ETH zürich

ETH Library

Notes on cumulus pricing and time-scale aspects of internet tariff design

Report

Author(s):
Reichl, Peter; Stiller, Burkhard

Publication date:
2000-11

Permanent link:
https://doi.org/10.3929/ethz-a-004284640

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
TIK Report 97

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Peter Reichl, Burkhard Stiller

Notes on Cumulus Pricing and Time-scale Aspects of Internet Tariff Design

TIK-Report
Nr. 97, November 2000

Institut für Technische Informatik und Kommunikationsnetze
Computer Engineering and Networks Laboratory

Peter Reichl, Burkhard Stiller:
Notes on Cumulus Pricing and Time-scale Aspects of Internet Tariff Design
November 2000
Version 1.0
TIK-Report Nr. 97

Computer Engineering and Networks Laboratory,
Swiss Federal Institute of Technology (ETH) Zurich

Institut für Technische Informatik und Kommunikationsnetze, TIK
Eidgenössische Technische Hochschule Zürich

Gloriastrasse 35, ETH-Zentrum, CH-8092 Zürich, Switzerland

Notes on Cumulus Pricing and Time-scale Aspects of Internet Tariff Design

Peter Reichl, Burkhard Stiller

Computer Engineering and Networks Laboratory, TIK, ETH Zürich, Switzerland

E-Mail: [reichl|stiller]@tik.ee.ethz.ch

November 20, 2000

Abstract

This report deals with design issues for Internet tariff schemes. It proposes a framework that is able to explain why current tariff proposals look like as they do, and why there cannot exist tariffs that are significantly different from the existing ones. Moreover, it is demonstrated how an extension of this framework allows to design a new tariff that eventually even solves the so-called “feasibility problem”, i.e. the trade-off between technical, economical and user-based requirements. To this end, the general focus is directed towards time-scales. There are four time-scales identified as being relevant for Internet tariffing, and the close relationship between these time-scales and existing tariffs is demonstrated. Afterwards, the notion of “tariff reaction” is introduced, and the Cumulus Pricing Scheme CPS is presented as leading example of the resulting new tariff structures. The second part of the report deals with some important mathematical results for CPS and various implementation aspects as well as a couple of open issues to be answered by simulation.

1 Introduction

With the exponential rise of the Internet over the last years, the question of designing suitable usage-based tariffs for Internet services has become of increasing interest. Despite of all efforts, a standard solution for such a tariff still does not exist. The major difficulty of this task depends on accounting the huge number of individual packets travelling through the network. There has been a number of proposals for reducing the incidental amount of data, especially by carefully choosing parameters, classes and accounting locations. However, these approaches which aim at reducing the complexity of the problem have not led to satisfying scalable and effective solutions. Therefore, in this report we propose a paradigm shift and argue that designing an Internet pricing scheme is not a problem of dealing with complexity, but rather a question of multidimensional mapping of time-scales.

2 The Feasibility Problem for Internet Pricing

2.1 General Requirements

Any proposal for tariffing Internet services has to cope with several general requirements. As described in [1], Figure 1 summarizes the three main requirement types (RT-1 to RT-3).

RT-1: Customer. Over the last years, there has been an extensive discussion on preferences customers show towards dynamic tariff schemes. E.g., within the INDEX [4], CATI [2], and M3I [3] projects it has turned out that especially transparency and predictability of charges are of major importance to the Internet customer.

RT-2: ISP - Economic. Economically, the ISP is interested in running the network efficiently, e.g. by maximizing network utilization or total revenue. Thus, pricing schemes represent an important interface to the customer, as prices may be used to indicate well-behavior or misbehavior of the customer or to signal the congestion state of the network, i.e. allow the ISP to communicate the relationship between current customer behavior and overall system status.

RT-3: ISP - Technical. Each Internet pricing scheme depends heavily on the existence of tools for technical accounting. There is a vast field of possibilities as to which detail data about the network status are to be obtained since technical conditions may vary enormously.

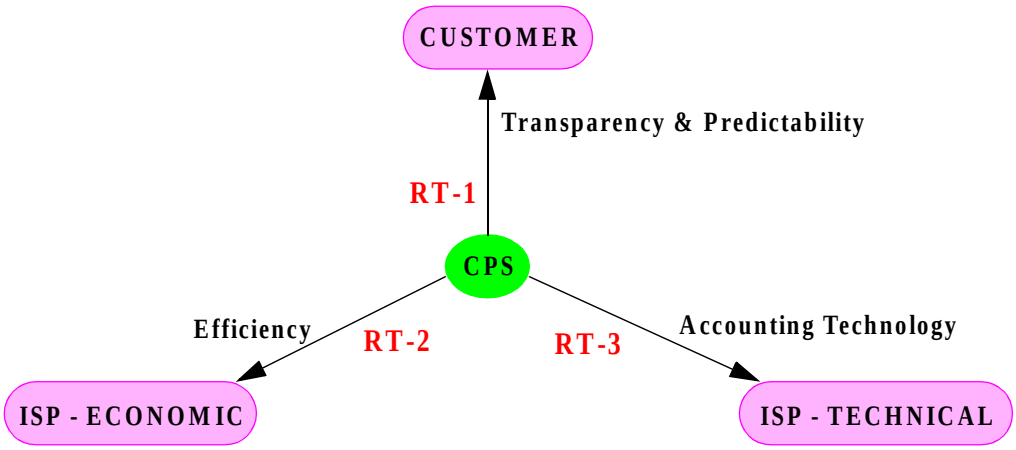

Figure 1: General Requirement Types

The quality and suitability of Internet pricing schemes is to be evaluated in terms of the respective balance between these three requirements. For example, flat rate pricing is excellent with respect to requirement types 1 and 3, but does not support economic efficiency at all, whereas Vickrey Auctions are efficient, but may yield unstable charges as well as very complex technical problems.

2.2 The “Feasibility Problem” of Internet Pricing

An even closer analysis in combination with practical experiences reveals that not all three requirements are equally important: Any scheme that does not fulfill the criterion of technical feasibility has no chance of being implemented for practical purposes. In this sense, pricing schemes may have different trade-offs between type 1 and 2, whereas requirement type 3 represents a hard criterion. This fact is termed the “Feasibility Problem” of Internet pricing. In the rest of the report we propose a solution by taking a fresh look to the general concept of pricing schemes. We no longer view them as being mappings from resource accounting parameters towards prices, but instead as mappings from very small time-scales to larger ones.

3 The Methodology of Time-Scales

This section introduces the so-called Methodology of Time-Scales (MTS). MTS is a framework describing tariff schemes formally as combination of various multi-dimensional mappings on different time-scales as it will be described in this section.

3.1 Time-Scales

In [1], we have already identified the following four different time scales as being relevant for Internet pricing schemes:

Atomic (communication-relevant): This involves sending packets, round-trip times, and managing feedback between sender(s) and receiver(s). These processes take place in the order of ms.

Short-term (application-relevant): This-time scale is concerned with the usual duration of applications like file-transfer, video-conferencing, or IP phone calls. The tasks of accounting and metering are closely related to these activities. Basic time units here are in the order of seconds to minutes.

Medium-term (billing-oriented): The time-scale for performing billing actions depends strongly on the usual human lifestyle habits of humans, e.g., monthly payments of rents, phone charges, or newspaper bills etc. Here, a week or a month is a good choice for a basic time unit.

Long-term (contract-specific): The largest time-scale in this context is the duration of contracts between customers and ISPs, which usually varies from several months to years as basic unit. Note that contracts between ISPs may be shorter.

Figure 2 sketches the relationship between these time-scales, the respective management tasks and communication contents as well as various pricing schemes.

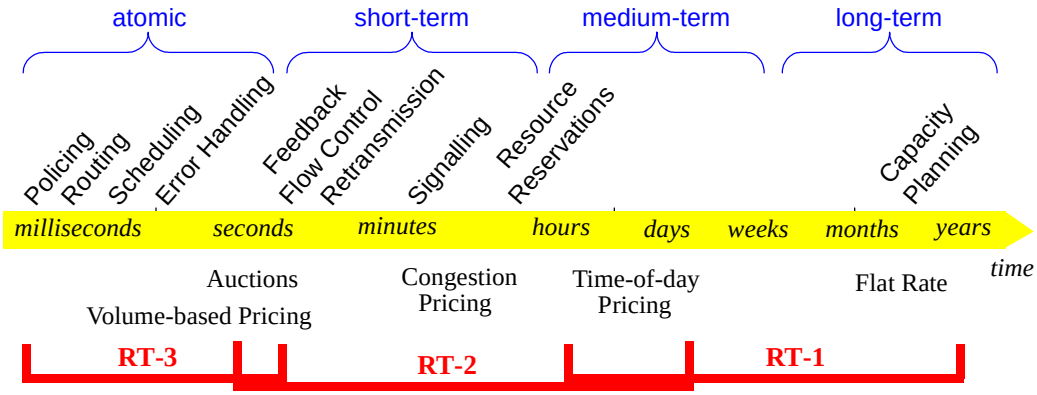

Figure 2: Time-scales

3.2 Charges, Prices, Tariffs

Prices are according to [1] defined as monetary value of a unit of delivered goods. As every time-scale as described in Section 3.1 offers unit goods in some wider sense, there may also be prices associated with each time-scale. Moreover, tariffs often are combined from different time-scales, e.g. in traditional telephony there is a monthly basic charge and additional call-based fees. Let us assume that the network offers one service only, then we can define the price to be a four-dimensional vector

$$\Pi = \left[\pi_1 \ \pi_2 \ \pi_3 \ \pi_4\right] \tag{1}$$

where each dimension represents the price to be paid if one basic unit of the respective time-scale is consumed. This can easily be generalized to the case where there are multiple unit goods asso-

ciated with one time-scale, e.g. various service classes w.r.t. different applications. In this case, π_2 representing the prices on time-scale 2, may be a multi-dimensional vector itself, with one component for each offered service.

The *charge* to be paid for a service etc. often depends linearly on the unit price, where the factor of proportionality is determined by the *tariff*. Put it in other words: Given a unit price π_j, the tariff τ_{ij} (we will come back to the indices in a second) basically is the number to multiply π_j with in order to get the total charge c_j for that particular service. Obviously, this number τ_{ij} may depend on a variety of input parameters. We will introduce the concept of tariffs formally in Section 3.3.

Hence, formally the resulting *charge* is calculated according to

$$C = C(t) = \left[c_1\ c_2\ c_3\ c_4\right] \tag{2}$$

with t denoting time-dependence and

$$c_j = \sum^{4} \tau_{ij}\pi_j\ . \tag{3}$$

Here, each component of C corresponds to a charge to be paid on the respective time-scale. As we try to conform to the design of usual human life-style, we will focus on charging schemes that are zero for all but one time-scale, i.e. time-scale 3 (monthly billing) wherever possible. But note that there may be exceptions to this rule (see Section 3.5.1 e.g.).

Note that the model presented here is much more general than the one proposed in [5], where pricing is a rather general concept that essentially describes the calculation of the monetary value of a certain delivered service. The notion of tariff is reduced to a database called “tariff directory” providing external information for the price calculation. Therefore, the notion of “price” according to [5] corresponds more or less to our concept of “charge”, described in the following summarizing definition.

Definition (Charge, Tariff, Price):

The *charge* for an Internet service corresponds to the monetary equivalent to be paid for its delivery. Usually, delivering a service may be divided conceptually to delivering a couple of basic (atomic) service units, each of which may have an individual (unit) *price*. The *tariff* describes the mapping of these prices (depending on additional input parameters) to the resulting charge. As Internet services are associated with different time-scales, all these notions may be multi-dimensional, with one or more dimension per time-scale each.

Some further remarks:

- Equation (1) implicitly assumes that prices are fixed in time. This appears to be a rather strong assumption, but in fact this only means that the time-dependence is shifted towards the concept of a tariff (see Section 3.3, where (5) describes that any tariff depends explicitly on time). Moreover, Section 3.5.6 looks at packet-based auctions as a pricing scheme with time-dependent prices, but it is a matter of discussion if this type of scheme still is an “Internet tariff”.

- Equation (3) describes charges as component-wise products of tariff and price. It must be noted that this is but the linear version of a more general concept of a mapping from a tariff and a price to a charge that must not necessarily be a linear one.

3.3 Formal Definition of a Tariff Scheme

On each of the time-scales identified in Section 3.1, there may be various input variables contributing to the tariff, as well as output variables that may determine the input parameters of the other time-scales. Hence, a *tariff T* may be described as a *4x4 matrix of mappings from (multi-dimensional) input-parameters to (multi-dimensional) output parameters*, the so-called Tariff Matrix, as shown in (4):

$$T = (\tau_{ij}) = \begin{bmatrix} \tau_{11} & \tau_{12} & \tau_{13} & \tau_{14} \\ \tau_{21} & \tau_{22} & \tau_{23} & \tau_{24} \\ \tau_{31} & \tau_{32} & \tau_{33} & \tau_{34} \\ \tau_{41} & \tau_{42} & \tau_{43} & \tau_{44} \end{bmatrix}. \qquad (4)$$

Here, each τ_{ij} is a function depending on several input parameters:

- Tariffs in general are time-dependent, i.e. may change over time. This feature is expressed by the four-dimensional vector $t = (t_1, t_2, t_3, t_4)$ describing the actual moment of validity in terms of all four time-scales. Suppose each time-scale to have a basic time unit (which could be a month in the case of time-scale 3, e.g.). Then, the vector (t_1, t_2, t_3, t_4) describes four-dimensionally the time unit that is actually charged (e.g. (0, 1, 3, 2) could stand for the first application in month 3 of the duration of contract number 2).

- Tariffs in general are utilization-dependent. Requesting twice the size of a resource usually should yield a higher charge than using it only once. Therefore, we introduce another independent vector $n = (n_1, n_2, n_3, n_4)$ corresponding to utilization on each of the four time-scales. Note that tariffs often depend linearly on n.

- Tariffs in general depend on some more input parameter, either pre-set parameters, output parameters from other time-scales or results from measurements. We distinguish these parameters again according to their relevant time-scale (e.g. measurement on packet-level, e.g. packet sizes, clearly belong to time-scale 1, measurements on application level, e.g. holding times of phone calls, belong to time-scale 2 etc.). We assume that any tariff may maximally depend on input from two different time-scales, and therefore we may describe this input by two vectors: let $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)})$ be the vector of *k* input variables relevant for time-scale i[1] and $x^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_l^{(j)})$ be the vector of *l* input variables relevant for time-scale *j*.

In total, we have therefore *k+l+8* input variables to a tariff. Then τ_{ij} maps these input variables to a vector of *m* output variables $y^{(j)} = (y_1^{(j)}, y_2^{(j)}, \ldots, y_m^{(j)})$ which again are associated with a time-scale (i.e. the time-scale on which the actual charging takes places). In most tariffs, $m = 1$, i.e. all the input parameters are mapped to one output parameter that e.g. may be multiplied with a

1. From now on, by $z^{(s)}$ we denote that the arbitrary parameter parameters z is associated to time-scale s.

respective price to yield a charge according to equation (3). But there may be also tariff schemes that use the output parameters as input for different time-scales, and therefore it is useful to allow $y^{(j)}$ to have more than one dimension.

Summarizing, a tariff as time-scale dependent mapping τ_{ij} looks therefore like

$$\tau_{ij}: \begin{cases} \Re^{k+l+8} \to \Re^{m} \\ \begin{pmatrix} x^{(i)} \\ x^{(j)} \\ t \\ n \end{pmatrix} \to y^{(j)} \end{cases} \tag{5}$$

with (possibly time-dependent) vectors $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)})$, $x^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_l^{(j)})$, $n = (n_1, n_2, n_3, n_4)$ and $t = (t_1, t_2, t_3, t_4)$ as input and $y^{(j)} = (y_1^{(j)}, y_2^{(j)}, \ldots, y_m^{(j)})$ as output.

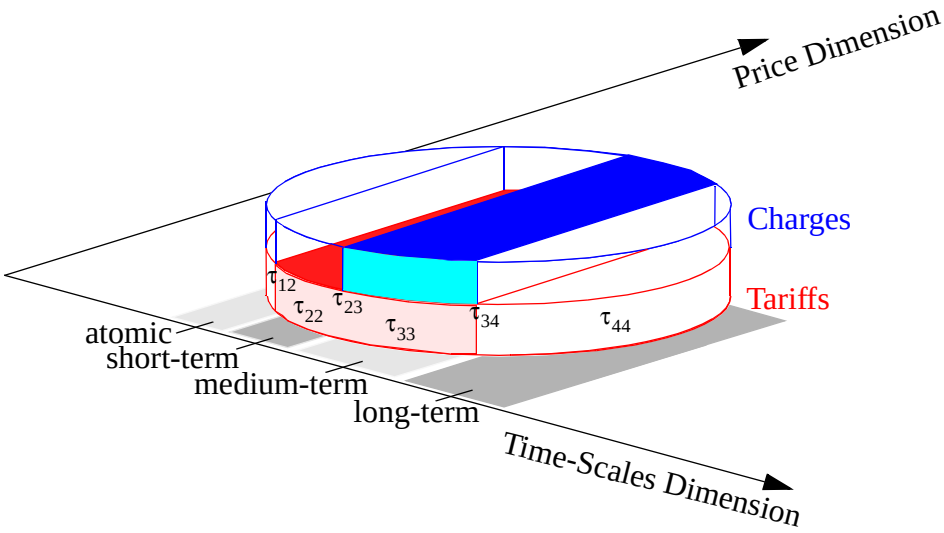

Figure 3: Tariffs as mappings from time-scales dependent input parameters and prices to charges. Example of a τ_{23} tariff, i.e. input from time-scales 2 and 3 (red), output (immediate charge) in time-scale 3 (blue).

3.4 The Form of the Tariff Matrix *T*

Obviously, the matrix of (4) can be reduced to a simpler form, because we may assume that input from one time-scale will usually have influence to the neighbor time-scale only, e.g. atomic measurements won't usually have a direct consequence for the monthly bill. Otherwise, one could always define an intermediate mapping from the atomic to the short-term time-scale and from there to the medium-term one etc. Therefore, we may cancel τ_{31}, τ_{41}, τ_{42}, τ_{13}, τ_{14} and τ_{24}, and (4) may be stated as a tri-diagonal matrix of the form

$$T = (\tau_{ij}) = \begin{bmatrix} \tau_{11} & \tau_{12} & 0 & 0 \\ \tau_{21} & \tau_{22} & \tau_{23} & 0 \\ 0 & \tau_{32} & \tau_{33} & \tau_{34} \\ 0 & 0 & \tau_{43} & \tau_{44} \end{bmatrix}. \tag{6}$$

where only the diagonal and the upper and lower secondary diagonal are non-trivial.

Let us consider now the lower secondary diagonal. There is plenty of reason to assume that τ_{21} and τ_{32} should vanish in every case, because if there is tariff information available from time-scale 2 or 3, respectively, then this information should be used to perform charging on this time-scale instead of going down to time-scale 1 or 2, respectively. This is a direct consequence of what we termed "feasibility problem". τ_{43} is a bit different, as it may make sense to perform charging on time-scale 3 rather than on time-scale 4, and in Section 3.5.2 we will actually see an example for this type of tariff.

Summarizing this argumentation, the Tariff Matrix should have the form

$$T = (\tau_{ij}) = \begin{bmatrix} \tau_{11} & \tau_{12} & 0 & 0 \\ 0 & \tau_{22} & \tau_{23} & 0 \\ 0 & 0 & \tau_{33} & \tau_{34} \\ 0 & 0 & \tau_{43} & \tau_{44} \end{bmatrix}. \tag{7}$$

Another important feature of the Tariff Matrix consists of the fact that for each tariff no more than one element per column should be different from 0. Simplicity on tariff schemes includes that

charging on one time-scale should not involve information from more than two time-scales. If the information comes from the same time-scale where the charging is performed, we have tariffs of type τ_{ii}, and this postulation is obvious. For tariffs of the form $\tau_{i,\,i+1}$, the definition of τ_{ij} according to (5) includes that information from both time-scale *i* and *i*+1 is used in the tariff. The third column in (7) represents the only remaining possible exception in the form of a (hypothetical) (τ_{23}, τ_{43}) tariff (i.e. a tariff mapping input from time-scales 2, 3 and 4 to output on time-scale 3). But this case is of minor importance, because it can easily expressed in terms of an equivalent (τ_{23}, τ_{44}) tariff[1].

Finally, note that a rough characterization of any tariff scheme is given by indicating the combination of all non-trivial τ_{ij}'s. E.g. we will later see that flat fee schemes are an example of (τ_{33}) tariffs, whereas traditional telephone tariffs belong to the class $(\tau_{22}, \tau_{33}, \tau_{44})$ etc.

3.5 Tariff Examples

Here are some examples of well-known tariffs revisited in the new framework. It is useful to have some idea about basic time-units on the different time-scales. Let us assume, contracts are on a 12 month base (hence basic time-unit on time-scale 4 is 1 year), payments take place every month (i.e. basic time-unit for time-scale 3 is 1 month), applications run maybe one minute (short phone calls, E-mail transfer) or one hour (video conferencing), and the communication processes of time-scale 1 are based on milliseconds.

3.5.1 Subscription only

This tariff consists of a one-time subscription charge *S* that allows unlimited resource usage for the duration of the contract. There are no measurements at all (because unlimited resource usage is guaranteed, and the tariff looks like

1. where e.g. the subscription fee is paid as a one-time charge instead of a monthly fee, cf. Section 3.5.1 for further details

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

i.e. $\tau_{44} = \tau_{44}(x^{(i)}, x^{(j)}, t, n) \equiv 1$ independent of any input parameters, time, utilization etc. With $\Pi = \begin{bmatrix} 0 & 0 & 0 & S \end{bmatrix}$ we end up with charging $C = \begin{bmatrix} 0 & 0 & 0 & S \end{bmatrix}$, i.e. paying S once per contract.

This scheme violates the requirement of paying monthly. Note that we can easily migrate to monthly billing if the contract has a limited duration of $x_1^{(3)} = h$ months. In this case,

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{h} & 0 \end{bmatrix} \tag{9}$$

with $\Pi = \begin{bmatrix} 0 & 0 & S & 0 \end{bmatrix}$ and $C = \begin{bmatrix} 0 & 0 & \frac{S}{h} & 0 \end{bmatrix}$ (each month, $1/h$ of the total subscription S has to be paid, when the contract is running for t months).

3.5.2 Leasing

A variation of the subscription-only scheme is termed leasing. Here, the user pays a monthly fee, but only for a limited number h of months, whereas the contract itself has no limited duration, i.e. after paying the total leasing charge L the service is for free. In this case, the tariff scheme for month ν shows an entry in the lower secondary matrix as

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1_{\nu \le h}}{h} & 0 \end{bmatrix} \tag{10}$$

with the usual indicator function $1_{\nu \le h}$, because the total leasing fee L is associated to time-scale 4 whereas the monthly fee is associated to time-scale 3. Thus we have $\Pi = \begin{bmatrix} 0 & 0 & L & 0 \end{bmatrix}$, hence $C = \begin{bmatrix} 0 & 0 & c_3 & 0 \end{bmatrix}$ with $c_3 = \sum^{4} \tau_{i3}\pi_3 = \tau_{43}\pi_3 = \frac{L}{h}$ during the first h months and $C = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$ afterwards.

3.5.3 Flat Rate

This tariff consists of a monthly flat rate F that allows unlimited resource usage during that month. This corresponds to

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{11}$$

with $\Pi = \begin{bmatrix} 0 & 0 & F & 0 \end{bmatrix}$, hence $C = \begin{bmatrix} 0 & 0 & F & 0 \end{bmatrix}$. The difference to the second case of the "subscription only scheme" is due to the fact that the flat rate scheme has no a priori duration of the contract.

3.5.4 Subscription + Flat Rate

This is an example how "basic tariff schemes" may be combined to yield more complex ones. Imagine the user has to pay S once for being connected to the Internet at all and an additional monthly flat rate of F. Then

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{12}$$

with $\Pi = \begin{bmatrix} 0 & 0 & F & S \end{bmatrix}$ and $C = \begin{bmatrix} 0 & 0 & F & S \end{bmatrix}$.

3.5.5 Volume-based Schemes

Let us consider a user running N identical applications (i.e. within one service type) that are characterized by their bandwidth $B = B(\theta)$ (that may vary over time θ) and their duration h, and assume the charge for this user shall be volume-based. Then we have a (τ_{22}) tariff with

$$\tau_{22} = \int_0^h B(\theta) d\theta \tag{13}$$

where τ_{22} basically represents the volume consumption per application. For applications with fixed bandwidth B (e.g. reserved bandwidth), (13) reduces to

$$\tau_{22} = \tau_{22}(B, h) = B \cdot h \tag{14}$$

With $\Pi = \left[0 \ \pi_2 \ 0 \ 0\right]$ (i.e. π_2 being the price for one unit of volume) we get

$$c_2 = N \cdot \tau_{22}(\pi_2) = N \cdot B \cdot h \cdot \pi_2 . \tag{15}$$

Note that if the applications are different, we have for application ν

$$\tau_{22}(\nu) = \tau_{22}(B(\nu), h(\nu)) = B(\nu) \cdot h(\nu) \tag{16}$$

and

$$c_2 = \sum_{\nu}^{N} \tau_{22}(\nu)\pi_2 = \sum_{\nu}^{N} B(\nu) \cdot h(\nu) \cdot \pi_2 . \tag{17}$$

3.5.6 Packet-based Auction Schemes

Assume before transmitting packet ν the user has to win an auction determining the price for the packet, i.e. $\pi_1(\nu)$. With N being the total number of packets transmitted, this gives a (τ_{11}) tariff with $\tau_{11} = 1$, $\Pi = \left[\pi_1(\nu) \ 0 \ 0 \ 0\right]$ and

$$c_1 = \sum_{\nu}^{N} \tau_{11}\pi_1(\nu) = \sum^{N} \pi_1(\nu). \tag{18}$$

Note that due to this variability in the price vector, it is definitely disputable whether auctioning is a tariff method at all (as in some sense there is no “tariff”, but only dynamic pricing).

3.5.7 ECN-based Pricing

According to [5], a pricing scheme based on Explicit Congestion Notification (ECN) marks has the general form of price π_1 per ECN mark times number of ECN marks received. This is clearly another (τ_{11}) tariff with $\tau_{11} = \sum_{\nu}^{N} 1_{(ECN)}(\nu)$, $\Pi = \begin{bmatrix}\pi_1 & 0 & 0 & 0\end{bmatrix}$ and $c_1 = \tau_{11} \cdot \pi_1$, where the for packet number ν the “ECN indicator function” is equivalent to the ECN bit itself, i.e.

$$1_{(ECN)}(\nu) = \begin{cases} 1 & \text{if the packet has an ECN mark} \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

indicates whether the ECN bit of the packet is set or not.

3.5.8 POTT (Plain Old Telephony Tariff)

The traditional well-known tariff for telephony is composed of several factors. Usually, there is a one-time subscription fee *S* caused by being assigned a number and the physical line being unlocked. Moreover, there is a monthly basic fee *F*, and there is a fee for each call, depending on its holding time *h*, the distance δ and the time of day t_d. Altogether, this yields a $(\tau_{22}, \tau_{33}, \tau_{44})$ tariff as follows:

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \tau_{22}(t) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{20}$$

where

$$\tau_{22}(t) = \tau_{22}(h(t), \delta(n), t_d(t)) \tag{21}$$

where $\tau_{22}(t)$ usually is linear in the call holding time, i.e. of the form

$$\tau_{22}(t) = h(t) \cdot \tilde{\tau}_{22}(\delta(t), t_d(t)), \tag{22}$$

with $\Pi = \left[0 \ \pi_2 \ F \ S\right]$, where the unit price $\pi_2(\delta(n), t(n))$ depending on distance and time-of-day usually are taken from a more or less transparent predefined table.

3.5.9 Coupled Tariffs

So far, essentially we have investigated tariffs with uncoupled time-scales, i.e. with zero functions in the secondary diagonals (except for Section 3.5.2 with its (τ_{43}) tariff). Our examples so far included tariffs of classes (τ_{ii}), i = 1, ..., 4, as well as combinations of them like POTT as representative of class $(\tau_{22}, \tau_{33}, \tau_{44})$. Now we are presenting an example of class (τ_{12}), where time-scale 1 and 2 are coupled in some sense.

For example, a mediator as described in the architecture of the ICCAS [1] reduces the amount of metered data e.g. by extracting their relevant statistical information. Therefore, a mediator could be expressed as a mapping of specific data on single packets to some sort of statistical values describing the flow characteristic of the respective application.

As an example, let us assume that a mediator meters each single packet of an application (e.g. an ftp transfer) and reduces this whole information to one parameter, i.e. the mean bandwidth consumption of this application. Therefore, let $x_1^{(1)}, x_2^{(1)}, \ldots, x_K^{(1)}$ be the sizes of the individual packets belonging to the application and h the duration of the application. Then

$$y_1^{(2)} = y_1^{(2)}(x_1^{(1)}, x_2^{(1)}, \ldots, x_K^{(1)}; h^{(2)}) = \frac{\sum_{\kappa=1}^{K} x_\kappa^{(1)}}{h} \tag{23}$$

represents the mean bandwidth consumption of the application in terms of kbit/s which is relevant for time-scale 2, i.e. the application-relevant time-scale. In this sense, $y_1^{(2)}$ is an example for a mapping from the atomic to the short-term time-scale, i.e. from time-scale 1 to time-scale 2.

Together with $\Pi = \begin{bmatrix} 0 & \pi_2 & 0 & 0 \end{bmatrix}$, where π_2 describes the price for bandwidth unit, *N* the number of applications, and the holding time *h* of application,

$$T = \begin{bmatrix} 0 & h \cdot y_1^{(2)} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{24}$$

is the equivalent to the volume-based scheme described by (13) and (15), but now based on input data from time-scale 1, even if the charging is restricted to time-scale 2.

Similarly, a coupling (τ_{23}) tariff may be constructed that aggregates measured applications in order to charge them monthly.

3.5.10 Preliminary (and Somehow Provocative) Conclusion

Summarizing Section 3.5, we have to note that it is possible to present examples for all types of relevant τ_{ij} tariffs as well as combinations of them. Even if the design of individual tariffs may be subject of further development (especially with respect to including new input parameters like delay, jitter etc.), the characteristic structure and principles for each τ_{ij} class will remain stable. In this sense, we have to take notice of the fact that, within this framework, apparently ***the design of new tariff schemes that look significantly different from the existing ones is not possible***.

4 Cumulus Pricing as an Example of a Tariff with Delayed Reaction

4.1 Tariff Reaction as Orthogonal Dimension

So far, in the case of dynamic tariffing (i.e. tariffs which may vary over time) the reaction of the tariff to the variation of input parameters was supposed to be immediate. E.g. with packet-based auctions in Section 3.5.6, the price was allowed to change from packet to packet, but each packet was associated with the price that has been actually valid at the moment (corresponding to t in (5)) the packet was issued. In the same sense, with volume-based pricing the price for a bandwidth unit could change from application to application, but was always associated with the actual running application. We will now relax this restriction in the sense that generally the situation at time t (on whatever time-scale) does not necessarily influence the price for that time t, but also the price for same later moment. Therefore, the input parameters of time t may cause a reaction only for a later period. E.g. Figure 4 sketches a tariff where the original input parameters come from time-scale 1 (red), are transformed by two subsequent tariffs τ_{12} and τ_{23} to yield some (symbolic) immediate charge which is depicted as red and green coins, but the relevant charge itself is affected only at a later stage through an additional tariff τ_{23}^*.

In this way, we add a new orthogonal dimension to the tariff schemes presented in Section 3. Whereas we may characterize these tariff schemes as “immediate tariffing schemes”, because the prices there are associated with the current system state, the new dimension allows for “delayed tariffing schemes” because now the current situation influences the prices for the following time as well. There is one case to be mentioned especially, i.e. if the actual situation has no influence on the actual prices at all, but only on the prices for the subsequent period. For the rest of this report we will focus on this special case, i.e. the delayed tariffing in a closer sense, and consider the characteristics of such a scheme by investigating the example of Cumulus Pricing.

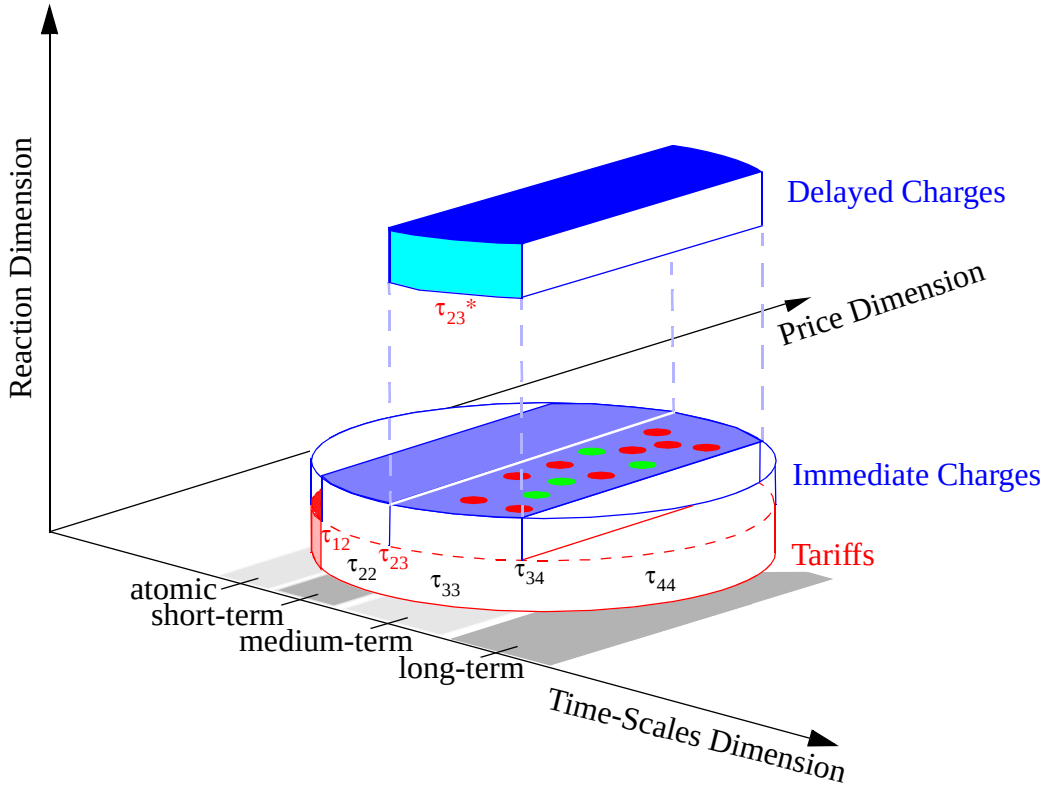

Figure 4: Tariff with Delayed Reaction

4.2 General Idea of Cumulus Pricing

The Cumulus Pricing Scheme CPS is described in [1] with great detail. Therefore, at this point we will only include a short refresher section on the general idea and provide the mathematical formulation as basis for our further investigation of selected topics. In this sense, CPS is basically a flat rate scheme (but rates may vary over long time-scales), it provides a feedback mechanism to bring market forces into play (where this feedback is not an immediate one, but requires the accumulation of discrete "flags" according to user behavior), and it allows a huge flexibility in terms of the technical prerequisites for metering and accounting mechanisms.

Characteristic to this scheme is the combination of an initial contract between customer and ISP (which contains information about expected usage patterns) with a feedback mechanism that interacts with the customer behavior on different time-scales. With CPS, *measurements* take place *over a short time-scale* and allow evidence about *user behavior on a medium time-scale*. This evidence is expressed in terms of discrete "Cumulus Points" (CPs), yet not triggering some sort of *reaction* by themselves, but only as a result of their accumulation *over a long time-scale*.

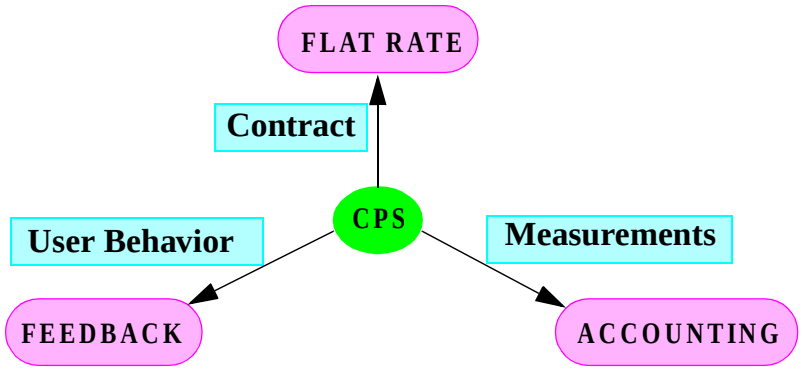

Figure 5: Basic Elements of Cumulus Pricing

The assignment of CPs works as follows: First of all, customer and ISP are supposed to agree on a contract specifying the expected user requirements in terms of bandwidth, delay etc. as well as a flat rate to be paid for this type of service. Following this agreement, the factual usage may not match the prediction given by the user (for whatever reason, be it e.g. an incorrect statement, changing habits, or new applications). As soon as these discrepancies exceed some threshold, the user receives feedback in terms of the mentioned CPs. They exist as red and green flags: a red CP indicates that the user has been overusing her capacities, a green one indicates the opposite, i.e. that the user might have been allowed to use more resources than she actually did. The larger the discrepancy between contract and reality, the more CPs may be assigned. CPs remain valid for a dedicated number of consecutive billing periods, and it is their accumulation that finally triggers certain consequences. Hence, receiving CPs requires no immediate reaction. However their successive accumulation over consecutive billing periods eventually may exceed a CP threshold and have consequences for the user, depending on ISP policies.

Figure 6 describes a typical example of how CPs are used. Customer *C* has stated her expected bandwidth requirements to be *x* MB/s, but the actual bandwidth consumption exceeds the agreed upon one slightly in January and heavily in February. Accordingly the consumer receives one red CP at the end of January and two additional red CPs at the end of February. Afterwards, her consumption falls below the expected value (one green CP in March), before it behaves exactly according to the contract in April (which is apparently the ideal situation). Later on, in May and

June this value is exceeded again. The accumulation of the CPs as of end of June sums up to five red CPs and eventually requires a renegotiation of the original contract.

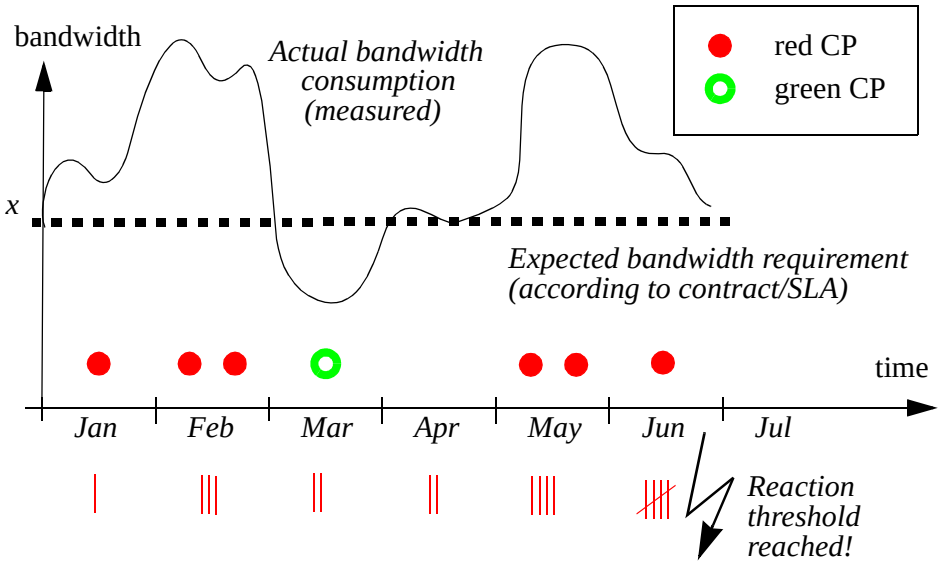

Figure 6: Red and Green Cumulus Points and Their Accumulation over Time

The time-scale view of CPS is sketched in Figure 7, mapping the basic CPS structure of Figure 5 into the time-scales framework of Figure 2. This mapping allows another description of the "Feasibility Problem" introduced in Section 2.2. According to Figure 7, the ultimate reason for the Feasibility Problem comes from the tension between the time-scales of the basic requirements RT-1 to RT-3 and the basic elements (i.e. tools) for operating a pricing scheme, i.e. measurements, user behavior and contract. Therefore, any solution to the Feasibility Problem has to be based on a suitable reconciliation between requirements and tools in terms of the relevant time-scales.

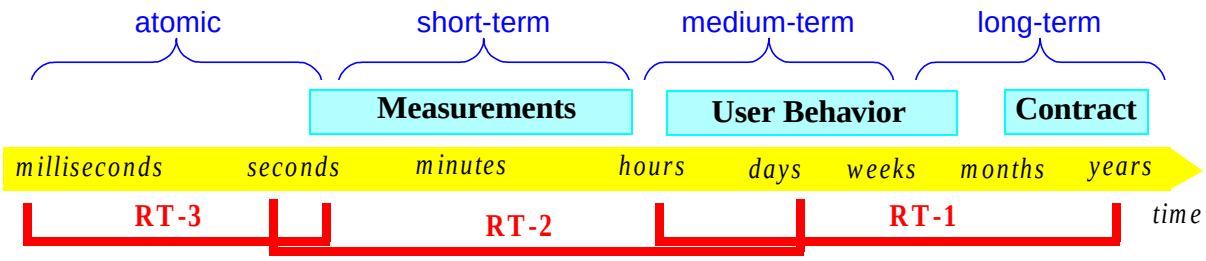

Figure 7: Time-Scale View

4.3 Formal Description of Cumulus Pricing

Following [1], we shortly describe the Cumulus Pricing Scheme formally:

Suppose that ISP *I* offers only one service, and initially customer *C* has stated her expected bandwidth requirements to be *x* MB/s, whereupon ISP *I* has offered a flat rate tariff of *a* $/month which customer *C* has accepted. In reality, the volume consumed by *C* is described by a function *V(t)* of time, which naturally may differ arbitrarily from the stated expected requirement *x*. Let $\Delta_i = \Delta(t_i)$ describe the monthly over- or under-utilization, respectively, of the customer with respect to her statement *x*, i.e.

$$\Delta_i = \int_{t_i}^{t_i} (V(t) - x)dt = \int_{t_i}^{t_i} V(t)dt - x(t_i - t_{i-1}) \tag{25}$$

where t_i describes the end of measurement 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 *I* according to a rule (the so-called "CP Rule") whose content is up to the ISP, but typically might look like the following:

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 measurement period. Then for measurement period i, the customer is assigned* c_i *cumulus points iff*

$$0 \le \theta_{c_i} \le \Delta_i < \theta_{c_i+1} \text{ or} \tag{26}$$

$$\theta_{c_i-1} < \Delta_i \le \theta_{c_i} \le 0\,, \tag{27}$$

the choice between (26) and (27) depending on $\operatorname{sgn}\Delta_i$.

Hence, if Δ_i is positive (i.e. overuse in period i) and lies between thresholds θ_c and θ_{c+1}, then *c* cumulus points are assigned. If Δ_i is negative and between thresholds θ_{c-1} and θ_c, then *c* cumu-

lus 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

$$\Gamma_n = \sum^{n} c_i , \tag{28}$$

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

The reaction to CP accumulation is again basically up to the ISP and is the content of a second rule, the so-called “Reaction Rule”, typically looking like this:

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

$$\left|\Gamma_n\right| \geq \Theta . \tag{29}$$

Depending on $\operatorname{sgn}\Gamma_n$, there may as well be two different thresholds Θ^+ and Θ^- 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 [1].

The rest of this report deals with some specific theoretical considerations about possible implementations of this pricing scheme.

5 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 "quasi-flat rate" $p(x)$ for this request (see Figure 8 for 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].

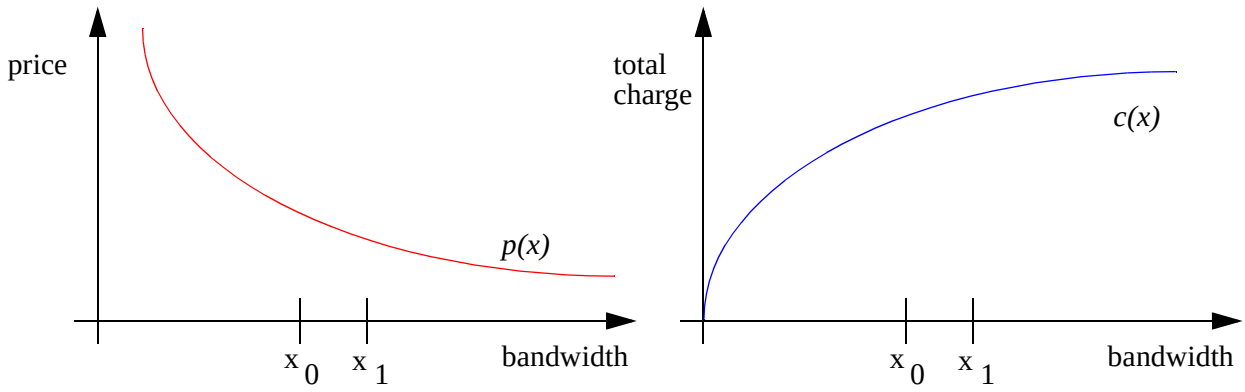

Figure 8: The Tariff Function $p(x)$ (left) and the resulting total charge $c(x)$ (right) according to (30) as a function of bandwidth consumption

1. It is intended to process the (somehow informal) SLAs automatically, therefore it is of interest to keep the number of parameters low. To this end, we aim 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 (30): in the case of bandwidth or volume, m as the total charge is obviously of the form " bandwidth $\times$ price per bandwidth unit ", whereas for parameters like delay or jitter the form of c is not immediately clear.

Assume the customer's expectation 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

$$c(x) = x \cdot p(x) \tag{30}$$

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

$$\delta = x_1 - x_0, \tag{31}$$

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 δ and hence on $p(\delta)$ for the period during which CPs have been accumulated. As the provider would like to drive the customer to behave incentive compatible, i.e. state her expected requirement according to her real expectations, the compound charge $c(x_0) + c(\delta)$ has to exceed the charge $c(x_1)$ for the correct statement. This difference between the charge $c(x_1)$ to be paid if the required bandwidth has been stated correctly and the compound charge consisting of the flat rate $c(x_0)$ for the stated required bandwidth and the extra charge $c(\delta)$ serves as "penalty function" $\Psi(x_0, x_1)$.

Hence define

$$\Psi(x_0, x_1) = c(x_1) - [c(x_0) + c(\delta)] = x_1 \cdot p(x_1) - [x_0 \cdot p(x_0) + \delta \cdot p(\delta)] . \tag{32}$$

(33)

Then we have the following requirements:

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. The tariff function should certainly not increase as in this case, instead of buying a large piece, the customer might buy a couple of smaller ones.

1. $c(x) = x \cdot p(x)$ is a monotonically increasing function, i.e. higher total bandwidth consumption yields higher total charge (see Figure 8 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 for 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. $\left|\Psi(x_0, x_1)\right| < \left|\Psi(\beta x_0, \beta x_1)\right| \leq \beta \cdot \left|\Psi(x_0, x_1)\right|$ for $\beta > 1$, i.e. a scaling property for increasing bandwidth: for high bandwidths, the same relative estimation error yields a penalty that is 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 σ_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 we end up with the square-root law

$$\Psi(Nx_0, N(x_0 + \delta)) = \Psi(Nx_0, Nx_0 + N\delta) \sim \frac{N\delta}{\sqrt{N} \cdot \sigma_0} = \sqrt{N}\frac{\delta}{\sigma_0} \sim \sqrt{N} \cdot \Psi(x_0, x_0 + \delta) \ . \quad (34)$$

Investigation of different candidates for $p(x)$ yields the following results:

- $p(x) = ax + b$ with $a \neq 0$: there is no linear function that is strictly antitonic and always positive, hence contradiction to requirement 0.

- $p(x) = \frac{1}{x}$: $\Psi(x_0, x_1) = \frac{x_1}{x_1} - \left\lceil \frac{x_0}{x_0} + \frac{\delta}{\delta} \right\rceil = -1$, contradicting requirement 3.

- $p(x) = \frac{1}{x^2}$: $m(x) = x \cdot p(x) = \frac{1}{x}$ monotonically decreasing, contradicting requirement 1.

- $p(x) = 1$: $\Psi(x_0, x_1) = x_1 - [x_0 + \delta] = 0$ contradicting requirement 2.

From these examples, we see that the class of possible candidates is $p(x) = \frac{1}{x^\alpha}$, $0 < \alpha < 1$. Having requirement 4 in mind, we propose therefore finally

$$p(x) = \frac{1}{\sqrt{x}}. \tag{35}$$

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:

It is well-known that $p(x) = \frac{1}{\sqrt{x}}$ is strictly antitonic and always positive.

Requirement 1 results from $c(x) = \sqrt{x}$.

Assume $\delta = x_1 - x_0 > 0$. Due to the binomial formula, $(a+b)^2 = a^2 + 2ab + b^2 > a^2 + b^2$ if $a, b > 0$, hence $a = \sqrt{x_0}$ and $b = \sqrt{\delta}$ yield $(\sqrt{x_0} + \sqrt{\delta})^2 > x_0 + \delta$, 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 δ 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. ❑

Having proved proposition 1 for $p(x) = \frac{1}{\sqrt{x}}$, it is worth to be noted that the same proof applies if the tariff function is scaled by an arbitrary positive factor $\lambda > 0$, i.e.

$$p_\lambda(x) = \frac{\lambda}{\sqrt{x}}. \tag{36}$$

In this case, we have to substitute c and Ψ 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 obviously does not concern the validity of the proof. Therefore, λ provides an additional degree of freedom we will come back to later.

6 Distance between Thresholds for the CP Rule

According to the precise mathematical description as presented in Section 4.3, Cumulus Points are assigned in close connection with violating certain thresholds. Now, the obvious question to be answered is how to define suitable thresholds. In this context, The following requirements are aimed at:

1. Use only a small number of thresholds either way (i.e. no more than 3 – 5).

2. Incorporate a hysteresis effect, i.e. prevent small oscillating deviations from resulting in unnecessary awards of CPs.

3. Choose the thresholds in a way that the resulting Cumulus Points are widely independent of the measurement method applied (i.e. the difference in terms of Cumulus Points must not exceed 1 if different measurement methods are used).

For the following considerations, assume that *V(t)* can be described by a stochastic process which is in equilibrium[1], fluctuating around mean ξ with (unknown) standard deviation σ. Any measurement performed on the resource then aims at estimating ξ, in order to be able to decide

between which two of the thresholds $\xi - x$ is lying. Estimating ξ with unknown σ from a normally distributed process is the one of the basic problems of the so-called statistical confidence estimation [6 p.685].

Assume X_n, n = 1, ..., N, to be sample values for *V(t)*. Define $\bar{X} = \frac{1}{N}\sum^{N} X_n$ 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 the term $\frac{\bar{X} - \xi}{S}\sqrt{N}$ suffices a Student-t distribution with $m = N - 1$ degrees of freedom. Without going too much into detail, this means the following:

Let α signify the confidence level of the estimation, i.e. the estimation procedure yields a so-called confidence interval $(\bar{X} - \varepsilon, \bar{X} + \varepsilon)$, and ξ lies within this interval with probability $1 - \alpha$. E.g. for α = 5%, the probability that ξ lies within the resulting interval equals 95%. Now, having chosen α, there are tables (e.g. [6 p. 22]) where it possible to determine a number $t_{\alpha,N-1}$ such that the interval

$$(\bar{X} - \varepsilon_{\alpha,N}, \bar{X} + \varepsilon_{\alpha,N}) \tag{37}$$

with

$$\varepsilon_{\alpha,N} = \frac{S}{\sqrt{N}} t_{\alpha,N-1} \tag{38}$$

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

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 starting point for reasons of mathematical tractability.

This investigation yields directly

Proposition 2:
Under steady-state assumptions, requirement 3 is fulfilled if every two neighboring thresholds have at least distance $2 \cdot \varepsilon_{\alpha,N}$.

Proof: see above. ❏

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

Moreover, proposition 2 solves also the issue of requirement 2: If the distance between θ_{-1} and θ_1 is larger than $2 \cdot \varepsilon_{\alpha,N}$, then there is no danger for getting an estimation of ξ that exceeds either threshold, hence in this case there won't be awarded any CPs.

Note for all these results that they always depend on the confidence level $1 - \alpha$, i.e. are only able to provide statistical guaranties. 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 simulation approach.

At the present stage, there is no obvious way to determine the size of the confidence interval and hence the distance between thresholds as described by proposition 2. Note, however, that (38) may be calculated using simulation results, where traffic sampling allows easily to calculate $\overline{X}$ and S in order to derive $\varepsilon_{\alpha,N}$. In fact, evaluating (38) in this way is one of the major tasks for the CPS simulation as described below.

7 Implementation Aspects

CPS as described and investigated so far basically is a framework that can be implemented in different ways. Implementations may e.g. differ by

- the way the initial contract is settled, especially in which way and to which detail the characteristics of expected traffic is provided;
- the way number and sizes of thresholds are determined for the CP Rule as well as for the Reaction Rule (i.e. the threshold indicating that the contract has to be renegotiated due to an obvious inconsistency between expected traffic and measured traffic as indicated by the number of red or green Cumulus Points);
- the type of interface between CPS and customer or ISP, respectively;
- the way legal aspects are taken into account, especially the question of non-repudiation of measurements.

This section presents some initial ideas about a possible implementation which is currently undertaken. Main issues considered here include the investigation of expected customer behavior during an initial “Probe Phase” before starting the actual charging scheme, a possible derivation of thresholds using the results of this Probe Phase and presents the first version of a GUI which may be used for simulation purposes.

7.1 The Probe Phase (PP)

As already stated above, CPS requires an initial contract (or SLA, respectively) between service provider and customer that contains the customer’s expected traffic characteristics. Finding out these characteristics may be done in different ways, e.g.

- the customer is offered to use the resources for free during one or more time units, whereas the provider uses this period to measure rather detailed the probable customer behavior, where the parameters on which the initial contract is based are provided by the ISP according to his findings;

- the customer is offered a special tool that she may use to determine the current traffic characteristics and extrapolate them into the future, where the parameters on which the initial contract is based are provided by the customer according to her findings – this approach could be viewed as outsourcing the parameter determination, as it is not included within the CPS implementation;

- the customer is allowed 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 then are used to determine the parameters on which the initial contract is based;

- any mixture of the above, especially approaches where both ISP and customers influence the parameters for the initial contract.

Offering the service initially for free and performing measurements during that period is a straightforward approach, but unfortunately practical experience (e.g. from the INDEX project) shows that in this case the users tend to strongly overuse the resources due to the fact that they are for free (e.g. by staying connected all the time) and change their behavior drastically as soon as they know they are charged now. In fact, it has turned out that there is no reliable way to relate such a free phase to the actual expected requirements, and therefore within INDEX this period could only be used to support the users becoming familiar with the system.

In terms of the legal issues, especially the question of non-repudiation, it appears to be reasonable that the customer should at least to some extent be involved within the process of parameter determination. According to the original CPS proposal, it is in fact the customer alone who is responsible for finding out his expected traffic characteristics, but obviously she might need some help while doing so. One possible approach might use outsourcing this problem to a third party or to some special tool offered to the customer by the ISP, but running under the responsibility of the customer. In this case, a detailed statistical evaluation about the current user behavior should be performed. Two basic parameters resulting from these measurements in our example are the mean value ξ and the standard deviation σ of the current resource requirements.

In the same way, ξ and σ could also be determined by the ISP during a so-called “Probe Phase” (PP). For the reasons presented above, the PP cannot be for free, but charging the customer a small amount nevertheless in a volume-based manner should be in the interest of both parties. The results of the PP, especially all measurements, could be made 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 a certain influence on the reliability of the traffic expectation, but the last responsibility is still with the customer. Typically, the duration of the PP should be small with respect to the duration of the contract, e.g. for our example of Section 4.2 it could be one month (i.e. month number -1). In our simulation, we use this approach in the form of a separate module yet integrated within the CPS.

Note that the PP approach already is based on customer and ISP participating in the parameter determination. There are a couple of other possibilities for mixing both parties, e.g. by performing independent measurements on both sides and averaging or combining them according to some predefined proportional scheme etc. In every case, it should be possible to determine at least mean and standard deviation of the expected traffic in a way that these numbers allow deriving the basic parameters of the initial contract in a way both customers and providers can trust.

7.2 Choosing Thresholds for the CP Rule

In Section 6, 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 there are some reasons for choosing a relatively low number of CPs that can be awarded for one time unit (perhaps 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 $\theta_n = 110\%$ (of ξ, e.g.);

- 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 σ into the process of calculating thresholds.

As CPS generally provides each customer with very individual contracts (i.e. SLAs), this provides at least the possibility to adjust also the 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 defacto 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 σ into the threshold determination. This is based on the following considerations.

Assume once again that the actual resource consumption may be viewed as a steady-state stochastic process. The main aim of awarding CPs consists of indicating whether the statement about expected resource consumption (still) is valid. A simple heuristic could work as follows: If ξ and σ 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 approx. 68.3% of the samples should be contained within the interval $[\xi - \sigma, \xi + \sigma]$, approx. 95.4% within $[\xi - 2\sigma, \xi + 2\sigma]$, approx. 99.7% within $[\xi - 3\sigma, \xi + 3\sigma]$ etc. Hence, if the sample measurement is "too often" outside $[\xi - \sigma, \xi + \sigma]$, this is an indication that maybe stating mean ξ 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 manner to use relative thresholds of the form

$$\theta_{\pm n} = 1 \pm n \cdot \frac{\sigma}{\xi} \tag{39}$$

i.e. one CP is awarded as soon as the measurement differs from ξ by more than σ, two CPs are awarded if the deviation is larger than 2σ etc. As an additional nice feature, this approach limits inherently the number of thresholds to be about 3 or 4 in each direction.

The evaluation of this proposal is left to simulation. A straightforward generalization of this approach yields non-linear thresholds of the more general form

$$\theta_{\pm n} = 1 \pm \gamma_n \cdot \frac{\sigma}{\xi} \tag{40}$$

with γ_n being any positive real numbers. Setting $\gamma_n = n$ reduces (40) to (39). Maybe it could be reasonable to set the first threshold not too far away from the expected mean in order to support the early warning feature of CPS, e.g. $\gamma_1 = 1{,}3$ (with one sample out of 10 exceeding this threshold), the second threshold at $\gamma_2 = 2{,}4$, where deviations of this order may still occur regularly (in about 1% of the cases), the third one at $\gamma_3 = 3{,}1$ where exceeding this threshold is rather improbable (approx. 0.1% of the cases) and maybe a fourth one at $\gamma_4 = 3{,}7$, i.e. at a point where only one sample out of 10,000 normally exceeds this threshold. Once again, the validation of these suggestions is left to simulation.

Nevertheless, this approach contains a fundamental problem. Assume two customers have contracts based on identical expected mean requirements $\xi_1 = \xi_2$, but differing standard deviations, e.g. $\sigma_2 \gg \sigma_1$. Further assume that both customers deviate from their expected requirements by an identical absolute value $\gamma_1 \cdot \sigma_1 < \delta < \gamma_1 \cdot \sigma_2$. In this case, customer 1 is awarded one CP whereas customer 2 stays o.k., despite of the fact that customer 1 in general behaves much more honestly (i.e. much less irregularly) than customer 2.

There is no obvious direct solution to this problem. Basically, if customer 1 has a deviation that is relatively large (compared to her usual behavior), this is justly recognized as strong indication that the initial expectation might be no longer valid, and there is no reason for customer 1 to complain.

Moreover, there is one possibility to account for this apparent iniquity, i.e. in terms of the tariff function. If the tariff function e.g. depends on σ, then it is rather straightforward to charge a customer with large fluctuations generally higher than a customer without fluctuations. In this sense, in (36) we may set $\lambda = \lambda(\sigma)$, e.g. $\lambda(\sigma) = \sigma$, resulting in a tariff function of the form

$$p_{\lambda(\sigma)}(x) = \frac{\sigma}{\sqrt{x}}, \tag{41}$$

respectively. If this suggestion yields satisfying results is another issue left open to simulation.

7.3 Further Research Issues

Summarizing Section 7, here is a list of open questions to be answered by simulation:

1. What are realistic values for $\varepsilon_{\alpha,N} = \frac{S}{\sqrt{N}} t_{\alpha,N-1}$ (38)?

2. How essential is the steady-state assumption for $V(t)$ as supposed in proposition 2?

3. Are thresholds calculated according to $\theta_{\pm n} = 1 \pm \gamma_n \cdot \frac{\sigma}{\xi}$ (40) reasonable? Especially, are they consistent to traffic that does not follow a normal distribution?

4. If so, how are the γ_n to be determined? Is the linear version $\gamma_n = n$ or a nonlinear version like the one sketched in Section 7.2 to be preferred?

5. Are the thresholds derived from 3. consistent to proposition 2 about their mutual distance[1]? How good is therefore the bound of proposition 2?

6. Or put 4. the other way round: using thresholds derived in 3., what is the confidence level that their distances are large enough to yield CP assignment being widely independent of the mea-

1. Note that proposition 2 is a statement derived from the requirement that the assignment of Cumulus Points should be widely independent of the measurement method, whereas the threshold calculation according to (40) is supposed to be a good heuristic. Hence, the issue to be investigated here is whether both of these approaches may co-exist, especially as the bounds for distances derived from proposition 2 may probably turn out not to be close to optimal ones.

surement process, as postulated in Section 4.2 requirement 3? Is this confidence level realistic?

7. How big is the influence of the number of measurements per customer? Is there a bound on this number due to the asymptotic nature of the Student-t distribution, where it does no longer make sense to increase the number of measurements beyond?

8. What is the result of tariff functions $p_{\lambda(\sigma)}(x)$ depending explicitly on σ? Is $\lambda(\sigma) = \sigma$ as in (41) a reasonable choice, and are there better alternatives?

8 Summary and Conclusions

This report has introduced a framework for designing Internet tariffs that is explicitly focussed on time-scale aspects. After describing shortly the four Internet relevant time-scales, Section 3 started with a definition of the concepts of charges, prices and tariffs, before proposing a formal view on tariffs as multi-dimensional mappings between different time-scales. It has been demonstrated how tariffs correspond to 4x4 matrices with several vanishing components. Using a large variety of examples, this concept has been proved to be general enough to describe all types of currently existing Internet tariffs. In Section 4, this framework has been extended by a new orthogonal dimension, describing how changing input parameters might have a delayed effect on the respective charges and prices. For the resulting new class of tariff schemes, the Cumulus Pricing Scheme as introduced in [1] turns out to be a prominent example, therefore its formal description is given, before the second part of the report deals with some special features of CPS. One important particle of CPS includes a function which describes the relationship between resource requirements and prices. In order to keep the complexity of information to be transmitted during SLA negotiation low, the same function is supposed to be used for determining extra fees in case of agreement violation. A thorough analysis of requirements on this function reveals in Section 5 that its form has to obey a “one-over-square root” law. The second specific CPS aspect analyzed in detail concerns the issue of how to set CPS thresholds. Section 6 gives a partial answer to this question by deriving a proposition about the necessary distance between thresholds in order to care for non-ambiguity of the CP assignment procedure. Finally, Section 7 provides a couple of

implementation aspects, including the proposal of a “Probe Phase” prior to starting the scheme, caring for determining resource requirements between users and providers, as well heuristic approach for threshold setting which is based on sound statistical deliberations.

This report complements the existing work on CPS in the following ways:

- The tariff function provides a simple but sound proposal for assigning charges to resource requirements, especially in the case of bandwidth or volume.

- The concept of a probe phase solves the most immediate critique on CPS, i.e. the question how to prevent customers from stating a zero request for the sake of the initial contract, then using unlimited resources for free and eventually changing to a new provider instead of adapting the contract. Introducing the probe phase yields to an initial contract that is realistic and acceptable for both customer and provider. Moreover, the characterization of expected resource requirements can be performed in much more detail.

- The heuristic concatenation of CPS thresholds and the standard deviation of expected traffic characterization provides a straightforward rationale for implementing both the CP rule and the reaction rule with only a small number of Cumulus Points.

- The MTS (Methodology of Time-Scales) framework allows to embed CPS into a more general “theory of Internet tariff design” and thus provides some valuable insight into what Internet tariffs are about at all. Especially the abstract formulation of tariffs in terms of the matrix (4) gives much room for adapting existing tariff schemes to new input parameters. For instance volume-based schemes as τ_{22} tariffs are well-known, and with MTS it is possible to extend their structure e.g. to delay-based or jitter-based tariffs as well.

- Finally, a clear and well-defined research plan has been sketched about which answers are to be expected from a simulative evaluation of CPS, i.e. a validation of the feasibility of the concept and some more information about the fine-tuning of the system.

Future work has to focus on implementation aspects of CPS. Moreover, there are important non-technical aspects to be covered as well. It is to be investigated how users react to this type of pric-

ing scheme, and which legal requirements the contract and SLA negotiation have to comply with, e.g. to which extent the provider must be able to verify his measurements, and whether the consumer may be entitled for continuation of her previous service even if she does not comply to the original contract. In this sense, the basic design of CPS can be viewed as being completed, and the next steps will consist of gaining and deepening practical experience with this new Internet pricing scheme.

Acknowledgements

The authors like very much to acknowledge many helpful discussions with Placi Flury, Jan Gerke, Hongguang Ma and Hasan. This work has been performed partially in the framework of EU IST project Market Managed Multi-service Internet (M3I, IST-1999-11429) with ETH Zurich being funded by the Swiss Bundesministerium für Bildung und Wissenschaft Bern.

Reference

[1] 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 Zurich, Switzerland, August 2000.

[2] CATI - Charging and Accounting Technology for the Internet. URL: http://www.tik.ee.ethz.ch/~cati

[3] M3I - Market Managed Multi-service Internet. URL: http://www.tik.ee.ethz.ch/~m3i

[4] The INDEX Project. URL: http://www.index.berkeley.edu

[5] M. Karsten (edt.): *Pricing Mechanism Design (PM)*. M3I Deliverable 3, version 1.0, June 30, 2000.

[6] I. N. Bronstein, K. A. Semendjaev: Taschenbuch der Mathematik. 23rd edition. Verlag Harri Deutsch, Frankfurt am Main, 1987.