic conditioning, charging, and accounting facilities.

3.1 Basic DiffServ Concepts

In DiffServ single application flows are assigned and accumulated to Behavior Aggregates (BAs) at the edges of the DiffServ domain. Packets belonging to a specific BA are identified by so-called DiffServ Code Points (DSCPs). The association and enforcement of the BA’s QoS is fulfilled by specific forwarding policies, the so-called Per Hop Behaviors (PHBs). Services in terms of network delivery characteristics can, thus, be described by BAs and their associated PHBs. Figure 2 depicts a situation, where four different flows (Flow 1 to 4) are accumulated into two different BAs. Since the DiffServ Border Router and the network dispose of only a limited amount of resources, flow set up has to be requested, and accepted flows have to comply to specified traffic characteristics. The access request, implying the commonly known admission control and conditioning of traffic, according to an allegation, i.e. a contract, demand for signaling and for appropriate contracts negotiations. The status quo of today’s DiffServ investigations [1] targets at the central Bandwidth Broker (BB) [32], which is responsible for signaling and admission control and configures a traffic conditioning entity at involved border router(s). The criteria for flow acceptance or rejection and for setting traffic conditioning parameters are deduced from agreements between the customer and the service provider.

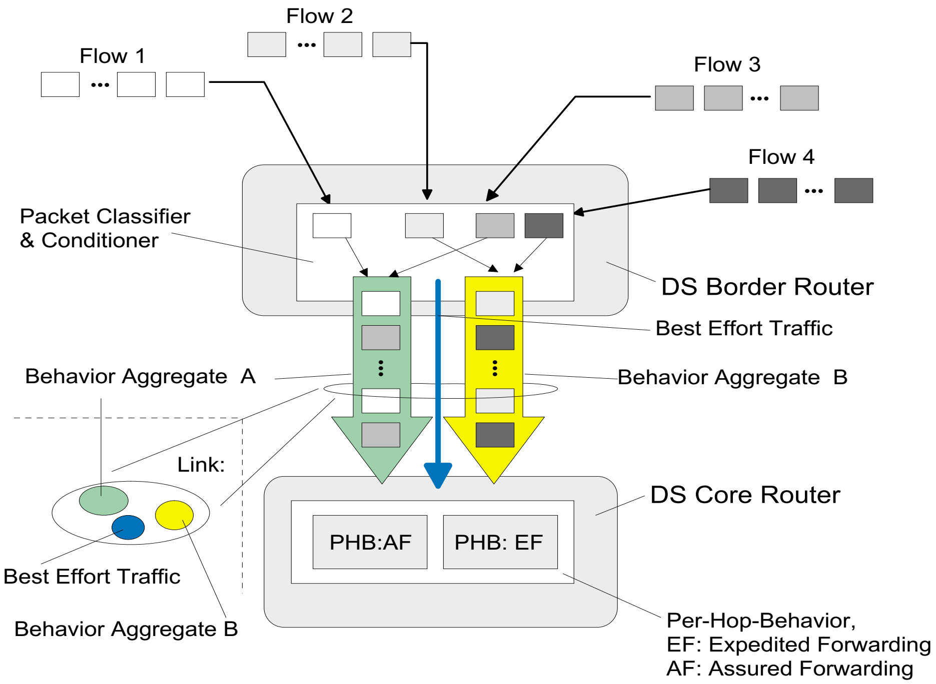


Figure 2: Flow Conditioning, Aggregation, and Per-Hop Behavior

3.2 DiffServ Service Classes

DiffServ quantifies services to a small number of different classes. Applications need to meet their QoS requirements by choosing from this pool of service classes. These different service classes either result from the combination of BAs and their PHBs or are simply best-effort services. In the current DiffServ specification two PHBs, i.e. Expedited Forwarding (EF) and Assured Forwarding (AF), are distinguished besides the best-effort class. EF provides facilities to create services with low time constraints, *e.g.*, low loss, low latency, low jitter, and with assured bandwidth, so that EF is often said to have the features of a *leased line*, i.e. a virtual leased line in a single DiffServ domain. Nevertheless, EF does not optimally satisfy all fields of application, since it provides only a rather weak support of bursts. Instead, AF is in fact able to cope with almost any bursty behavior, however, at the expense of the delivery of IP packets, which only can be guaranteed with certain high probabilities. Furthermore, several AF classes can be distinguished to allow for a fine-grained handling of different application requirements. Services and applications with ordinary QoS requirements preferably continue to deploy best-effort forwarding. Note that best-effort in DiffServ subtly differs from best-effort of the early days of the Internet, since congestion control mechanisms of TCP are undermined by traffic shaping mechanisms of DiffServ, *e.g.*, AF packets may be degraded to best-effort packets in case of congestion. Thus, best-effort generated from DiffServ congestion policies preempts ordinary application best-effort.

4 Pricing Differentiated Services

4.1 Charging Considered Necessary

Charging in general may be viewed as consequence of the simple fact that users are not alike. They differ with respect to personal judgement of QoS, to personal satisfaction with a given QoS, to their applications with various QoS requirements, and to their willingness-to-pay for a service. Therefore, users will utilize different QoS classes, if those are offered and priced suitably. As initial results indicate [11], users will not use similar service classes all the time, but change service classes depending on their application in use, on the current QoS they perceive, and on their current mood. For instance, a professional Internet user transmitting high quality real-time video streams over the Internet shows very specific requirements. She also has an estimation of how often she uses the network for such transmissions and can specify these requirements in an aggregate level, *e.g.*, within an SLA, which has been negotiated with her provider. However, these estimations will not be always consistent with the real traffic. During vacations, the user will produce less or no traffic at all, whereas periods of intensive working may involve more video transmissions than usual, thus excessing a contracted traffic threshold. If the amount of her traffic misses her estimations frequently, the contract needs to be adapted.

Concerning the destination of such traffic, it does not necessarily sink in her access provider’s domain, but traverses through to other domains. Therefore, just as a single customer has an estimation of her data traffic, an ISP is able to provide a similar estimation based on the sum of all his customers. These customers can be either single users or other providers, who inject data into his domain. It is rather probable (due to statistical multiplexing gains) that errors in users’ estimations will not yield fatal consequences for the provider traffic, since smaller and larger estimations will balance out each other. Furthermore, it is an ISP’s policy to meter heavy users and change their contract and SLA in time.

4.2 The Concept of Retrospective Pricing

In contrast to the previous work , e.g. [24][28], here we view the solution of the Feasibility Problem from the perspective of customer’s expectation on a suitable charging scheme. Users want to be charged predictably, transparently and in a simple way. Taking these requirements serious excludes a couple of tariff proposals right away, including e.g. schemes that account for too fine details in the resource consumption (e.g. usage-/volume-based schemes), that use too complicated theoretical derivations (e.g. effective bandwidths), that require complicated interactions with the user (e.g. auctions) etc. In this sense, flat rate schemes indeed appear to be optimal from the user’s point of view. On the other hand, including any economic incentive inevitably requires a feedback to the actual situation in the network (e.g. load, congestion etc.).

This dilemma cannot be resolved easily in the current Internet tariffing frameworks. [22] sketches a possible resort by proposing the “Time Scales Methodology” which starts from a characterization of Internet tariff schemes in terms of relevant time-scales and extends this framework by introducing additional dimensions that may lead directly to new types of tariff schemes. One such possible new dimension concerns the reaction the user is required to perform according to actual feedback from the network. Especially the idea of delaying this reaction to a later moment has turned out to be rather fruitful. One prominent example of this type of “retrospective pricing” is Odlyzko’s “Paris Metro Pricing” PMP [21]. PMP is based on a fixed subdivision of the network into several independent subnetworks that deliver identical functionalities, but are charged differently. As a consequence, the more expensive subnetworks will be less crowded and therefore are expected to provide better QoS than the cheaper ones, but without giving any specific guarantees. User’s interests are respected in this scheme as the charge for the subnetwork (i.e. “traffic class”) chosen remains constant (and is balanced by changing QoS). PMP is a typical example of user reaction delay: initially, any user will chose the traffic class that she expects to suit best her demand. As there are no QoS guarantees available, only time can tell whether the QoS experienced is indeed sufficient to the user. Only retrospectively the user can decide whether the delivered QoS has been sufficient for her, and subsequently change the traffic class in case of future network use.

In this sense, PMP can be characterized as a flat rate scheme that shifts the trade-off of flat charging to changing QoS conditions. Therefore, user expectations are fulfilled in terms of the constant and transparent charge, but not with respect to the delivered QoS.

A second retrospective approach, the “Cumulus Pricing Scheme” as introduced in [24], shifts the “Black Peter” to the side of the ISP. Here, as usual with flate rate schemes, it is the ISP to pay the price for flat charging, but has the possibility to correct the initial contract with the customer in case this agreement turns out to be severely out of balance.

The rest of this paper deals with results of some recent investigations of CPS. After giving a brief survey, we demonstrate the parametrization of CPS applied to a DiffServ scenario specified in great detail. Afterwards, a simulation tool designed specifically for CPS is introduced, and a couple of simulation results are presented. Finally, we describe in detail a lot of specific open questions that are currently tackeled and will be hopefully resolved in the near future.

4.3 Introduction of CPS

The Cumulus Pricing Scheme (CPS) recently has been established in [24] and [29]. The driving rationale behind its design can be summarized as follows: Users generally appear to have a strong inclination towards flat-rate schemes, therefore CPS originally offers a flat tariff, which has to be agreed upon in an initial traffic contract between customer and ISP, the so-called “CP Contract” (CPC) that may show the form of an SLA. This CPC contracts the size of expected resource requirements. On the other hand, pure flat-rate schemes have turned out to be infeasible, as they do not include any economic incentives. Hence, CPS allows the charge rate to be changed, if there is an obvious imbalance between the CPC and the actual resource usage. Again, for reasons of user acceptance, this imbalance is signalled repeatedly to the user, using so-called Cumulus Points (CP), which indicate over- or underutilization of resources with respect to the initial requirement’s declaration and are typically assigned on a rather long time-scale, *e.g.*, monthly. Hence, the CP assignment takes place according to the so-called “CP Rule”, if the actual resource usage violates thresholds stated in the SLA, and the assignment should be based on monitoring of resource usage, where the manner in which measurements are performed is entirely up to the ISP. Receiving CPs requires no immediate reaction on the user side, i.e. the flat-rate remains stable, although the resource usage currently does not meet the original expectation. CPs appear in two flavors: red ones for over- and green ones for underutilization. They may be specified either as relative (i.e. as factor ϑ_i relative to the CPC specification) or absolute ones (*e.g.*, as overutilization in terms of θ_i resource units[1]). They are accumulated over time. Eventually, if this accumulation violates a different threshold, a second rule, the “Reaction Rule”, states the imbalance between declaration and actual usage and urges the customer and the ISP to start a CPC renegotiation.

As introduced formally in [24], CPS can be expressed with two rules related to different time-scales. Suppose that initially the customer has stated her expected resource requirements to be *x*, whereas the actual resource consumption is described by a function *V(t)* of time. Let $\Delta_i = \Delta(t_i)$ describe the monthly over- or underutilization of the customer with respect to her statement *x*, i.e.

$$\Delta_i = \int_{t_{i-1}}^{t_i} (V(t) - x)dt = \int_{t_{i-1}}^{t_i} V(t)dt - x(t_i - t_{i-1}) \tag{1}$$

where t_i describes the end of monitoring period *i*, *e.g.*, the end of month, *i* = 0,1,2... Cumulus Points are assigned by the ISP according to the following

---

[1] Note the tiny difference between the Greek letters.

CP Rule:

Define θ_n, $n = -N,\ldots,-1,0,1,2,\ldots,N$, *to be the CP thresholds,* $\theta_0 = 0$ *and* $\theta_{\pm(N+1)} = \pm\infty$ *where N describes the maximal number of CPs that could possibly be assigned for one monitoring period. Then for monitoring period i, the customer is assigned* c_i *cumulus points iff*

$$0 \le \theta_{c_i} \le \Delta_i < \theta_{c_i+1} \text{ or} \tag{2}$$

$$\theta_{c_i-1} < \Delta_i \le \theta_{c_i} \le 0, \tag{3}$$

the choice between (2) *and* (3) *depending on* $\operatorname{sgn}\Delta_i$.

Now the cumulus points c_i are accumulated over time according to

$$\Gamma_n = \sum_{i=1}^{n} c_i, \tag{4}$$

hence, Γ_n describes the total sum of cumulus points assigned since the start of the contract.

The reaction to CP accumulation is the content of a second rule, the so-called “Reaction Rule”:

Reaction Rule:

Define Θ *to be the reaction threshold. Then the contract between customer and ISP is in the state of imbalance and needs to be renegotiated after period n if*

$$|\Gamma_n| \ge \Theta. \tag{5}$$

The concept of accumulating feedback marks that are delivered in a rather coarse-grained manner has a long-standing and “successful” tradition especially in Germany or Switzerland. The most prominent example has already been mentioned: In the city of Flensburg in the very northern of Germany, a central record of German traffic offenders is maintained which is based on so-called “Flensburg points”. These points are awarded as soon as somebody does not behave on the road as he should and the policeman is just round the corner. The points are accumulated, and as soon as a certain threshold is exceeded within two years, the respective driver looses his licence. Another example comes from some major Swiss supermarket chains where with each buy you receive some points on your “Cumulus Card”, and excessing the threshold here allows you to get some extra price for special products.

Summarizing, the CPC fixes expected resource usage and a respective flat-rate. The actual consumption is monitored and in case of exceeding the initial specification, the user receives a monthly feedback in form of CPs. After a negotiated duration of *n* months, the accumulated balance of CPs may exceed the reaction threshold, thus, eventually requiring ISP and customer to adapt the CPC accordingly.

4.4 DiffServ Charged by CPS

Based on a realistic application and topology scenario for a single DiffServ domain, this section discusses how to derive particular parameter and threshold values for CPS with respect to Differentiated Service classes.

4.4.1 Application and Technology Scenario

Consider a set of example applications being supported in a single DiffServ domain as introduced in Section 3.1 and suppose it offers five different service levels, each of which is handled by a single queue per router. The best-effort service class (BE) supports adaptive applications, e.g. TCP connections. Each of these connections will try to utilize as much bandwidth as possible to gain the largest bandwidth share according to the TCP congestion control algorithm. As discussed in Section 3.2, the Expedited Forwarding (EF) class may support telnet applications to ensure that the end-to-end delay can be guaranteed and suitable reaction times are achieved for a customer. Each of these telnet applications generates Variable Bit Rate (VBR) traffic, which is substantially lower than the full rate reserved for EF. Additional applications may require EF, such as remotely collected sensor data or a board of executive's Voice-over-IP with 32 kbit/s (Platinum VoIP) conference. The unused bandwidth may be borrowed by other service classes. Three additional queues may exist for three types of Assured Forwarding (AF) classes. Types AF_1 and AF_2 support Constant Bit Rate (CBR) traffic in an on/off mode with varying demand between 75% and 125% of the reserved rate. This includes regular VoIP with 16 kbit/s CBR in class AF_1 as well as video-streaming of 2 Mbit/s as a CBR application for AF_2, moreover, class AF_3 is used for bursty MPEG-coded video, ranging from 1.8 Mbit/s to 8 Mbit/s VBR. If each of these five queues are assumed to be handled equally, they share the capacity of an STM-1 link (155 Mbit/s) equally as well, yielding $q^* = 31\text{ Mbit/s}$ per queue. Besides these service classes offered and the applications utilizing them at the edges of a DiffServ domain, charging and network technology is required to be set-up and configured accordingly.

Assuming a smaller DiffServ domain with four Border Routers (BR) and an unlimited number of internal routers, Figure 3 depicts the technology required to support DiffServ charging with a suitable charging platform (here for the case of the Internet Charge Calculation and Accounting System ICCAS as presented in [29]). A single ICCAS for a DiffServ domain supervises and allows for the configuration of BR's metering and mediation components. Depending on the size of the DiffServ domain, multiple ICCAS or components of it may exist in a replicated fashion to offer a scalable solution.

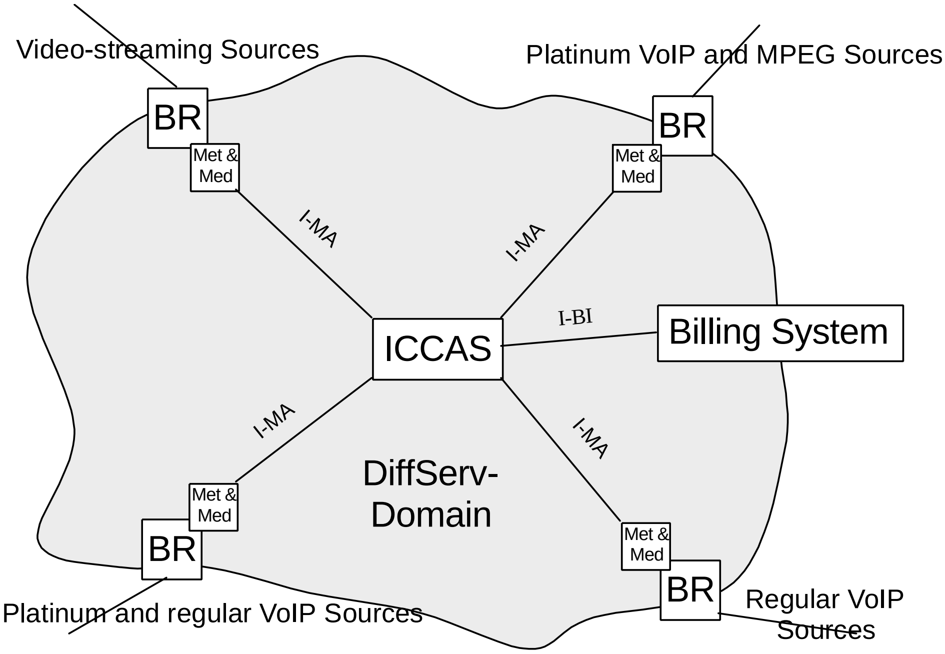


Figure 3: Transmission of Charging Data and Different Application Sources in a DiffServ Domain

4.4.2 CPS Parameter and Threshold Values

CPS is originally based on the concept of individual contracts between each customer and their ISPs. On one hand, these contracts contain (1) an expected resource requirement specification (*e.g.*, in terms of bandwidth), (2) a flat-rate related to this specification, and (3) thresholds allowing for the observation of a specification's validity. On the other hand, an ISP operating a DiffServ domain offers several standardized traffic classes with predefined QoS parameters, as, *e.g.*, described in the scenario above. Hence, the application of CPS particularly to this DiffServ scenario restricts the freedom in negotiating the CPC as DiffServ classes have to be taken into account, *e.g.*, in this example for best-effort, EF for telnet and platinum VoIP, AF_1 for regular VoIP, AF_2 for video-streaming, and AF_3 for MPEG-coded video. For each of these classes the customer has to specify the average resource requirements, *e.g.* as number of applications he expects to operate. This information is sufficient to fix the according CPC, especially the flat-rate for this expected resource usage and the various CPS thresholds. For reasons of simplicity, we will discuss a scheme involving two threshold for red CPs (ϑ_1 for light and ϑ_2 for heavy overutilization) as well as one "green" threshold ϑ_{-1}. These thresholds may be determined as follows.

Starting with AF_2, it is supposed that in this class CBR traffic is transmitted, which originates from video-streaming applications requiring 2 Mbit/s each and corresponding to a considerable share of the total bandwidth available for AF_2 traffic. Assume ξ to be the proportion between the bandwidth for one CBR application and the total capacity available in the respective AF class, i.e. for the above scenario $\xi(AF_2) = 2Mbps/31Mbps = 0{,}065$, moreover let n be the expected number of applications as specified in the CPC. As in this scenario the general demand for AF_2 traffic is supposed to vary between 75% and 125% of the reserved rate, it is reasonable to assume that this should be valid for the individual traffic, too. Therefore, it is proposed to set the relative CP thresholds ϑ_1 and ϑ_{-1} to 1.25 and 0.75, respectively, i.e., if the measured traffic exceeds the CPC specification by a factor of 1.25 or more, a red CP is assigned, etc. As long as the actual traffic lies within these limits, no CPs are distributed. *E.g.*, if a customer A books $n(A) = 5$ continuous video streaming "channels", he may well use sometimes six of them without afflicting the network as a whole (because the resulting overbooking of the AF_2 class merely reduces the available BE capacity), whereas customer B having specified only $n(B) = 1$ connection, but using two, should be assigned one red CP. If customer B runs a total of three applications, he should receive one more red CP (as excessing the specification by a factor of 3 certainly indicates heavy overutilization), yielding, *e.g.*, $\vartheta_2(A) = 2{,}1$. For customer A, this threshold must be defined smaller, because in her case running one more application does not correspond to a comparable relative increase of bandwidth requirements. Therefore it is proposed to set $\vartheta_2(A) = 1{,}8$ (heavy overutilization, if 9 instead of 5 applications are running, because in this case customer A already claims more than 50% of the total available AF_2 bandwidth, due to $\xi = 6{,}5\%$ being relatively high). Therefore, as a first result the value of a relative CP threshold results from a trade off between absolute and relative increase in resource usage, if the number of running applications is increased by one and, thus, depends on both ξ and n.

Due to the granularity of applications, CP thresholds for the AF_2 class appear to be relatively large (this may be compensated, *e.g.*, by a relatively high flat rate per Mbit/s). For AF_1 traffic (i.e. Platinum VoIP requiring CBR traffic of 32 kbit/s), the general situation is similar, but the granularity of these channels is much finer, as now $\xi \approx 0{,}1\%$. Therefore, the second (and maybe higher) thresholds can be set smaller than for the AF_2 case. Still, traffic varying from 75% to 125% of the reserved capacity does not harm the network, which leads to the setting of $\vartheta_1 = 1{,}25$ and $\vartheta_{-1} = 0{,}75$ again. For customer B ordering a single Platinum VoIP channel, $\vartheta_2(B)$ is again set to 2.1 (the argumentation that excessing the specification by a factor of more than 2 corresponds to high overutilization still is valid), but for customer A specifying $n = 100$ Platinum VoIP calls (*e.g.*, with traffic varying with standard deviation $\sigma = 10$) it is reasonable to apply the "σ-approach" sketched in [23], i.e. let $\vartheta_2(A)$ depend on σ, which, *e.g.*, yields $\vartheta_2(A) = 1.25 + \frac{\sigma}{n} = 1.35$ in this case.

Summarizing both AF cases formally, the following result is achieved: For ξ » 1%, i.e. running one application requires a significant share of the total available bandwidth q^*, ϑ_2 is monotonically decreasing with the number n of specified applications, i.e. in this example scenario from $\vartheta_2(n = 1) = 2{,}5$ to $\vartheta_2(n \approx q^*/\xi) \approx 1{,}25$, as in the latter case the customer's CPC already has claimed all available bandwidth of that class. Assuming a quadratic decrease, the rationale presented above for the case of this scenario presented yields finally a function for $\vartheta_2(n)$ as sketched in Figure 4.

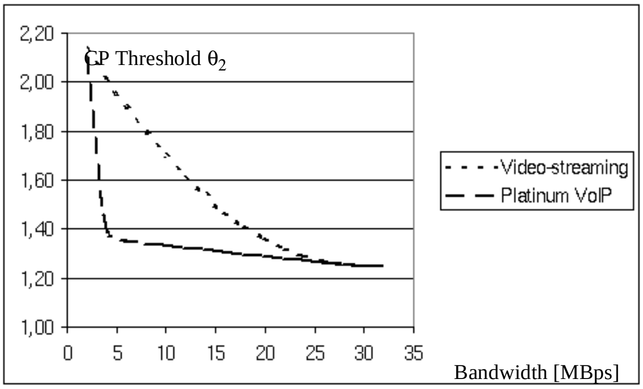


Figure 4: ϑ_2 as a Function of User's Consumption

For ξ « 1% different regimes are obtained. For small n, the crucial parameter still is the relative increase of resource requirement, hence again $\vartheta_2(n = 1) = 2{,}5$, whereas for moderate and high values of n, *e.g.*, $n \geq 0{,}1q^*$, the mentioned "σ-approach" may be used and it is assumed that the standard deviation follows the usual $\sqrt{n}$ law (cf. Figure 4).

The case of AF_3 requires a completely different approach, as thresholds have to deal with highly bursty VBR. In this case, it is useful to fall back on the "effective bandwidth approach" as introduced by [18]. Without going too much into details, the CPC traffic characterization now consists of effective and peak bandwidth x_e and x_p, respectively, and therefore, includes the degree of burstiness, i.e. extremely bursty traffic yields an effective bandwidth close to the peak rate of the traffic, whereas traffic without burstiness results in the traffic average to be the effective bandwidth. Then the flat rate is derived from the effective bandwidth. Again, $\vartheta_1 = 1{,}25$ and $\vartheta_{-1} = 0{,}75$ are obtained, whereas now ϑ_2 can be related to the peak bandwidth, *e.g.*, such as $\vartheta_2 = 1 + \frac{x_p - x_e}{x_e}$. Note, however, that these approaches proposed are still part of ongoing investigations.

Treating the EF case is much easier, especially if it is assumed that only VBR traffic to be assigned to this class. In this case, the entire threshold scheme can be performed based on fixed deviations of the expected mean bandwidth consumption as stated in the CPC, while it is aimed to keep the actually used total capacity rather close to the specified one. Thus, either all thresholds depend on a statistical value of standard deviation per typical demand or thresholds are set manually, *e.g.*, as $\vartheta_1 = 1.05$ and $\vartheta_2 = 1.1$.

Finally, BE is charged very little or even provided for free, whereas there are no QoS guarantees corresponding to this traffic class. Therefore, CP thresholds are set basically to yield a proportional fairness between all users, but can have comparably large values, *e.g.*, $\vartheta_1 = 10$ and $\vartheta_2 = 100$. Here, it is still helpful for the ISP to get an initial estimation on possible demand for capacity, but the customer hardly ever should be punished for transmitting more BE traffic than specified. However, BE traffic may be discarded in a network due to overload in higher service classes.

4.4.3 Implementation Technology

Besides these essential determinations of CPS parameters and thresholds, which are depending on the application scenario and its SLAs, the ICCAS can be supported by existing technology. There are several products available, which are able to fulfill a subset of charging system functions, while another important subset of functions has been implemented. A close networking hardware dependency is observed for the metering component. *E.g.*, NetFlow [6] is able to collect and transmit flow data, if they are metered at Cisco routers. The open software solution NeTraMet [4] performs metering functions for a desktop computer. To handle information gained by the metering and process mediation tasks, the Smart Internet Usage (SIU) package [26] can be applied. SIU is able to access NetFlow routers directly and can be adapted to access a NeTraMet meter via the Simple Network Management Protocol (SNMP). However, one of the major restrictions imposed by these systems is the non-existence of a real-time capability, required for short-termed, feedback-driven, and usage-based pricing schemes. Measuring data only in the time-scale of seconds or even minutes, however, is not a problem with the integration of the ICCAS, DiffServ, and CPS. While the CPS has been designed with the end-customer and providers in mind, it also offers the major advantage to be implementable with metering technology available today, since it merely presupposes the existence of tools and means for metering and accounting traffic data without posing any critical requirement on their granularity (even if it is straightforward that fine-granular measurements will yield more precise results). Therefore, these longer time-scales due to technical feasibility do not yield a problem for CPS. Finally, the following additional considerations are to be taken into account in order to set up effective databases for the scenario of DiffServ with CPS:

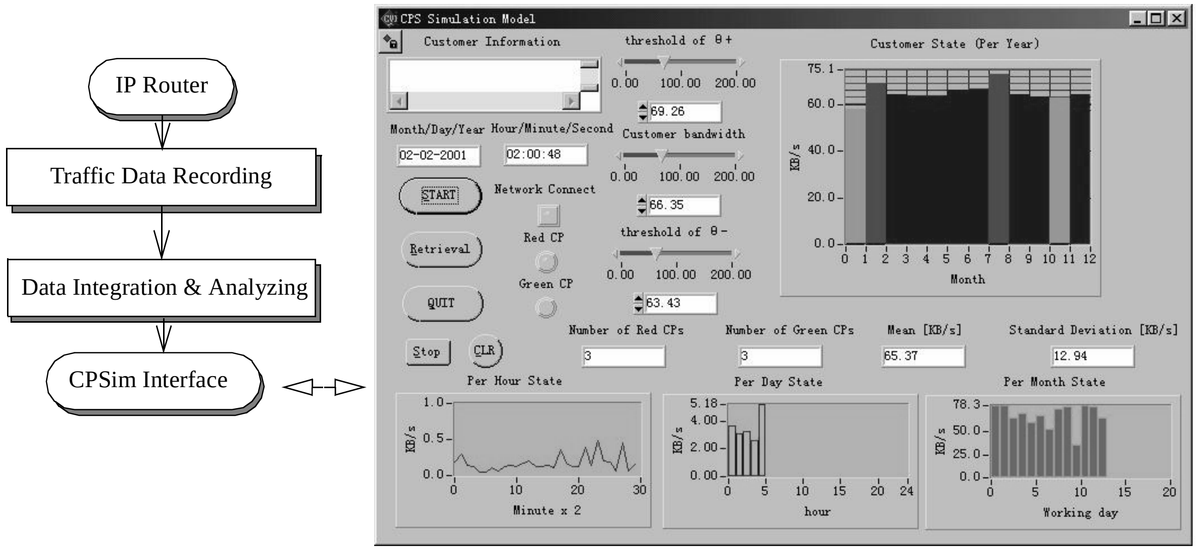
- The traffic descriptor that is used to identify a certain traffic of a session comprises of source and destination IP addresses, source and destination port numbers, and DSCP.
- Users can be assigned fixed IP addresses to map sessions to users, otherwise dynamic IP addresses have to be saved together with the user identification each time this user initiates a session.
- CPS thresholds are part of the contract information, which are to be stored together with the agreed upon service classes in the customer database.
- The resource usage to be measured for CPS, according to the ISP’s policy chosen, are traffic volume and duration.
- Cumulus Points are calculated on a per-customer basis and stored in the charging database.

5 CPSim - Design and Simulation Results

5.1 The Simulation Tool CPSim

CPSim is a simulation tool expecially designed for the further simulative investigation and evaluation of the Cumulus Pricing Scheme. It has been developed to serve two different purposes: (1) investigating CPS with respect to its general characteristics and (2) determining suitable quantitative numbers for its parametrization, *e.g.*, for the threshold levels. For that reason, two different forms of data are generated and processed, respectively: (1) Random traffic is generated by applying the generalized TES traffic model as introduced in [10], and (2) real traffic is obtained from an Internet router. Therefore, the simulation model CPSim acts as an evaluation and management tool for Internet traffic. It consists of the following three modules as depicted in Figure 5(a): (1) The Traffic Data Recording module (TDR) unit preprocesses router data, *e.g.*, obtaining a list of volume-based usage data in terms of transferred bytes. (2) The Data Integration and Analyzing module (DIA) implements all CPS rules and functions. While the latter one can be executed with any type of volume-based data, the former one needs access to an existing router. (3) Finally, the user interface (UI), as shown in Figure 5(b), acts as an interface for configuring thresholds, obtaining characteristics and statistics of red and green CP assignments.

In the UI, four graphs display traffic data on different time scales (minutes, hours, days and months). Four command buttons on the panel allow to manipulate the simulation model, whereas customer bandwidth and thresholds are controlled via respective sliders. For a simulation these parameters can be set either manually or automatically. Characteristics of customers will be computed during the simulation. These results, mean value and standard deviation of customer's data over one month, will be displayed in the respective data fields as well as the numbers of red and green CPs; if the latter ones exceed certain limits, additionally text messages may appear. The system timer inside the simulation model controls the integration length according to a calendar. For online data processing, the current date and time are displayed on two string indicators. Figure 5(b) shows as well the graphical representation of a simulation run for one year. The CP state over the year is depicted as a histogram using different colors: dark indicates the usage of bandwidth according to the initial traffic specification, grey indicates that the customer has used less traffic, whereas darkgrey represents an overutilization per month. Therefore, for this year 3 green and 3 red CPs have been assigned, and the traffic showed a mean value of 65.35 kbit/s as well as deviation of 12.94.



Month/Day/Year: 02-02-2001
Hour/Minute/Second: 02:00:48
threshold of θ+: 69.26
Customer bandwidth: 66.35
threshold of θ-: 63.43
Number of Red CPs: 3
Number of Green CPs: 3
Mean [KB/s]: 65.37
Standard Deviation [KB/s]: 12.94

Figure 5: (a) CPSim Tool Architecture and (b) User Interface of CPSim including an Example Result

5.2 Simulation Results

CPSim has been used to perform an exhaustive simulative evaluation of major CPS aspects. The following results are based on real network data traffic over 240 working days, i.e. one entire year. These data have been obtained from long-term SNMP measurements at ETH Zurich's LAN-WAN access routers

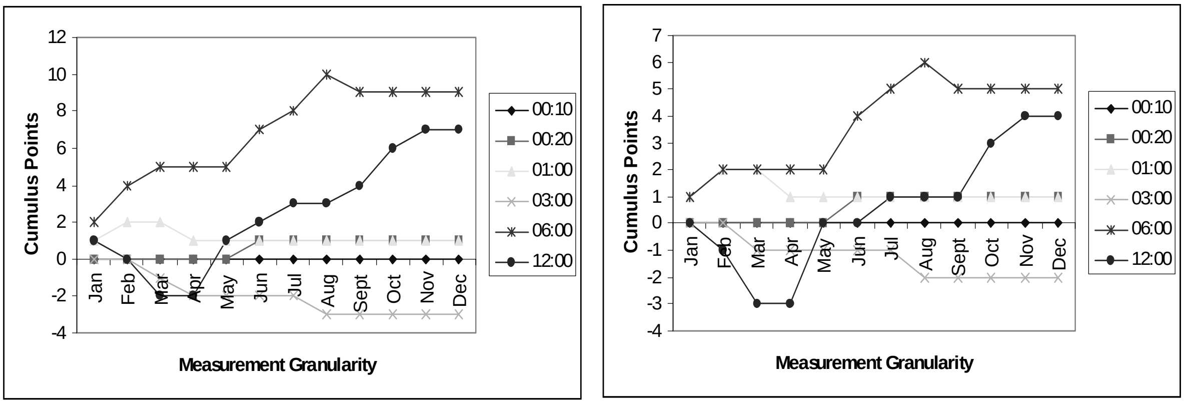


Figure 7: Stability of CPS with Respect to Measurement Granularity for Cases A and B

For these scenarios, Figure 7 depicts the influence of measurement granularity, whereas Figure 8 investigates effects of under- and overestimation in the user's requirement specification.

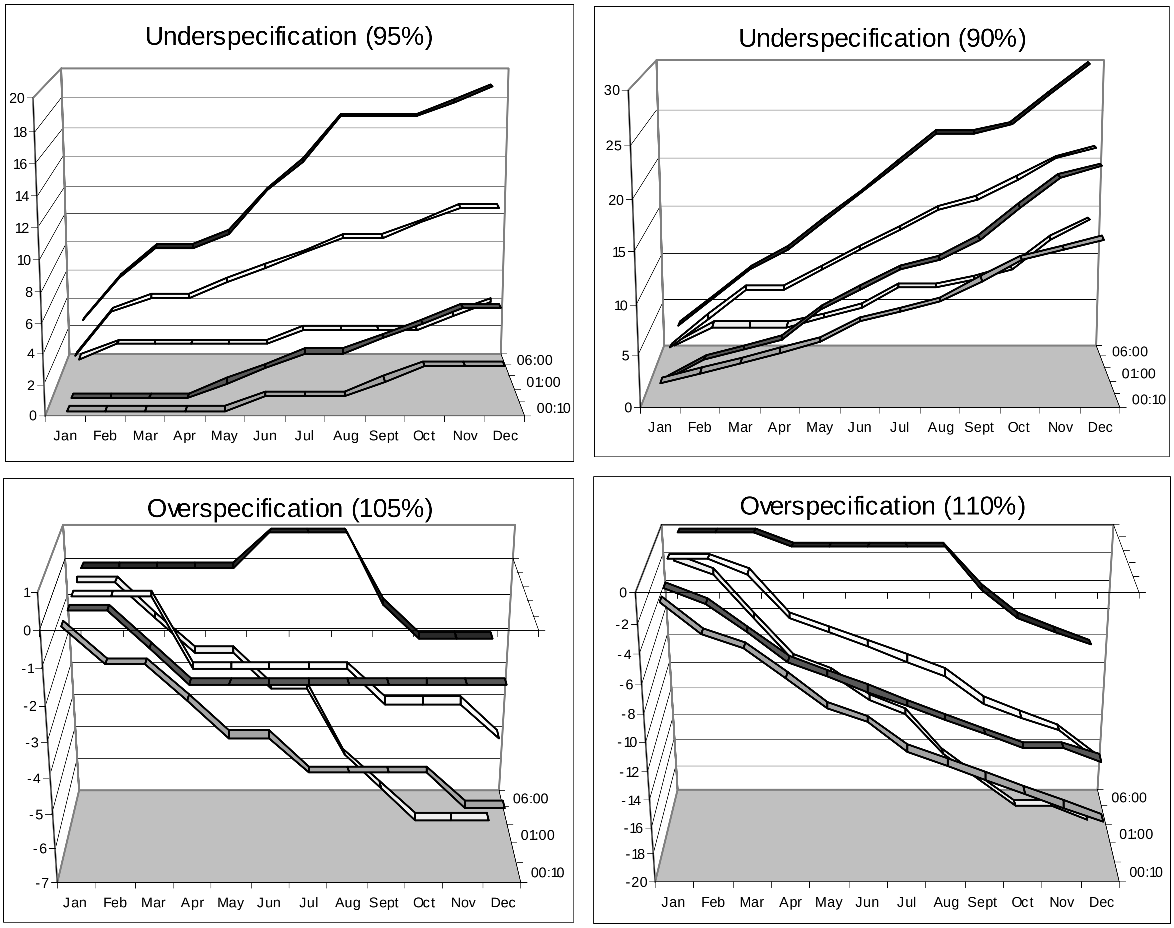


Figure 8: Scenario Case B for varying under- and overestimation of the initial specification

The following class of measurement methods has been applied in this simulation: For various granularities (ranging from 2 minutes to 12 hours), the current bandwidth consumption is determined once per granularity interval, *e.g.*, once within each 2 min interval at a randomly chosen instant. It is assumed that the customer has delivered an initial specification (*i.e.* mean μ and standard deviation σ) of the expected traffic, and threshold levels have been chosen to be σ, 2σ and 3σ in reference case A (where hence the probabilities that the measured traffic is within $[\theta_{-i},\theta_i]$ equals 68%, 95% and 99.7% for i = 1, 2, 3, resp.), and $1{,}3\sigma$, $2{,}4\sigma$, and $3{,}1\sigma$ in reference case B (with corresponding probabilities of 90%, 99% and 99.9%).

Let us first have a closer look at Figure 7, illustrating the evolution of the accumulated number of Cumulus Points (y-axis) in the course of one year (x-axis) for various granularities of the measurement procedure. Granularity (ranging from 10 and 20 minutes to 1, 3, 6 and 12 hours) in this context corresponds to the minimal period in time during which one measurement of the actual resource consumption is made at an randomly chosen instant. Furthermore we assume here that the user delivers a correct initial specification of resource requirements. For case A as well as case B we see that the number of CPs remains pretty stable, except for the two largest granularities (6 and 12 hours) where the corresponding curves lie significantly above the zero-axis. From this and similar results it can be concluded that for a correct specification, CPS is very stable for granularities up to 2–3 hours (corresponding to at least 8–12 suitably distributed measurements per day), whereas for rougher granularities the outcome is less predictable. This gives us an idea about the almost negligible requirements with respect to the measurement tools.

Figure 8 abandons the assumption of correct user specifications. To this end, we assume that the customer delivers an estimate that is either 105% or 110% of the eventual true resource consumption (this is termed “overspecification”) or 95% or 90% (termed “underspecification”). The five curves depicted per figure correspond again to different measurement granularities (10 min, 30 min, 1 hour, 3 hours, 6 hours) with 10 min in the front and 6 hours in the back of the picture. For these cases, the reaction of CPS on all granularity levels is rather consistent, where the number of CPs apparently is increasing the faster the larger the discrepancy between user specification and real resource consumption. Once again, the scheme is very stable with respect to different granularities, and in every case, a suitable reaction threshold level, *e.g.*, 5 CPs, is reached at least after 3–5 months of continuously exceeding the specified requirements. Note that Figure 8 presents simulation results for reference case B, whereas case A yields a similar outcome.

6 Summary and Discussion

The Differentiated Services Architecture offers multiple service classes, which are the result of a certain packet handling strategy within the network, based on Behavior Aggregates and Per-hop Behavior. Service Level Agreements contain technical, financial, and legal aspects which describe particular service classes offered to a customer or another provider. Due to the fact that only suitable pricing will provide the incentive for choosing the “best” service class, mainly in terms of a customer-specific Quality-of-Service/price rate, a DiffServ-capable pricing model and adequate Internet charging technology are required in a commercialized, multi-service Internet.

This paper continues the discussion of the Cumulus Pricing scheme as a special example for the concept of retrospective pricing and complements existing literature on this scheme, e.g. [23], [24]. A sound application scenario on top of a Differentiated Services domain has been introduced and applied to define exemplarily Cumulus Pricing Scheme thresholds according to the service class offerings. Summarizing the results derived from this approach shows that (1) the Cumulus Pricing Scheme is applicable to Differentiated Services, (2) it minimizes the technical effort for metering and accounting on the Internet Service Provider side, while keeping current technology boundaries low, and (3) it obtains an incentive-compatible characteristic for resource usage as discussed in [23].

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 the ETH Zürich has been funded by the Swiss Bundesministerium für Bildung und Wissenschaft, Bern, No. 99.0536, and partially in the framework of the Austrian Kplus Competence Center Program.

References

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss: An Architecture for Differentiated Services. I*nternet Engineering Task Force, RFC 2475, December 1998.*

[2] R. Braden, D. Clark, S. Shenker: Integrated Services in the Internet Architecture: An Overview. *Internet Engineering Task Force, RFC 1633, June 1994.*

[3] B. Briscoe, M. Rizzo, J. Tassel, K. Damianakis: Lightweight Policing and Charging for Packet Networks. *Third IEEE Conference on Open Architectures and Network Programming (OpenArch 2000), Tel Aviv, Israel, March 26-27, 2000.*

[4] N. Brownlee: Traffic Flow Measurement: Experiences with NeTraMet. *Internet Engineering Task Force, RFC 2123, March 1997.*

[5] D. Clark: Combining Sender and Receiver Payments in the Internet. *Telecommunications Research Policy Conference, October 1996.*

[6] Cisco Systems: Service Providers New World News - Pricing for Profitability. *Packet, Cisco Systems, 3rd Quarter 2000, pp 35 – 39.*

[7] R. Cocchi, D. Estrin, S. Shenker, L. Zhang: Pricing in Computer Networks: Motivation, Formulation and Example. *IEEE/ACM Transactions on Networking, Vol. 1, No. 6, December 1993, pp 614 – 627.*

[8] S.A. Cotton (edt.): Network Data Management – Usage (NDM-U) for IP-Based Services. *IPDR Specification Version 1.1, June 2000.*

[9] G. Dermler, M. Günter, T. Braun, B. Stiller: Towards a Scalable System for Per-flow Charging in the Internet. *Applied Telecommunication Symposium, Washington D.C., U.S.A., April 17-19, 2000.*

[10] R. Edell, N. McKeown, P. P. Varaiya: Billing Users and Pricing for TCP. *IEEE Journal an Selected Areas in Communications, Vol. 13, No. 7, July 1995, pp 1162 – 1175.*

[11] R. Edell, P. P. Varaiya: Providing Internet Access: What we learn from the INDEX trial. *Keynote Talk Infocom'99 New York, INDEX Project Report #99-010W. URL: http://www.index.berkeley.edu/99-010W.*

[12] G. Fankhauser, B. Stiller, C. Vögtli, B. Plattner: Reservation-based Charging in an Integrated Services Network. *4th INFORMS Telecommunications Conference, Bocca Raton, Florida, U.S.A., March 1998.*

[13] M. Günter, T. Braun: Internet Service Delivery Control with Mobile Agents. *2nd International Symposium on Agent Systems and Applications/Mobile Agents (ASA/MA 2000), Zürich, Switzerland, September 13-15, 2000.*

[14] A. Gupta, D. O. Stahl, A. B. Whinston: Managing the Internet as an Economic System. *CISM, University of Texas at Austin, U.S.A., July 1994.*

[15] J. MacKie-Mason, H. Varian: Pricing Congestible Network Resources. *IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, 1995, pp 1141 – 1149.*

[16] M. Karsten: QoS Signalling and Charging in a Multi-service Internet Using RSVP. *Ph.D. Thesis, Technical University Darmstadt, Germany, May 2000.*

[17] M. Karsten, J. Schmitt, L. Wolf, R. Steinmetz: An Embedded Charging Approach for RSVP. *6th IEEE/IFIP International Workshop on QoS (IWQoS'98), Napa, California, U.S.A., May 18-20, 1998.*

[18] F. P. Kelly: Notes on Effective Bandwidths. In: Stochastic Networks: Theory and Applications (Edts F.P. Kelly, S. Zachary, I.B. Ziedins). *Royal Statistical Society Lecture Notes Series, 4. Oxford University Press, 1996, pp 141-168.*

[19] M3I: Market Managed Multi-service Internet. *5th Framework EU Project No. IST-1999-11429, URL: http://www.m3i.org.*

[20] K. Nichols, V. Jacobson, L. Zhang: A Two-bit Differentiated Services Architecture for the Internet. *Internet Engineering Task Force, RFC 2638, July 1999.*

[21] A. Odlyzko: Paris Metro Pricing: The Minimalist Differentiated Services Solution. *7th International Workshop on QoS (IWQoS'99), London, U.K., June 1-4,1999, pp 159-161.*

[22] P. Reichl, B. Stiller: Nil nove sub sole? Why Internet Tariff Schemes look like as they do. *4th Berlin Internet Economic Workshop IEW 2001.*

[23] P. Reichl, B. Stiller: Notes on Cumulus Pricing and Time-scale Aspects of Internet Tariff Design. *TIK Report No. 97, ETH Zürich, Switzerland, November 2000.*

[24] P. Reichl, P. Flury, J. Gerke, B. Stiller: How to Overcome the Feasibility Problem for Tariffing Internet Services: The Cumulus Pricing Scheme. I*EEE International Conference on Communications, Helsinki, Finland, June 11-14, 2001.*

[25] S. Shenker, D. Clark, D. Estrin, S. Herzog: Pricing in Computer Networks: Reshaping the Research Agenda. *ACM Computer Communication Review, Vol. 26, No. 2, April 1996, pp 19 – 43.*

[26] SIU: Smart Internet Usage. *Hewlett Packard, http://communications.hp.com/smartinternet/, January 2001.*

[27] 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.*

[28] B. Stiller, J. Gerke, P. Reichl, P. Flury: The Cumulus Pricing Scheme and its Integration into a Generic and Modular Internet Charging System for Differentiated Services. *TIK Report No. 96, ETH Zürich, Switzerland, September 2000.*

[29] B. Stiller, J. Gerke, P. Reichl, P. Flury: Management of Differentiated Service 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.*

[30] 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.*

[31] I. Stoica, H. Zhang: Providing Guaranteed Services Without Per-flow Management. *ACM Computer Communication Review (SIGCOMM'99), Vol. 29, No. 4, October 1999, pp 81-94.*

[32] B. Teitelbaum, P. Chimento: QBone Bandwidth Broker Work Group. *URL: http://qbone.ctit.utwente.nl/BBroker/, August 2000.*