

Nil nove sub sole?

Why Internet tariff schemes look like as they do

Peter Reichl¹, Burkhard Stiller²

¹Telecommunications Research Center Vienna, FTW, Maderstraße 1, A-1040 Vienna, Austria

²Computer Engineering and Networks Laboratory TIK, ETH Zürich, Gloriastrasse 35, CH-8092 Zurich, Switzerland

E-Mail: reichl@ftw.at, stiller@tik.ee.ethz.ch

Abstract

Pricing schemes for Internet services have turned out to be of crucial influence for the competitiveness of Internet Service Providers (ISP). Despite of this urgent demand, there is still no standard tariff scheme that solves the so-called “feasibility problem”, i.e. ensures simultaneously technical feasibility, economic efficiency and user acceptance. This paper investigates the reason for the apparent shortage of “new schemes under the sun” and deals with the inherent time-scale based structure of Internet pricing schemes. Tariffs are viewed in terms of mappings between different time-scales that can be expressed in the form of Tariff Matrices with a particular tri-diagonal form. It turns out that all established proposals fit perfectly into the resulting framework, and that on the other hand there is no room for novel approaches left until a new (temporal) dimension is added. Introducing the dimension of “delay of charging” into the framework gives rise to a completely new class of dynamic tariffs, as is shown by means of two promising examples for the resulting schemes, i.e. the Cumulus Pricing Scheme (CPS) for the Differentiated Service (DiffServ) architecture and the Connection-Holder-is-Preferred-Scheme (CHiPS) for dynamic multiprovider auctions.

Keywords: *Internet Charging, Tariffs, Cumulus Pricing Scheme, Multiprovider Auctions*

1 Introduction

The increasing deregulation of the telecommunications market and an emerging business orientation of Internet services drive the need for appropriate pricing models for packet-based communications, which are independent of regulated aspects. Well-known and widely accepted pricing models for communication networks offering a single network service, e.g., telephony or X.25, are provider-centric, i.e. based on fixed values that are re-issued whenever provider costs or regulations change. However, in an increasingly competitive environment, this approach of charging models is too slow. Furthermore, the deregulation opens the field of pricing in particular communication services within an open market approach [12].

Many projects covering Internet charging functionality on the network level, including services and sometimes content as well, intend to achieve a complete independence from pricing models, as the selection of a suitable pricing model for the Internet remains an open problem. The major difficulty here 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. In contrast to these approaches, this paper deals with a recent proposal to shift the perspective and view the design of Internet pricing schemes not as a problem of reducing complexity, but rather as a question of multi-dimensional interdependences between time-scales.

It is interesting to note that despite of the apparently urging demand for scalable and efficient Internet pricing schemes, no satisfying standard solution has been proposed so far. In order to find out why there is “nothing new under the sun” (as the title may suggest), this paper investigates the inherent time-scale based structure of pricing models for the Internet and, presenting a matrix-based notion for describing existing proposals, discovers a number of unique characteristics, which lead to the rather provocative conclusion that it is impossible to propose structurally novel pricing schemes, unless a new temporal dimension is introduced which allows to delay the charging reaction to changing market situations.

This paper is organized as follows. Section 2 introduces time-scales and the “Feasibility Problem” for Internet tariff design. The Methodology of Time-Scales is developed in Section 3, formally defining charges, prices, tariff, the tariff matrix, and elaborating on various well-known tariff examples. The new tariff dimension “Delay” is introduced in Section 4 as well as two examples for the resulting new tariff schemes. Finally, Section 5 summarizes the work and draws conclusions.

2 Time-Scales and the Feasibility Problem for Internet Tariff Design

In [20], we have already identified four different time-scales as being relevant for Internet tariff schemes. They are summarized in the following Table 1:

Table 1: Relevant Time-Scales

Time-scale	Relevance	Typical Actions	Units
atomic	communications	packets, roundtrip times	~ms
short-term	applications	ftp, IP phone calls	sec/min
medium-term	Billing	phone bills, rents	~week
long-term	Contract	ISP-customer contracts	~year

Moreover, in [16] we have analyzed basic requirements to be fulfilled by any useful proposal for tariffing Internet services. The resulting three general requirement types (RT) are summarized in Table 2.

Table 2: Requirement Types and Characteristics

Type	Target	Characteristics
RT-1	customer	transparency, predictability
RT-2	ISP (economical)	network efficiency
RT-3	ISP (technical)	accounting technology

The quality and suitability of Internet tariff schemes depends on the balance between these three requirements. Flat rate pricing, e.g., is excellent with respect to RT-1 and RT-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 [14].

A closer analysis together with practical experiences reveals that not all three requirements are of equal importance: Any scheme that does not fulfill the criterion of technical feasibility has no chance of being realized for practical purposes. In this sense, pricing schemes may have different trade-offs between RT-1 and RT-2, whereas RT-3 represents a hard criterion. In [16], this fact has been termed the ‘‘Feasibility Problem’’ of Internet Pricing.

Now we may assign the three requirement types to appropriate time-scales. In doing so, we notice that they are apparently placed at the transition points between scales: Transparency and predictability is relevant on medium- to long-term scales, economic efficiency refers to short- and medium-term scales and the accounting technology deals with atomic and short-

term events. Thus, the problem of balancing these requirements turns out to be an issue of reconciling the different time-scales. Hence, we can now deduce a new description of the Feasibility Problem: According to Figure 1, 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 for operating a tariff scheme, i.e. measurements, user behavior and contract. Therefore, any solution to this problem has to be based on a suitable reconciliation between requirements and tools in terms of the relevant time-scales.

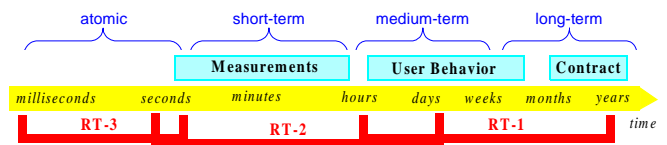


Figure 1: Time-scales, Requirements, and Tariff Elements

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 Charges, Prices, Tariffs

Prices are according to [19] defined as monetary value of a unit of delivered goods. As every time-scale as described in Section 2 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

$$\Pi = [\pi_1 \ \pi_2 \ \pi_3 \ \pi_4] \quad (1)$$

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 associated with one time-scale, e.g. various service classes w.r.t. different applications. In this case, π_2 represent-

ing 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} ¹ 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 3.2.

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

$$C = C(t) = [c_1 \ c_2 \ c_3 \ c_4] \quad (2)$$

with t denoting time-dependence and

$$c_j = \sum_{i=1}^4 \tau_{ij} \pi_j. \quad (3)$$

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-of-thumb (see Section 3.4.1 e.g.).

Note that (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 3.2, where (5) describes that any tariff depends explicitly on time). Moreover, Section 3.4.6 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’’.

Furthermore note that (3) describes charges as component-wise (Hadamard) 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.

1. we will come back to the indices in Section 3.2.

3.2 Formal Definition of a Tariff Scheme

On each of the time-scales identified in Section 2, 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):

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

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 $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_k^{(i)})$ be the vector of k input variables relevant for time-scale i^1 and $x^{(j)} = (x_1^{(j)}, x_2^{(j)}, \dots, x_l^{(j)})$ 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 $y^{(j)} = (y_1^{(j)}, y_2^{(j)}, \dots, y_m^{(j)})$ 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 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

$$\tau_{ij}: \begin{cases} \mathfrak{R}^{k+l+8} \rightarrow \mathfrak{R}^m \\ \begin{pmatrix} x^{(i)} \\ x^{(j)} \\ t \\ n \end{pmatrix} \rightarrow y^{(j)} \end{cases} \quad (5)$$

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

Figure 2 tries to sketch² this view of tariffs as time-scale dependent mappings from input parameters (here from time-scales 2 and 3) and prices to charges (here on time-scale 3).

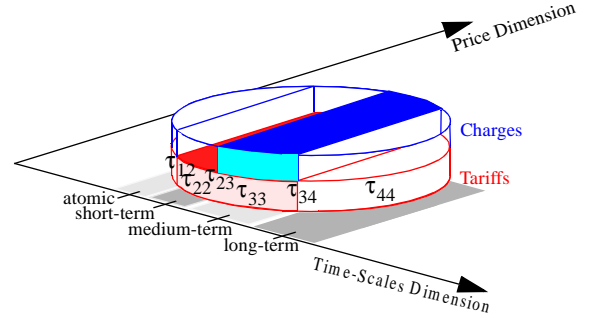


Figure 2: Tariffs as Mappings from Time-scales Dependent Input Parameters and Prices to Charges.

3.3 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

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

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 3.4.2 we will actually see an example for this type of tariff.

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

2. e.g. for the mentioned telephone tariff (monthly charge plus call-based fee)

Summarizing this argumentation, the Tariff Matrix should have the form

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

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}) tariff¹.

Finally, note that any tariff scheme now may be characterized by indicating the respective 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 in its full size belong to the class $(\tau_{22}, \tau_{33}, \tau_{44})$ etc.

3.4 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.

3.4.1 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

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (8)$$

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

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{h} & 0 \end{bmatrix} \quad (9)$$

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

3.4.2 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 v shows an entry in the lower secondary matrix as

1. where e.g. the subscription fee is paid as a one-time charge instead of a monthly fee, cf. Section 3.4.1 for further details

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1_{v \leq h}}{h} & 0 \end{bmatrix} \quad (10)$$

with the usual indicator function $1_{v \leq h}$, 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

$$\Pi = [0 \ 0 \ L \ 0], \quad \text{hence} \quad C = [0 \ 0 \ c_3 \ 0] \quad \text{with}$$

$$c_3 = \sum_{i=1}^4 \tau_{i3} \pi_3 = \tau_{43} \pi_3 = \frac{L}{h} \quad \text{during the first } h \text{ months and}$$

$$C = [0 \ 0 \ 0 \ 0] \quad \text{afterwards.}$$

3.4.3 Flat Rate

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

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (11)$$

with $\Pi = [0 \ 0 \ F \ 0]$, hence $C = [0 \ 0 \ F \ 0]$. 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.

3.4.4 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

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (12)$$

with $\Pi = [0 \ 0 \ F \ S]$ and $C = [0 \ 0 \ F \ S]$.

3.4.5 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

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

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

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

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

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

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

$$\tau_{22}(v) = \tau_{22}(B(v), h(v)) = B(v) \cdot h(v) \quad (16)$$

and

$$c_2 = \sum_{v=1}^N \tau_{22}(v) \pi_2 = \sum_{v=1}^N B(v) \cdot h(v) \cdot \pi_2. \quad (17)$$

3.4.6 Packet-based Auction Schemes

Assume before transmitting packet v the user has to win an auction determining the price for the packet, i.e. $\pi_1(v)$. With N being the total number of packets transmitted, this gives a

(τ_{11}) tariff with $\tau_{11} = 1$, $\Pi = [\pi_1(v) \ 0 \ 0 \ 0]$ and

$$c_1 = \sum_{v=1}^N \tau_{11} \pi_1(v) = \sum_{v=1}^N \pi_1(v). \quad (18)$$

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

3.4.7 Flow-based Auction Schemes

Except for flows instead of packets being the good to be auctioned, this is a strict analogy to Section 3.4.6. Therefore assume before starting a new flow v , consisting of a sequence of individual packets, the user has to win an auction determining the price for that flow, i.e. $\pi_2(v)$. With M being the total number of flows to be sent, this yields a (τ_{22}) tariff with

$$\tau_{22} = 1, \Pi = \begin{bmatrix} 0 & \pi_2(v) & 0 & 0 \end{bmatrix} \text{ and}$$

$$c_2 = \sum_{v=1}^N \tau_{22} \pi_2(v) = \sum_{v=1}^N \pi_2(v). \quad (19)$$

3.4.8 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

$$\text{is clearly another } (\tau_{11}) \text{ tariff with } \tau_{11} = \sum_{v=1}^N 1_{(\text{ECN})}(v),$$

$\Pi = \begin{bmatrix} \pi_1 & 0 & 0 & 0 \end{bmatrix}$ and $c_1 = \tau_{11} \cdot \pi_1$, where the for packet number v the “ECN indicator function” is equivalent to the ECN bit itself, i.e.

$$1_{(\text{ECN})}(v) = \begin{cases} 1 & \text{if the packet has an ECN mark} \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

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

3.4.9 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:

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \tau_{22}(t) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (21)$$

where

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

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

$$\tau_{22}(t) = h(t) \cdot \tilde{\tau}_{22}(\delta(n), t_d(n)), \quad (23)$$

with $\Pi = \begin{bmatrix} 0 & \pi_2 & F & S \end{bmatrix}$, where the unit price $\pi_2(\delta(n), t(n))$, depending on distance and time-of-day of call n , usually is taken from a more or less transparent predefined table.

3.4.10 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 3.4.2 with its (τ_{43}) tariff). Our examples so far included tariffs of classes (τ_{ii}) , $i = 1, \dots, 4$, 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 [19] 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 $x_1^{(1)}, x_2^{(1)}, \dots, x_K^{(1)}$ be the sizes of the individual packets

belonging to the application and h the duration of the application. Then

$$y_1^{(2)} = y_1^{(2)}(x_1^{(1)}, x_2^{(1)}, \dots, x_K^{(1)}; h^{(2)}) = \frac{\sum_{\kappa=1}^K x_{\kappa}^{(1)}}{h} \quad (24)$$

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,

$$T = \begin{bmatrix} 0 & h \cdot y_1^{(2)} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (25)$$

is the equivalent to the volume-based scheme described by equations (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.

3.4.11 Intermediate and Provocative Conclusion

Summarizing Section 3.4, 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 The Dimension of Delay

The dimensions of Internet pricing schemes discussed above lack an important further alternative, the ‘‘Delay’’ dimension. Based on a general introduction of delaying charges on purpose, two examples of resulting approaches are discussed.

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 3.4.6, 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 the price at same later moment. Therefore, the input parameters of time t may cause a reaction only for a later period. E.g., starting from the situation of Figure 2, Figure 3 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 now the relevant charge itself is affected only at a later stage through an additional mapping 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 periods, i.e. the future, too. 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. In Section 4.2 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.

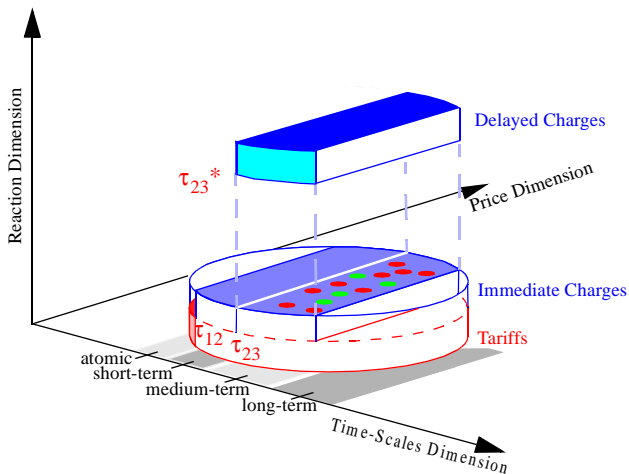


Figure 3: Tariffs with Delayed Reaction: The CPS Example

4.2 Example I: Cumulus Pricing

The recently (cf [19], [20]) established Cumulus Pricing Scheme CPS for the Differentiated Services (DiffServ) architecture is described in [16] with great detail. Therefore, at this point we will only include a short refresher section on the general idea. 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*.

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 4 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.

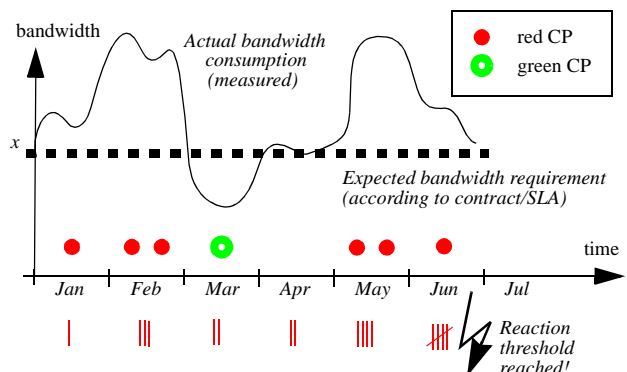


Figure 4: Red and Green Cumulus Points and Their Accumulation over Time

4.3 Example II: Multiprovider Auction Schemes

A second example for including delayed reactions into Internet pricing comes from the area of flow-based auction schemes. Here, the “Smart Market” approach introduced in [6], [7] and [8] has obtained a dominant position, due to the incentive-compatibility of the second price auctions (Generalized Vickrey Auctions, GVA) used. The characteristic property of GVAs concerns the fact that the users with the highest bids win the auction, but do not pay their bid (as one would expect), instead they are charged the highest bid submitted by a losing bidder. Applying GVAs to IntServ scenarios including connections traversing more than one provider poses a couple of additional requirements, e.g. synchronisation issues between the single auctions at the different providers and especially the fact that losing one local auction immediately may have fatal consequences for the globally established connection.

To solve these issues for a RSVP based environment, [14] has proposed the so-called Connection-Holder-is-Preferred-Scheme (CHiPS). Whereas regular GVAs decouple the win of an auction from the price to be paid for getting the resource, CHiPS goes one step further, decoupling the win of an auction also from getting the resource to be auctioned. The basic idea of CHiPS is to perform a regular GVA in order to determine the correct market price, but as far as resource assignment is concerned, connections that are already established are given a small preference, i.e. if such a connection loses one local auction due to some market turbulences, this connection should get a second chance to deliver a-posteriori a bid that would have been sufficient to win also this particular local auction. As pointed out in [14], this is possible due to the particular property of GVAs that the price to be paid for the good usually is lower than the auction bid, i.e. there is a mismatch between the global budget (spending cap) and the money that is spent indeed, which may be used for this a-posteriori increase of bids at critical auction locations.

In Section 3.4.7 we have seen that regular flow-based auctions may be described as (τ_{22}) tariffs. Figure 5 demonstrates that the a-posteriori re-bidding for established connections may be viewed as another example of delayed charging.

Whereas in the CPS case of Section 4.2, the charging has been delayed with respect to the next larger time-scale (i.e. using services on time-scale 2 has consequences on time-scale 3, and not even for the actual medium-term period, but only for the following one), in the CHiPS example we stay within the same time-scale (short-term), where using a services during one RSVP period eventually has consequences for the next RSVP period. In this sense, CHiPS is an example for a tariff of type (τ_{22}^*) .

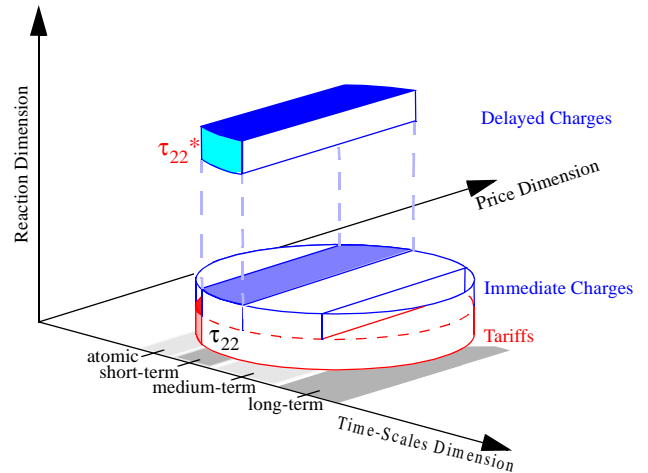


Figure 5: Tariffs with Delayed Reaction: CHiPS

As an additional remark note that [14] has demonstrated the feasibility of CHiPS in terms of implementation and overhead, but has explicitly restricted the scheme to the case of already established connections, whereas the question of how to establish a flow through multiple providers by means of an auction mechanism has been factored out so far. Recently, [2] has provided a suitable approach to solve this problem by applying an open descending multi-unit auction with prices for the single links dropping with different speeds, according to the varying demand on these links. Moreover, various dropping policies, e.g. the “decrement rates” policy, the “prize freezing” policy and the “instant price adaptation” policy are introduced and evaluated in [2]. This proposal can be applied to RSVP scenarios in a straightforward manner. Therefore, combining the descending multi-unit auction approach with CHiPS eventually forms a coherent auction mechanism for RSVP-based multiprovider environments.

5 Summary and Conclusions

This paper has introduced a framework for designing Internet tariffs that is explicitly focussed on time-scale aspects. After describing shortly all four Internet relevant time-scales, a definition of the concepts of charges, prices, and tariffs has been presented. This was followed by a proposal of 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. Finally, this framework has been extended by the new orthogonal “Delay” 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 [16] and a flow-based auction turns out to serve as two prominent examples.

Future work with respect to Internet pricing schemes has to focus on implementation aspects of CPS. Moreover, there are important non-technical aspects to be covered as well. It has to be investigated how users react to this type of pricing 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 next steps will consist of gaining and deepening practical experience with this new Internet pricing scheme.

Acknowledgments

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, Grant No. 99.0536, and has been partially funded in the framework of the Austrian Kplus Competence Center Program.

References

- [1] D. Clark, W. Fang: *Explicit Allocation of Best-Effort Packet Delivery Service*; IEEE/ACM Transaction on Networking, Vol. 6, No. 4, August 1998, pp 362-373.
- [2] C. Courcoubetis, M. Dramitinos, G. Stamoulis: *An Auction Mechanism for Bandwidth Allocation Over Paths*. Submitted to ITC-17, International Teletraffic Congress, Salvador da Bahia, Brazil, Sept. 2001
- [3] 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.
- [4] R. Edell, P. P. Varaiya: *Providing Internet Access: What we learn from the INDEX trial*; Keynote Talk Infocom'99 New York, INDEX Project Report #99-010W. URL: <http://www.index.berkeley.edu/99-010W>.
- [5] M. Karsten (ed.): *Pricing Mechanism Design (PM)*. M3I Deliverable 3, version 1.0, June 30, 2000.
- [6] J. MacKie-Mason, H. Varian: *Pricing Congestible Network Resources*; IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, 1995, pp 1141 – 1149.
- [7] J. MacKie-Mason, H. Varian: *Pricing the Internet*; In: Public Access to the Internet, B. Kahn, J. Keller (eds.), Prentice Hall, Englewood Cliffs, New Jersey, U.S.A., 1995.
- [8] J. MacKie-Mason: *A Smart Market for Resource Reservation in a Multiple Quality-of-Service Information Network*. University of Michigan, Sept. 1997.
- [9] N. McIntosh: *Heavy Surfers Pay the Price*. The Guardian, England, U.K., August 3, 2000.
- [10] A. J. Mund: *AOL's Breakdown: A Harbinger for the Internet's Future?* Cable TV and New Media - Law and Finance, Vol. 14, No. 12, February 1997.
- [11] M3I: *Market Managed Multi-service Internet*; 5th Framework EU Project, IST Program, No. IST-1999-11429, URL: <http://www.m3i.org>, January 2001.
- [12] D. N. Newbery: *Privatization, Restructuring, and Regulation of Network Utilities*; The MIT Press; Cambridge, Massachusetts, U.S.A., 1999.
- [13] A. Odlyzko: *Paris Metro Pricing: The Minimalist Differentiated Services Solution*; 7th International Workshop on QoS (IWQoS'99), London, U.K., June 1-4, 1999, pp 159-161.
- [14] P. Reichl, G. Fankhauser, B. Stiller: *Auction Models for Multiprovider Internet Connections*; Messung, Modellierung und Bewertung von Rechensystemen (MMB'99), Trier, Germany, September 21-22, 1999.
- [15] P. Reichl, B. Stiller: *Notes on Cumulus Pricing and Time-Scale Aspects of Internet Tariff Design*. Technical Report No. 97, TIK, ETH Zürich, November 2000.
- [16] P. Reichl, P. Flury, J. Gerke, B. Stiller: *How to Overcome the Feasibility Problem for Tariffing Internet Services: The Cumulus Pricing Scheme*. International Conference on Communications (ICC 2001), Helsinki, Finland, June 11-15, 2001.
- [17] 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.
- [18] 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.
- [19] 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, TIK, ETH Zürich, September 2000.
- [20] B. Stiller, J. Gerke, P. Reichl, P. Flury: *Management of Differentiated Services Usage by the Cumulus Pricing Scheme and a Generic Internet Charging System*. 7th IEEE/IFIP Integrated Network Management Symposium, Seattle, Washington, U.S.A., May 14-18, 2001.
- [21] B. Stiller, P. Reichl, S. Leinen: *Pricing and Cost Recovery for Internet Services: Practical Review, Classification, and application of Relevant Models*; to appear: Netnomics - Economic Research and Electronic Networking, Vol. 3, No. 1, March 2001.