# Edge Pricing in Space and Time: Theoretical and Practical Aspects of the Cumulus Pricing Scheme

Peter Reichl[2,1], Burkhard Stiller[1]

[1]Computer Engineering and Networks Laboratory TIK, Swiss Federal Institute of Technology, ETH Zürich, Switzerland, E-Mail: stiller@tik.ee.ethz.ch

[2]Telecommunications Research Center Vienna, FTW, Vienna, Austria, E-Mail: reichl@ftw.at

***Abstract***

*In the context of the rapidly increasing commercialization of the Internet, the design of suitable pricing models for packet-based networks becomes a necessity for offering approaches for financially-driven sharing of global resources between multiple customers with varying requirements. Addressing such differentiated services both from a customer and a provider point of view has lead to the Cumulus Pricing Scheme (CPS) framework recently established. In this paper, important aspects of CPS are investigated in further detail. Together with a number of simulation results, these investigations demonstrate that CPS is an excellent example for a simple, transparent, market-managed, and feasible Internet pricing scheme that eventually extends the well-known spatial edge pricing paradigm into the time dimension also.*

## 1 Introduction

The variety of qualities that are required to enable different communication services like multimedia, workflow, or teleconferencing applications has lead to introducing and developing so-called Quality-of-Service (QoS) concepts for the Internet. As soon as guaranteed services become a necessity, the networking technology has to provide mechanisms that allow for a clear decision, whether a newly arriving flow shall be permitted to use an Internet service with higher qualities. This permission is essential, since better QoS requires more resources from the network to be supplied and utilized for a particular flow. Based on traditional QoS access control mechanisms, the permission will be driven by the willingness-to-pay of a particular Internet application or an aggregated traffic stream originating in enterprises. Therefore, the combination of up-to-date networking control technology and economic traffic management including suitable pricing models, which allow for the calculation of charges being based on various degrees of network usage, will be relevant for tomorrow's Internet.

The location of such economic control functions determines a crucial issue, both from the customer and the provider point of view. In terms of space, the edge-pricing approach proposed by [5] argues for shifting all

charging functionality towards single points, preferably at the edge between a customer and his wide area backbone network or the edge between two different core networks, respectively. Assuming that the backbone network's technology will follow the Differentiated Services Architecture (DiffServ) approach [1], this edge will be implemented by a DiffServ border router. Distinguishing between single customers, such as a residential customer, and enterprises, comprising of multiple users hidden behind a single financial customer, so-called Service Level Agreements (SLA) [20] are required to define and specify services, which are delivered from an Internet Service Provider (ISP) to customers.

Whereas edge pricing in this "spatial" sense currently dominates the still ongoing discussions about the suitable design of usage-based pricing schemes for Internet services, temporal aspects of pricing mechanisms have not experienced similar attention. In this paper it is argued that this is the reason for the fact that a standard solution for pricing Internet services so far is not in sight, because under these premises the major difficulty of this task depends on accounting a huge number of individual packets traveling through the network. A number of proposals for increasing technical efficiency by reducing the amount of accounting data, especially by carefully choosing parameters, classes, and accounting locations have not led to satisfying scalable and effective solutions. Therefore, in order to solve this "Feasibility Problem of Internet Tariffing" (i.e. the trade-off between customers and their interest in predictable and transparent tariffs and Internet Service Providers (ISPs) trying to achieve economic efficiency for operating the network and technical efficiency for the accounting operation), [11] has proposed a strong paradigm shift and argues that designing an Internet pricing scheme is not a problem of dealing with complexity, but rather a question of multi-dimensional mapping of time-scales. As an example for resulting solutions, the Cumulus Pricing Scheme CPS has been established in [11], [16], and [18] as a newly developed framework for edge pricing in space and time that succeeds in suitably balancing the mentioned trade-off between customer, economical, and technical efficiency. This work is extended in the present paper by detailed investigations of important theoretical and practical aspects of CPS.

The growing interest in Internet pricing schemes is reflected in some recent survey papers, *e.g.*, [6], [19]. One one hand, although usage-based pricing seems to impose an overhead on telecommunication systems [8], [13], any form of usage-based pricing is important, since utilized resources (like frequencies, cables, routers/switches, and most notable operating personnel) are rare and costly. On the other hand, since the traditional, flat-fee Internet pricing model has been constantly critiqued for its economic draw-backs of not being incentive-compatible [7], [13], the Cumulus Pricing Scheme [11] as investigated here appears to be one of the few real novelties in that area. Furthermore, recent charging projects, like M3I [9], CATI [15], and INDEX [4], investigate(d) many different technology and economics questions for the Internet case.

This paper is organized as follows. Section 2 summarizes basic properties of the Cumulus Pricing Scheme (CPS). While Section 3 discusses the developed tariff function and proves its essential requirements, Section 4 deals with assignment levels for Cumulus Points (CP), so-called CP thresholds. Implementation aspects for CPS are presented in Section 5 and the design of a simulation environment for CPS as well as simulation results are added in Section 6. Finally, Section 7 summarizes and draws conclusions.

## 2 The Cumulus Pricing Scheme CPS

The Cumulus Pricing Scheme as established in [11], [16] and [18] is characterized by compound actions on different time-scale levels. Its basic idea consists of three steps: (a) first some flat-rate-type contract is established between customer and ISP, based on the expected resource requirements as specified by the customer; (b) during the lifetime of the contract, the actual resource consumption is monitored and reported back periodically through a rough feedback mechanism (the so-called "Cumulus Points" indicating whether and to which extent the user is violating her resource requirement specification); (c) if the aggregation of the Cumulus Points over some periods exceeds a predefined threshold, the whole contract has to be renegotiated. Up to this point, the customer in any case pays the flat rate as agreed upon; the renegotiation can be performed, *e.g.*, as determination of some extra fee for overutilization, change of expected requirement, and rate.

As introduced formally in [18], 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, respectively, 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, $i = 0,1,2...$ (note that $t_0$ describes the start of the contract between ISP and customer).

Cumulus Points are assigned by the ISP according to the following CP Rule:

**CP Rule**:

*Define* $\theta_n$, $n = -N,\ldots,-1,0,1,2,\ldots,N$, *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*

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

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

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

Hence, if $\Delta_i$ is positive (i.e. overuse in period i) and lies between thresholds $\theta_c$ and $\theta_{c+1}$, then *c* cumulus points are assigned. If $\Delta_i$ is negative and between thresholds $\theta_{c-1}$ and $\theta_c$, then *c* cumulus points are assigned, where *c* now is a negative number, hence the cumulus points are referred to as "green" ones, whereas for positive *c* the cumulus points are "red".

Now the cumulus points $c_i$ are accumulated over time according to

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

hence, $\Gamma_n$ 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* $\Theta$ *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*

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

Depending on $\operatorname{sgn}\Gamma_n$, there may as well be two different thresholds $\Theta^+$ and $\Theta^-$ for red and green CPs, respectively. For further details about the renegotiation as well as the other degrees of freedom within this scheme, please refer to [11] and [12].

The remainder of this paper deals with specific theoretical and practical considerations of CPS as well as a possible implementation of this pricing scheme. For that reason, the term price will be applied for a number and unit currency defined for a particular resource, such as bandwidth, volume, or time. A tariff defines a function, which applies prices in a combined fashion for a given customer and service.

## 3 The Tariff Function $p(x)$

The starting point of CPS consists of an initial contract between service provider and customer, *e.g.*, in the form of an SLA (Service Level Agreement). During this negotiation, the customer has to state her expected resource requirements $x$, and the provider offers a flat rate $p(x)$ for this request. Figure 1 shows a sketch of the general shape of $p(x)$ and the resulting increase of total charge from, *e.g.*, bandwidth $x_0$ to $x_1$.

Moreover, if the customer exceeds her expected requirements, the same function $p(x)$ is supposed to be used while determining eventual extra fees the customer has to pay[1]. Therefore, the derivation of $p(x)$ is of central importance.

Generally, any QoS parameter or combination of QoS parameters may be the subject of the SLA and subsequently requesting a tariff function of its own. To simplify considerations, for the time being we restrict ourselves to the case of one-dimensional QoS parameters that are time-independent (*e.g.*, bandwidth, volume); the extension towards time-dependent parameters (delay, jitter etc.) as well as the investigation of multidimensional QoS vectors is subject of further work[2].

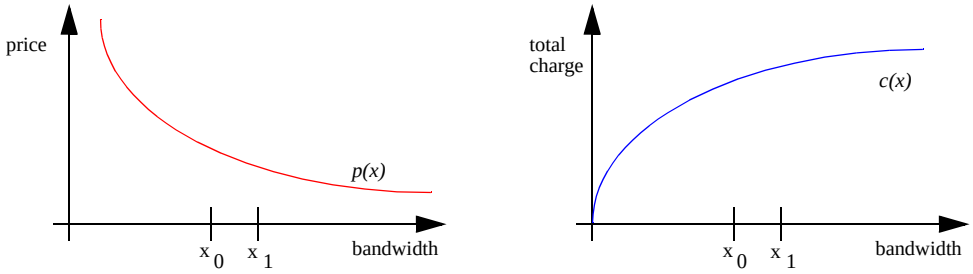


Figure 1: The Tariff Function $p(x)$ (Left) and the Resulting Total Charge $c(x)$ (Right) According to (6) as a Function of Bandwidth Consumption.

Assume the customer's expected resource consumption to be $x_0$, whereas her measured requirement equals $x_1 > x_0$. Assume for instance that we focus on bandwidth pricing. Then let $p(x)$ be the price per bandwidth unit if bandwidth $x$ is agreed upon, and define

$$c(x) = x \cdot p(x) \tag{6}$$

to be the total charge for this bandwidth. Note that in general there is always a function $c(x)$ describing the total charge for using resources $x$; in our bandwidth example this function equals the product of size of resource consumption and price per resource unit.

If the measured customer requirements differ from the expected ones by

$$\delta = x_1 - x_0, \tag{7}$$

1. It is intended to process SLAs automatically, therefore, it is of interest to keep the number of parameters low. To this end, it is aims at deriving a suitable class of candidate functions among which one or two parameters are sufficient to characterize the tariff unanimously. Moreover, using the same tariff function for calculating extra fees further simplifies the processing of SLAs.
2. The distinction between time-independent and time-dependent QoS parameters is necessary as there is a direct influence on the *c*-function introduced in (6): in the case of bandwidth or volume, *m* as the total charge is obviously of the form " bandwidth × price per bandwidth unit ", whereas for parameters like delay or jitter the form of *c* is not clear immediately.

then, after accumulating a sufficient number of red Cumulus Points, the customer (in order to extinguish her red CPs) may be charged an extra fee. Obviously, this extra fee should be based on her additional requirement $\delta$ and hence on $p(\delta)$ for the period during which CPs have been accumulated. In order to provide correct economic incentives, the compound charge $c(x_0) + c(\delta)$ has to exceed the charge $c(x_1)$ for the correct statement, where the difference serves as "penalty function" $\Psi(x_0, x_1)$:

$$\Psi(x_0, x_1) = c(x_1) - [c(x_0) + c(\delta)] = x_1 \cdot p(x_1) - [x_0 \cdot p(x_0) + \delta \cdot p(\delta)] \ . \qquad (8)$$

Then we have following requirements for the tariff function:

0. $p(x) > 0$ is a positive function that is monotonically decreasing, due to the usual provision of discount for increasing size of goods sold.
1. $c(x) = x \cdot p(x)$ is a monotonically increasing function, i.e. higher total bandwidth consumption yields higher total charge (see Figure 1 for the relationship between $p(x)$ and $c(x)$);
2. $\Psi(x_0, x_1) < 0$ if $x_0 \neq x_1$, and $\Psi(x_0, x_1) = 0$ if $x_0 = x_1$, i.e. the customer is to be punished if expected and measured requirements do not coincide, whereas stating the resource consumption correctly minimizes the resulting absolute value of the penalty function;
3. $\Psi(x_0, x_0 + \delta)$ is monotonically decreasing in $\delta = x_1 - x_0$, i.e. the larger the deviation from the expected requirement $x_0$, the larger the absolute value of the penalty function;
4. $|\Psi(x_0, x_1)| < |\Psi(\beta x_0, \beta x_1)| \leq \beta \cdot |\Psi(x_0, x_1)|$ for $\beta > 1$, i.e. a scaling property for increasing bandwidth: for high bandwidth, a similar relative estimation error yields a penalty being higher than for low bandwidths (because the absolute deviation is larger), but the penalty per unit of deviation is not larger than for lower bandwidth (i.e. the penalty does not grow more than linearly with the scaling factor).

Whereas requirements 0. – 3. are somehow straightforward, requirement 4 needs some additional comment. For getting a better intuition, assume, *e.g.*, $\Psi(x_0, x_0 + \delta)$ to be the penalty function for estimating the expected resource requirements wrongly by $\frac{\delta \cdot 100}{x_0}$ %. Making the same relative error for the case of a much larger size of expected resource requirement, $100 \cdot x_0$ say, should not result in a penalty value that is more than 100 times as high as the original one, hence $\Psi(100x_0, 100(x_0 + \delta)) \leq 100 \cdot \Psi(x_0, x_0 + \delta)$, otherwise splitting up the requirement into smaller pieces could bring an advantage. On the other hand, the penalty value should certainly be higher than for the original case, simply because the absolute size of deviation is 100 times as high as in the first case.

In fact, there is a strong argument for the idea that the scaling should obey a square-root law, i.e. $\Psi(\beta x_0, \beta x_1) = \sqrt{\beta} \cdot \Psi(x_0, x_1)$: Assume a source that requires a mean of $x_0$ and whose "uncertainty" is characterized by a standard deviation of $\sigma_0$. According to standard probability theory, multiplexing $N$ independent such sources yields one aggregated source with mean $N \cdot x_0$ and standard deviation $\sqrt{N} \cdot \sigma_0$. Now, if we assume that the penalty $\Psi(x_0, x_0 + \delta)$ should be proportional to the relative error of the estimation w.r.t. the standard deviation, i.e. $\Psi(x_0, x_0 + \delta) \sim \frac{\delta}{\sigma_0}$, then applying twice this relation yields

$$\Psi(Nx_0, N(x_0 + \delta)) = \Psi(Nx_0, Nx_0 + N\delta) \sim \frac{N\delta}{\sqrt{N} \cdot \sigma_0} = \sqrt{N}\frac{\delta}{\sigma_0} \sim \sqrt{N} \cdot \Psi(x_0, x_0 + \delta) \ . \tag{9}$$

Investigating different candidates for $p(x)$ like $p(x) = 1$, $p(x) = ax + b$, $p(x) = \frac{1}{x}$, $p(x) = \frac{1}{x^2}$ as performed in [12] shows that the class of possible candidates must have the form $p(x) = \frac{1}{x^\alpha}$, $0 < \alpha < 1$. Having requirement 4 in mind, we propose therefore finally the following form of the tariff function:

$$p(x) = \frac{1}{\sqrt{x}}. \tag{10}$$

In this case, $\Psi(x_0, x_1) = \sqrt{x_1} - [\sqrt{x_0} + \sqrt{\delta}] = \sqrt{x_1} - \sqrt{x_0} - \sqrt{x_1 - x_0}$.

**Proposition 1**: $p(x) = \frac{1}{\sqrt{x}}$ fulfills requirements 0. – 4.

**Proof:**

$p(x) = \frac{1}{\sqrt{x}}$ is strictly antitonic and always positive, moreover requirement 1 results from $c(x) = \sqrt{x}$.

Now let $\delta = x_1 - x_0 > 0$. As $(a + b)^2 > a^2 + b^2$ if $a, b > 0$, $(\sqrt{x_0} + \sqrt{\delta})^2 > x_0 + \delta$, i.e. $\sqrt{x_0} + \sqrt{\delta} > \sqrt{x_0 + \delta}$ and hence $\Psi(x_0, x_1) = \sqrt{x_1} - \sqrt{x_0} - \sqrt{x_1 - x_0} < 0$. Therefore, requirement 2 is fulfilled.

Deriving $\Psi(x_0, x_0 + \delta) = \sqrt{x_0 + \delta} - [\sqrt{x_0} + \sqrt{\delta}] = (\sqrt{x_0 + \delta} - \sqrt{\delta}) - \sqrt{x_0}$ with respect to $\delta$ yields $\frac{d}{d\delta}(\sqrt{x_0 + \delta} - \sqrt{\delta}) = \frac{1}{2}\left(\frac{1}{\sqrt{x_0 + \delta}} - \frac{1}{\sqrt{\delta}}\right)$. As $\delta > 0$, we have $\sqrt{x_0 + \delta} > \sqrt{\delta}$, hence $\frac{1}{\sqrt{x_0 + \delta}} < \frac{1}{\sqrt{\delta}}$ and therefore, $\frac{d}{d\delta}(\sqrt{x_0 + \delta} - \sqrt{\delta}) < 0$, thus validating requirement 3.

Finally, $\Psi(\beta x_0, \beta x_1) = \sqrt{\beta x_1} - [\sqrt{\beta x_0} + \sqrt{\beta\delta}] = \sqrt{\beta}(\sqrt{x_1} - [\sqrt{x_0} + \sqrt{\delta}]) = \sqrt{\beta} \cdot \Psi(x_0, x_1)$ is consistent to requirement 4. ❑

Note that the same proof applies if the tariff function is scaled by an arbitrary positive factor $\lambda > 0$, i.e.

$$p_\lambda(x) = \frac{\lambda}{\sqrt{x}}. \tag{11}$$

In this case, we have to substitute $c$ and $\Psi$ according to $c_\lambda(x) = x \cdot p_\lambda(x) = \lambda \cdot c(x)$ and $\Psi_\lambda(x_0, x_1) = c_\lambda(x_1) - [c_\lambda(x_0) + c_\lambda(\delta)] = \lambda \cdot \Psi(x_0, x_1)$, but this linear scaling does not concern the validity of the proof. Therefore, $\lambda$ provides an additional degree of freedom.

# 4 Distance between Thresholds for the CP Rule

According to the formal description as presented in Section 2, Cumulus Points are assigned in close connection with violating certain thresholds that have to fulfill the following requirements:

1. Use only a small number of thresholds either way (i.e. no more than 3 – 5).
2. Small oscillating deviations should not result in unnecessary CP assignments (hysteresis effect).
3. Choose the thresholds such that the CPs are more or less independent of the measurement method applied (i.e. the CP difference for using different measurement methods must not exceed 1).

For the following considerations, assume that *V(t)* can be described by a stochastic process which is in equilibrium[1], fluctuating around mean $\xi$ with (unknown) standard deviation $\sigma$. Any measurement performed on the resource then aims at estimating $\xi$ for deciding which thresholds are exceeded by $\xi - x$. Estimating $\xi$ with unknown $\sigma$ from a normally distributed process is the one of the basic problems of the so-called statistical confidence estimation [2].

Assume $X_n$, $n = 1, ..., N$, to be sample values for *V(t)* and $\bar{X} = \frac{1}{N}\sum^{N} X_n$ and $S^2 = \frac{1}{N-1}\sum^{N} (X_n - \bar{X})^2$ to be the well-known standard estimators for mean and variance. The solution of the confidence estimation problem is based on the fact that $\frac{\bar{X} - \xi}{S}\sqrt{N}$ suffices a Student-t distribution with $m = N - 1$ degrees of freedom. Without going too much into detail, this means the following:

1. Note that this assumption may certainly be subject to discussion, especially with respect to all the recent work on self-similar traffic in the Internet. Nevertheless, the steady-state case may serve as a feasible starting point.

With $\alpha$ describing the confidence level of the estimation, the estimation procedure yields a confidence interval $(\bar{X} - \varepsilon, \bar{X} + \varepsilon)$, i.e. $\xi$ lies within this interval with probability $1 - \alpha$. Now, having chosen $\alpha$, there are tables (*e.g.*, [2] on page 22) that allow to determine a number $t_{\alpha,N-1}$ such that the interval

$$(\bar{X} - \varepsilon_{\alpha,N}, \bar{X} + \varepsilon_{\alpha,N}) \tag{12}$$

with

$$\varepsilon_{\alpha,N} = \frac{S}{\sqrt{N}} t_{\alpha,N-1} \tag{13}$$

is a confidence estimation of $\xi$ for confidence level $1 - \alpha$.

This rationale is summarized in the following proposition:

**Proposition 2:** Under steady-state assumptions, requirement 3 is fulfilled if every two neighboring thresholds have at least distance $2 \cdot \varepsilon_{\alpha,N}$ as described by (13). ❑

Hence, we can devise at least a minimal distance between each two thresholds. This is especially interesting, because the Student-t distribution possesses an asymptotic limit, i.e. for sample sizes larger than 30 $t_{\alpha,N-1}$ does no longer vary essentially. This asymptotic case corresponds to observing continuously the process *V(t)*.

Moreover, proposition 2 solves also the issue of requirement 2: If the distance between $\theta_{-1}$ and $\theta_1$ is larger than $2 \cdot \varepsilon_{\alpha,N}$, then there is no danger for getting an estimation of $\xi$ that exceeds either threshold, hence in this case there won't be assigned any CPs.

Note, these results always depend on the confidence level $1 - \alpha$, i.e. are only able to provide statistical guarantees. Certainly, assuming steady-state conditions is another limiting constraint, but proposition 2 allows at least some statement about minimal distances between suitable thresholds. Their number as well as their relative sizes are still open issues and subject to further work, but there is a good chance to derive useful answers to these questions from a simulative evaluation of (13), where traffic sampling allows easily to calculate $\bar{X}$ and $S$ in order to derive $\varepsilon_{\alpha,N}$.

# 5 Implementation Aspects

CPS as described and investigated so far basically is a framework that can be realized in various ways. This section presents ideas about a possible implementation as currently undertaken within the M3I project [9] and includes the investigation of expected customer behavior during an initial "Probe Phase" before starting the actual charging scheme and the derivation of thresholds using the results of this Probe Phase.

## 5.1 The Probe Phase (PP)

As already stated above, CPS requires an initial contract (*e.g.* in the form of an SLA) between service provider and customer that contains the customer's expected traffic characteristics. Due to certain legal issues, especially the question of non-repudiation, it appears to be necessary that the customer is involved in the process of parameter determination. According to the original CPS proposal [11], it is in fact the customer alone who is responsible for finding out his expected traffic characteristics, but obviously she might need help in doing so. One possible approach (for others see [12]) consists of allowing the customer to use the network for a short "Probe Phase" where she pays a rather low flat rate (or gets refund if her usage is very small), during which the ISP (and maybe the customer too) performs detailed measurements, which afterwards are confirmed by the customer and are used to determine parameters on which the initial contract is based. Generally there are two possibilities: (1) outsourcing this problem to a third party or to a special tool offered to the customer by the ISP, but running under the responsibility of the customer, or (2) measurements performed by the ISP with results accessible to the customer in order to allow her deriving a statement of expected resource requirements for the initial contract (in this way, the ISP has an influence on the reliability of the traffic expectation, but the last responsibility is still with the customer). In any case, a detailed statistical evaluation on the current user behavior should be performed, determining at least mean $\xi$ and standard deviation $\sigma$ of current resource requirements in order to estimate expected requirements and thus provide basic parameters for the initial contract in a way both customers and providers can trust.

## 5.2 Choosing Thresholds for the CP Rule

In Section 4, we have presented some considerations concerning the distance between the thresholds in order to make the CP assignment procedure widely independent on the measurement method used by the ISP. This is apparently only a partial answer to the question about how to determine these thresholds in practice. The choices here include, *e.g.*,

- the number of thresholds – we have already discussed that it is reasonable to choose a relatively low number of CPs that can be awarded for one time unit (*e.g.*, 3 to maximally 5 CPs per direction);
- absolute vs. relative thresholds, i.e. the thresholds may have either the form $\theta_n = 5$ [MByte/s] or $\theta_n = 110\%$ (*e.g.*, of $\xi$);
- linear vs. non-linear thresholds, i.e. should the distances between the neighboring thresholds (either relative or absolute) be equal or not;
- thresholds depending on additional parameters vs. independent thresholds – *e.g.*, one could imagine to include the standard deviation $\sigma$ into the process of calculating thresholds.

As CPS generally provides each customer with very individual contracts (*i.e.* SLAs), this provides the possibility to adjust thresholds exactly to the specific customer needs. In this sense, relative thresholds provide a transparent way of adapting the thresholds in case of changing requirements. If the first threshold bites, *e.g.*, if the actual traffic deviates by more than 10% from the expected value, this may be left valid even if the statement of expected resource requirements has to be changed. On the other hand, the de facto value of the relative threshold probably should consider the individual situation of the customer. Therefore, our current simulation scenario opts for including the standard deviation $\sigma$ into the threshold determination.

Assume once again that the actual resource consumption may be viewed as a steady-state stochastic process. The main aim of assigning CPs consists of indicating whether the statement about expected resource consumption (still) is valid. A simple heuristic could work as follows: If $\xi$ and $\sigma$ are reasonable estimations for mean and standard deviation, and if the mentioned stochastic process is reasonably close to normal distribution (which is a standard assumption for our case), then according to standard probability theory approximately 68.3% of the samples should be contained within the interval $[\xi - \sigma, \xi + \sigma]$, approximately 95.4% within $[\xi - 2\sigma, \xi + 2\sigma]$ etc. Hence, if the sample measurement is "too often" outside $[\xi - \sigma, \xi + \sigma]$, this is an indication that maybe stating mean $\xi$ is no longer valid. If the samples are outside $[\xi - 2\sigma, \xi + 2\sigma]$ or even $[\xi - 3\sigma, \xi + 3\sigma]$, this indication is even stronger. This suggests in a straightforward generalization to use relative thresholds of the form

$$\theta_{\pm n} = 1 \pm \gamma_n \cdot \frac{\sigma}{\xi} \tag{14}$$

i.e. one CP is awarded as soon as the measurement differs from $\xi$ by more than $\gamma_1 \cdot \sigma$, two CPs are awarded if the deviation is larger than $\gamma_2 \cdot \sigma$ etc. Setting $\gamma_n = n$ yields linear thresholds with properties as described above. One proposal for non-linear thresholds might be $\gamma_1 = 1{,}3$, $\gamma_2 = 2{,}4$, $\gamma_3 = 3{,}1$ and maybe $\gamma_4 = 3{,}7$, where the numbers are chosen in a way that assuming normal distribution one sample out of ten can be expected to exceed regularly the first threshold, one out of 100 the second one, one out of 1000 the third one etc., thus supporting the early warning feature of CPS. As an additional feature, both these approaches limit inherently the number of thresholds to be about 3 or 4 in each direction.

# 6 The Simulation Tool CPSim

In order to investigate CPS not only theoretically, *e.g.*, by the mathematical proofs of important characteristics as described in earlier sections of this paper, also a simulation model has been designed. It will 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 2(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 2(b), acts as an interface for configuring thresholds, obtaining characteristics and statistics of red and green CP assignments.

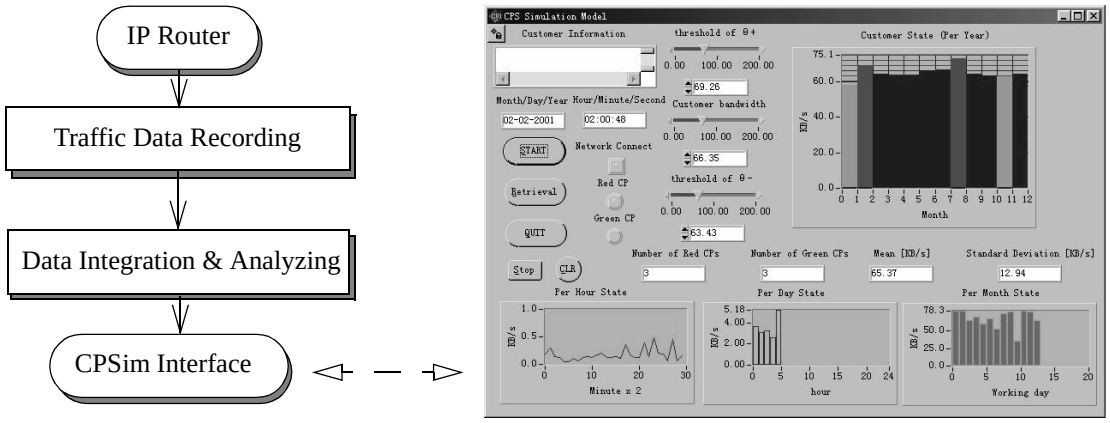


Figure 2: (a) CPSim Tool Architecture and (b) User Interface of CPSim including an Example Result

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 2(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.

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, and the following class of measurement methods: 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 a correct initial specification (*i.e.* mean $\mu$ and standard deviation $\sigma$) of the expected traffic, and threshold levels have been chosen to be $1{,}3\sigma$, $2{,}4\sigma$, and $3{,}1\sigma$. For this scenario, Figure 3 (left) depicts the influence of measurement granularity. From this and similar results it can be concluded that 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. Figure 3 (right) investigates the situation, where the customer specification is underestimating the real requirements by about 10%. In this case, the reaction of CPS on all granularity levels is rather consistent, 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.

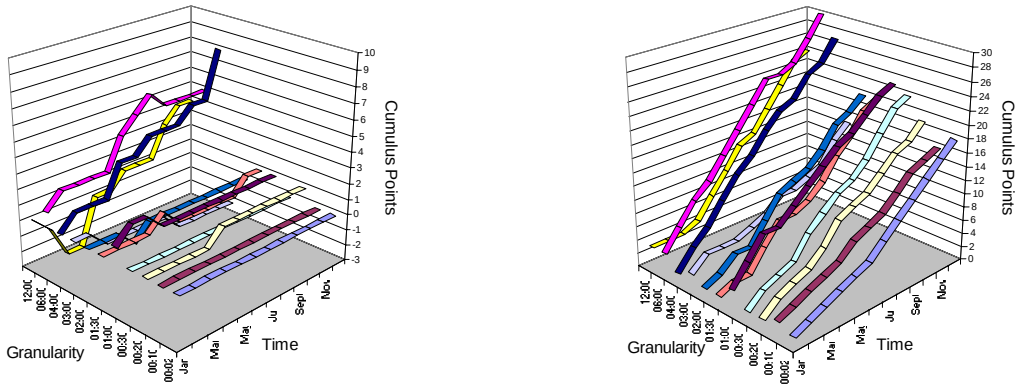


Figure 3: Cumulus Points for Different Measurement Methods with Varying Granularities

## 7 Summary and Conclusions

The Cumulus Pricing Scheme (CPS) provides a flexible design framework for a flat-fee, transparent, predictable, and multi-provider-capable pricing scheme, which may be implemented in different ways. Depending on a monthly over- or underutilization of an initial estimation, red or green Cumulus Points (CP) are assigned to customers, where this assignment is based on threshold levels. As it has been shown, the tariff function offering a flat-fee varies over time according to resource usage, more specifically, incremental usage changes. However, it ensures that later adjustments of traffic characteristics, basically not meeting the initial estimation, will end up in paying a higher price for this communication service.

Due to the definition of thresholds for these CP assignments, it becomes possible to use only a small number of thresholds for red and green CPs, in the range of about 3-5 levels. Built-in hysteresis effects prevent small oscillating deviations, resulting in unnecessary awards of CPs. Finally, the selection of these thresholds assigns CPs that are widely independent of the technical measurement method applied.

The initial estimation of traffic characteristics forms the basis for CPS and allows for an economically and technically efficient approach. In addition, the flat-fee approach and the possibility for contract renegotiation (restricted to a longer time-scale) provides the desired degree of price stability and predictability for customers. The Probe Phase has been proposed as one feasible mechanism to obtain a specification of the expected traffic that allows setting up and renegotiating the contract between provider and customer. Note that it is this temporal restriction of tariff and contract changes to specific points in time which eventually provides the temporal analog to the (originally spatial) edge pricing paradigm, which itself is fulfilled by CPS in a straightforward manner.

Future work includes the follow-on development of the CPSim tool as well as a set of simulations, based on additional real data measurements of LAN-WAN access routers. This will enable a calibration of additional CPS parameters, such as exact values for CP thresholds based on the mean and standard deviations, for a given networking environment. In addition, user reaction will be studied in a practical environment, mainly within the M3I project context, by implementing a CPS scenario based on a developed Internet Charging System.

**Acknowledgements**

The authors very much like to acknowledge Hongguang Ma for implementing the CPSim tool and thank Placi Flury, Jan Gerke, and Hasan for many lively discussions. This work has been performed partially in the framework of the EU IST project Market Managed Multi-service Internet (M3I, IST-1999-11429) with ETH Zürich being 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.

**References**

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

[2] I. N. Bronstein, K. A. Semendjaev: *Taschenbuch der Mathematik;* (in German) 23rd Edition, Verlag Harri Deutsch, Frankfurt/Main, Germany, 1987.

[3] CATI: *Charging and Accounting Technology for the Internet;* Available at the URL: http://www.tik.ee.ethz.ch/~cati, March 2000.

[4] K. Chu, J. Altmann: *Demand for Different Qualities of Service for Internet Access: A Review of INDEX Findings;* Network Modelling in the 21st Century, Royal Society, London, December 1999

[5] 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.

[6] M. Falkner, M. Devetsikiotis, I. Lamdadaris: *An Overview of Pricing Concepts for Broadband IP Networks;* IEEE Communications Surveys, 2nd Quarter 2000, pp 2-13.

[7] 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.

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

[9] M3I: *Market Managed Multi-service Internet;* 5th Framework EU Project, IST Program, No. 11429, Available at the URL: http://www.tik.ee.ethz.ch/~m3i, January 2001.

[10] P. Reichl: *A GTES Model for Periodic Traffic;* International Conference on Computer Communications (ICC'98), Atlanta, GA, U.S.A., June 8-10, 1998.

[11] P. Reichl, P. Flury, J. Gerke, B. Stiller: *How to Overcome the Feasibility Problem for Tariffing Internet Services: The Cumulus Pricing Scheme;* To appear: IEEE International Conference on Communications (ICC'2001), Helsinki, Finland, June 11-15, 2001.

[12] P. Reichl, B. Stiller: *Notes on Cumulus Pricing and Time-Scale Aspects of Internet Tariff Design;* Technical Report No. 97, Computer Engineering and Networks Laboratory TIK, ETH Zürich, Switzerland, November 2000.

[13] 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.

[14] D. Songhurst (edt): *Charging Communication Networks - From Theory to Practice;* Elsevier, Amsterdam, The Netherlands, 1999.

[15] 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, LNCS, Springer Verlag, Heidelberg, Vol. 1629, pp 281-296.

[16] B. Stiller, J. Gerke, P. Flury, P. Reichl, Hasan: *Charging Distributed Services of a Computational Grid Architecture;* International Workshop on Internet QoS for the Global Computing (IQ'2001), Brisbane, Australia, May 16-18, 2001.

[17] B. Stiller, J. Gerke, P. Reichl, P. Flury: *The Cumulus Pricing Scheme and Its Integration into a Generic and Modular Charging and Accounting System for Differentiated Services*; Technical Report No. 96, Computer Engineering and Networks Laboratory TIK, ETH Zürich, Switzerland, August 2000.

[18] B. Stiller, J. Gerke, P. Reichl, P. Flury: *Management of Differentiated Services Usage by the Cumulus Pricing Scheme and a Generic Internet Charging System;* To appear: IEEE/IFIP Symposium on Integrated Network Management (IM'2001), Seattle, Washington, U.S.A., May 14-17, 2001.

[19] B. Stiller, P. Reichl, S. Leinen: *A Practical Review of Pricing and Cost Recovery for Internet Services;* to appear: NETNOMICS - Economic Research and Electronic Networking, Vol. 3, No. 1, March 2001.

[20] B. Teitelbaum, P. Chimento: *QBone Bandwidth Broker Work Group;* Available at URL: http://qbone.internet2.edu/, August 2000.