Prepaid Charging for QoS-enabled IP Services based on Time Intervals

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Pascal Kurtansky, Burkhard Stiller

Prepaid Charging for QoS-enabled IP Services based on Time Intervals

TIK-Report
Nr. 222, June 2005

Institut für Technische Informatik und Kommunikationsnetze
Computer Engineering and Networks Laboratory

Pascal Kurtansky, Burkhard Stiller
Prepaid Charging for QoS-enabled IP Services based on Time Intervals
June 2005
Version 1
TIK-Report Nr. 222

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

Institut für Technische Informatik und Kommunikationsnetze,
Eidgenössische Technische Hochschule Zürich

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

Prepaid Charging for QoS-enabled IP Services based on Time Intervals

Pascal Kurtansky[1], Burkhard Stiller [2, 1]
[1] Computer Engineering and Networks Laboratory TIK, ETH Zürich, Switzerland
[2] Department of Informatics, University of Zürich, Switzerland
E-mail: [kurtansky|stiller]@tik.ee.ethz.ch

***Abstract*—Based on the analysis of existing prepaid charging solutions with their limitations, a novel approach prepaid charging for Internet Protocol (IP)-based services based on time intervals is introduced. This approach combines key advantages of hot billing with a tariff flexibility and the guarantee of minimal fraud for providers. The technical effort of real-time charging is reduced to a minimum and, therefore, ensures economic efficiency as well as scalability in large-scale environments for providers who achieve a better customer binding even in prepaid cases. Driven by the theoretical background being developed an application scenario shows the practicability of this approach of prepaid charging based on time intervals.**

***Index Terms*—Prepaid Charging, Zero Fraud, QoS-enabled, IP Services**

I. Background

During many years, the role of charging for telecommunication services has been studied thoroughly [39, 15], where charging is defined as a process that is applied by service providers and network operators on the usage of services. In contrast to telecommunication systems and their standards, charging in IP networks is very often referred to as pricing and the two terms are used interchangeably. An overview of pricing (i.e. charging) in IP networks is given in [26, 24].

A noticeable development is that the two worlds of telecommunication systems and IP networks will merge together into a so-called all-IP network [20, 35]. From a charging perspective, this merging means to combine the most sophisticated features and functionalities from either system. This paper takes one widely-used feature found in mobile telecommunication systems—i.e. prepaid charging—and shows how it can be applied for QoS-enabled all-IP networks.

To distinguish charging from other processes connected with it, Karsten et al. [29] define a layered model, where charging is positioned on top of metering and accounting, but below billing. Metering determines the particular usage of resources on the network and application layer. Accounting defines the summarized technical usage information in relation to a users' service consumption. Charging maps accounting data into monetary units by applying a tariff, i.e. evaluating a tariff function. Billing defines the collection of charging records and the delivery of a bill to the user.

Orthogonal to this layered model different charging options, namely prepaid and postpaid, have to be distinguished. However these two options are named differently in mobile telecommunication standards. As defined in [10], the former is referred to as online charging and the latter as offline charging. Within postpaid approaches, charges for service consumptions are added to a bill, which is delivered to the user at the end of a billing period, *e.g.*, after one month. Within the prepaid approach, the customer is buying a certain amount of credits—typically represented in monetary units—prior the service usage. These credits are kept at the providers side and are referred to as the customers' balance, Lin et al. [32].

If a customer is requesting a prepaid service, then an initial amount of credits is deducted from the users' balance, given that sufficient credits are left. This can be done either directly or in a reservation-based manner. In the former case, granting of service usage also includes the deduction of the corresponding credits. In the latter case, credits are reserved prior to service usage and if they are not used up entirely, restored to the users' balance after the service usage. Depending on the tariff and amount of credits deducted, the customer can use the whole or only a part of the service. Thus, this makes it necessary to re-check and re-request credits during service consumption in real-time to prevent overuse of credits. The real-time characteristics of this process are the reason why telecommunication standard bodies are referring to prepaid as online charging. Performing this online charging for a large group of customers and a big variety of services is very resource consuming for a provider as shown in [22].

To reduce the costs of real-time charging, a hybrid approach—i.e. between pre- and postpaid—the so-called *hot billing* was introduced. As described by Lin et al. [32] in hot billing, during service consumption the charging is not done in real-time, instead it is performed at a certain point in time—*e.g.*, once within 24 hours—or after the service usage. In the period between successive points in time of credit checking, the customer may overuse his credits, i.e. the balance may go below zero. In hot billing environments this time period is called fraud window. Due to this post-processing of credit checking, some operators reported to have revenue losses up to 20%.

To combine the advantage of hot billing, i.e. avoiding the real-time charging effort, with a fast information on the charges to be applied, the concept developed within this paper defines a reservation-based prepaid scheme. The technical overhead imposed due to the real-time charging is minimized,

while the risk of fraud is eliminated, i.e. a zero fraud window is guaranteed.

The remainder of the paper is organized as follows. While Section II shows existing solutions and their limitations, Section III lists all requirements that need to be met. Section IV is dedicated to services supported and Section V develops the mathematical background of time intervals. Driven by the application of time intervals within Section VI, a real-world scenario is outlined in Section VII. Finally, Section VIII draws conclusions.

II. Existing Solutions

In mobile telecommunication networks prepaid systems are well established and they can be categorized into four groups [32, 17]: (1) Service Node, (2) Intelligent Network (IN), (3) Handset-Based (SIM card-based), and (4) the concept of Hot Billing. These prepaid systems are well suited to handle services being offered by 2G and 3G networks, such as voice or messaging services. From the standardization perspective, there are two important projects dealing with 3G networks—3GPP [1] and its sister project 3GPP2 [2]. Both projects are based on the ITU IMT-2000 standards [14], whereas 3GPP focuses on the european standards—IMT-DS/-TC (UMTS)—and 3GPP2 on the american and asian standard—IMT-MC (CDMA2000). The 3G systems in place today, UMTS [9] and CDMA2000 [4, 3] are able to handle classical circuit-switched services such as voice calls and basic packet-switched services based on IP. The evolution of todays 3G networks is towards an all-IP architecture [8, 6], which will be in place within the next years. Therefore, concerning the future of prepaid charging in 3G networks, the support for IP services will become very important.

An approach to cover services from mobile telecommunication and IP networks is done by the non-profit consortium Parlay/OSA [38]. Parlay/OSA defines an API providing all functionalities needed from defining a service, handling authorization and authentication, through support of mobility, session control and finally charging [11]. Parlay/OSA acts like an additional layer below the application layer hiding all different kind of networks and giving the impression of an "all-Parlay" architecture. This is surely a reasonable way to go to today, but it's more likely that the future brings all-IP instead of all-Parlay.

On the other hand, for IP networks, many existing solutions today are proprietary and for very specific services, *e.g.*, some VoIP service providers are selling prepaid VoIP cards. Non-commercial standards for prepaid charging in IP-networks are being developed by the two IETF working groups, AAA [12] and RADEXT [13]. The former extends the diameter base protocol [18] with the so-called *Credit-Control Application* (CCA) [27]. The extensions made by the latter [33] focus on prepaid charging only, while CCA can also be used for postpaid charging. Besides differences in the protocol messages, the two extensions handle prepaid charging quite similarly and interoperability between them is provided, too.

These extensions define new network elements and interactions between them to handle prepaid charging. A thorough analysis shown in Section IV will reveal the limitations of those approaches.

Today, prepaid charging for services is far more common in mobile telecommunication networks than in IP networks. An analysis of available services in todays 3G networks shows, that prepaid customers have limited service offerings to choose from compared to postpaid customers. This is mainly due to the fact that (1) older prepaid systems have been designed for voice and messaging services only and (2) extending existing systems to support data sessions is very complex and expensive. Even for this limited set of wireless services, Lilge [31] shows that prepaid charging has become very popular in the last years and the increasing trend still continues. From a commercial and economic perspective, there exist many white papers by major companies—like [16, 36, 25]—stating that future networks must be able to support prepaid for a variety of data and content-based services with much more flexible tariffs. Thus, there exists a clear requirement to be able to support prepaid charging not only for pure voice and messaging services, but also for advanced IP services.

Except for hot billing, existing reservation-based prepaid systems are technically based on so-called *alarms*. Alarms represent thresholds of accounting data a customer is allowed to use up, *e.g.*, number of seconds for a duration of a voice call or an amount of volume allowed for a download. Alarm-based prepaid systems work as follows: First, upon the request for a service, credits in monetary units are being checked and a certain amount of these credits is reserved. Second, thresholds (i.e alarms) have to be calculated. Calculating thresholds is conceptually achieved by reversely applying tariff functions. For a given amount of credits, measured here in monetary units, the charging system has to find the corresponding amount of technical accounting data—i.e. the charge for these accounting data equals exactly those credits. Third, during service consumption thresholds are monitored in real-time by the accounting system. Forth, if a threshold is reached, an alarm is generated. Fifth, the alarm is communicated to the charging system, which has to do the charging for these resources consumed. If there is a number of sufficient credits left, a new threshold is calculated and used for further processing. Otherwise, the service will be stopped.

The service node approach is one of the most widely deployed prepaid solution. One important difference to the process described above is that the service node approach uses a fixed and a priori defined threshold for the alarm. The alarm itself is represented in monetary units and is independent of the service tariff, *e.g.*, the cost per minute of a voice call. Therefore, it's possible that a customer is overusing his credits and that the operator loses revenue. To reduce the possible bad debt, an operator can decrease the threshold of the alarm but at the expense of an increased load of the service node, because credit checks have to done more often. This dilemma is analyzed thoroughly by Chang et al. [22]. For the voice call service, they show how to set an optimal credit checking frequency, i.e. the equilibrium between costs for credit checking and the amount of revenue loss.

A similar analysis has been performed by the same authors for prepaid systems based on hot billing, [21]. In hot billing, the charging data records (CDR) are always processed after the service usage, *e.g*, after the call has been released. Due to the signalling costs that occur when processing CDR's, an operator should try to reduce the number of CDR's, i.e. to increase the number of completed calls accumulated in one CDR. But reducing the CDR transmission frequency increases the potential bad debt. As for the service node approach, Chang et al. provide an optimal CDR transmission frequency for the voice call service.

Therefore, prepaid systems can be optimized for a single service, i.e. the voice service. But one of the major requirement imposed on future prepaid systems, is that they are able to cope with a multi-service environment. Towards this direction goes the analysis of miscellaneous data services in 3G networks by Khan et al. [28], like web browsing or e-mail. According to scenarios with different customer types and their traffic characteristics, the authors show how to set an optimal CDR transmission frequency. Unfortunately, the authors don't present a formal solution to the problem.

The requirements of reducing the amount events produced to check credits and the need to minimize the possible revenue loss for an operator, are diametrically opposed. Basically, two core problems are connected with this issue. First, the proper mapping or translation of the credits represented in monetary units into thresholds that can be used to set alarms. Then, secondly, a way has to be found to minimize the number of message exchanges needed to control these alarms in real-time. The first problem increases in a multi-service environment, where services can be used in parallel. It then becomes necessary to apportion the prepaid credits among the services, but avoiding credit fragmentation. The first problem also increases, when more complex non-linear tariffs functions depending on several parameters are used, e.g., a combined abc-tariff as introduced by Kelly [30]. However, typical today's tariffs are much simpler in the sense that they only depend on one parameter, *e.g.*, the volume downloaded. Since providers are in search of more flexible tariffs that can be applied for various IP services in the prepaid charging option, new schemes are essential. Solving the second problem, must also involve the type of service and the characteristics of tariff functions as analyses above have shown.

The newly developed prepaid charging approach based on time intervals, solves the two problems elegantly. A novel tariff modelling allows for nearly arbitrary complex tariff functions and to support multi-service environments, the concept of service bundles is introduced. For each service bundle the calculated time interval allows to reduce the number of thresholds to be monitored in real-time.

III. Requirements

As this work on prepaid charging is targeted at IP-based services, the underlying All-IP assumptions are outlined and the key functionality of accounting is defined to be applied in the network.

A. Architecture: All-IP

The core requirement is that an *all-IP* network architecture supporting QoS is in place, for instance [8, 6]. There are technical and commercial indications that future network architectures will overcome the heterogeneity of today's communication networks and that they will merge to one single communication technology, i.e. IP. The concept of all-IP defines separate IP multimedia domains interconnected with and operating themselves on IP.

In Europe, the today's 3G architecture is based on UMTS (Universal Mobile Telecommunications System) release 99 [9], which contains two logical transmission planes—circuit- and packet-switched domains. To extend capabilities for supporting packet-switched services with QoS (Quality-of-Service) and AAA provisioning, a so-called *IP Multimedia Subsystem* (IMS) [19] was introduced in latest UMTS release [8]. This extension is often being referred to as "the 3G all-IP architecture", as for instance in [20, 35]. A similar development is taking place for the CDMA2000 standards, which are very widespread outside Europe. Currently, CDMA2000 1xRTT systems represent a 2.5G solution [4], which together with the extension CDMA2000 1xEV-DO Release 0 [3] constitute the 3G counterpart to UMTS. The upcoming release A of CDMA2000 1xEV-DO [5] introduces end-to-end QoS for packet-switched services and can be seen as a first instance of an 3G all-IP architecture.

An all-IP architecture has also been developed within the Daidalos project [23]. The IMS architecture and the Daidalos all-IP architecture are quite similar, cf. Daidalos deliverables [23]. A simplified Daidalos architecture with respect to charging is depicted in Fig. 1. with its core components listed in Table I.

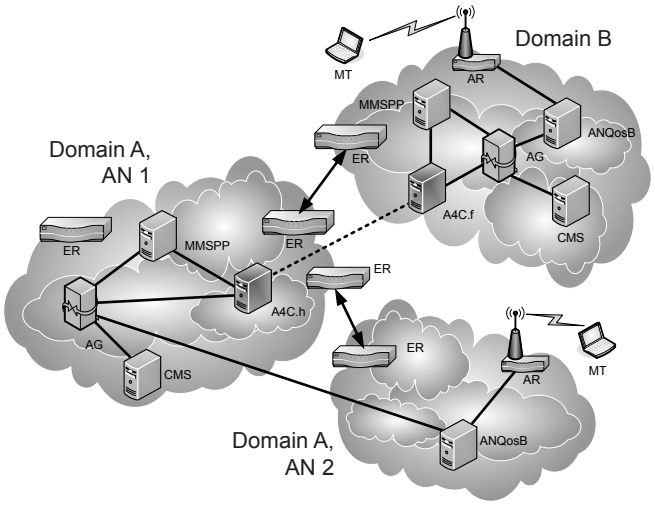

Fig. 1. Example of an all-IP Architecture.

Fig. 1 shows a scenario with two providers—A and B—and their administrative domains with corresponding access networks (AN). Between those two providers a business-to-business relation exists so that roaming between them is possible From a providers' A perspective, the domain A is the home

and domain B the foreign domain for its customers. Therefore, the A4C systems (Authorization, Authentication, Accounting, Auditing, and Charging) are named accordingly, i.e. *A4C.h* and *A4C.f*.

TABLE I: CHARGING RELEVANT CORE COMPONENTS

	Component	Main Functions
A4C	Authorization, Authentication, Accounting, Auditing, and Charging	Core component for A4C functions, for prepaid charging: Credit Control Server.
AG	Accounting Gateway	Correlation of accounting events, for prepaid charging: Credit Control Client.
AN QoSB	Access Network QoS-Broker	Ensuring QoS in the access network.
AR	Access Router	Physical connectivity in the access network based on: Ethernet, WLAN, W-CDMA, DVB-H
CMS	Central Monitoring System	Layer 3 metering.
ER	Edge Router	Interconnecting administrative domains (IPv6)
MMSPP	Multi-Media-Service-Provisioning-Platform	Handling of multimedia services.
MT	Mobile Terminal	User Equipment.

A4C systems are based on the IETF AAA Architecture, [12] with extensions to support auditing and charging. To support the prepaid charging option, the diameter credit control application is used [27], whereas the Accounting Gateway is acting as the client and A4C as the server. During any service consumption, the Central Monitoring System is constantly collecting metering information and is forwarding it to the AG. The process of charging, i.e. the mapping of accounting data into monetary units is done within the A4C.

All services offered over this all-IP architecture are implicitly based on QoS. There exists a well-defined mapping of services into distinct traffic classes and technical QoS parameters, *e.g*., bandwidth specifications, priorities or Differentiated Services Code Point (DSCP) values. Network resources needed—wired or wireless—to provide services are always controlled and reserved by the ANQoSB. As a consequence, there exists no pure best-effort traffic class without any reservations of resources. Best-effort traffic is treated as a service with low QoS, i.e. very small bandwidth and low priority.

B. Accounting

The key requirements on service accounting are as follows: (1) on the transport level, accounting must provide detailed information on all QoS parameters of the traffic class used by the service, (2) on the application layer, high level service usage data must be accounted for, *e.g.*, the number of messages sent, (3) the ability to support inter- and intra-domain mobility is demanded, and (4) the correlation of accounting data and events per session and customer has to be performed. The Daidalos' accounting system and the charging system for the IMS allow the correlation of accounting data or CDRs.

Therefore, the remainder of this paper assumes that this all-IP architecture including such a type of accounting system meeting all these requirements are in place.

IV. IP SERVICES AND TARIFF MODELING

Based on the definition of IP services the notion of service bundles is introduced. Those bundles as well as the model of tariffs defined serves as the important basis for the time interval prepaid approach.

A. IP Services

Within the all-IP architecture any kind of services based on IP are supported. Service types reach from network up to application services. The characteristics of services, the tariffs to be applied and QoS-parameters are kept within the Service Level Agreement (SLA). To identify a service unambiguously, each service is associated with a unique service ID, denoted as s_i. Typical network services consist basically of *IP with QoS*. They are used as the transportation basis for application services, *e.g.*, a video call service. Both, network and application services need to be identified in a unique way. This identifier is not only necessary for authentication and authorization, but also for accounting and charging: It is the service identifier that is used as a unique key to associate accounting data with a service and to apply the corresponding tariff. On a per-user or per-customer segment basis, every service is associated with a tariff function, which maps accounting data onto monetary units. The provider charges for services by applying the tariff with the charging option specified in the customers' contract, normally bilaterally defined beforehand in SLA's.

Application services are further categorized into session- and event-based groups. The latter consists of services whose consumption is reflected by one-time events. Typical event-based services are sending messages, querying traffic information or buying a song. Many existing tariff schemes for event-based services charge per event only, i.e. the transport of data corresponding to the event is not charged additionally. Therefore, the metering and accounting system within the all-IP architecture must be able (1) to detect every event and (2) to correlate properly accounting data to those events.

Session-based services have well-defined session setup-, communication-, and session release-phases. Typical session-based services are VoIP-calls, video-/audio-streaming, or video-conferencing. Independent of the signalling protocol used to establish and release sessions, the accounting system in the all-IP architecture must be able to detect these events and generate accounting data accordingly.

B. Service Bundling

With an all-IP architecture in place a provider is able to offer a variety of IP-based services to its customers. For prepaid charging, such an multi-service environment introduces a new problem. The credits in monetary units represent a singu-

lar value, which has to be shared among all the services with prepaid charging option. If a customer is using *n* of these services in parallel, then up to *n* distributed processes are accessing and modifying the credits. Apart from the problem of mutual exclusion, the real-time checking and updating of credits has to be done for all *n* services independently. Since providers have typically thousands of customers, the load imposed on the prepaid system increases dramatically. This effect is intensified, because the prepaid system is part of the A4C, which is a central component.

To overcome this scalability problem for such a multi-service environment, so-called *service bundles* are introduced. A service bundle consists of *I* services $s_1, s_2, ..., s_I$ with prepaid charging option. The customer can use any subset of all *I* services from the bundle in parallel. Instead of performing credit checking of the services independently, it will be done on a per bundle basis. To be able to do so, a so-called *time interval* is associated with every bundle (cf. Section V). To calculate the time interval for each bundle, the tariff functions of each service s_i have first to be transformed into time-based functions. In the second step, the time interval is being calculated and applied on the basis of the remaining credits of the customer (cf. Section VI). The key benefit of this concept is that credit checking and updating needs only to be done on the basis of this time interval instead of independently for each service from the bundle.

This service bundle concept is comparable with the credit- and resource-pools from diameter CCA [27] and the radius extensions for prepaid [33] respectively. However, the basis for these approaches are very simple linear tariff functions depending on one variable. For this kind of functions, the rule of proportion can always easily be applied to calculate so-called multipliers M_i for every service *i* belonging to the pool. The used resources per service *i* are labelled C_i, and thus the service units used per pool can be calculated by summing up all products M_i*C_i. The total service units *S* in the pool are calculated according to the initial quotas Q_i associated to every service $_i$, i.e. summing up over all products M_i*Q_i. The main concern about these approaches is that they don't solve the problem of sharing credits smoothly among services. This becomes evident by a deeper inspection of flow IX in [27]: From initially *0* units, *S* is increased by M_i*Q_i, everytime a new service *i* from the pool is requested by the customer. The Q_i are calculated by making reservations of *some* $ from the customers balance, *e.g.*, $5 for service 1 yielding to $Q_1 = 50$ minutes. But this is the actual problem to be solved, i.e. how much is *some* $ for every service *i*? Sure, the services from the pool are then sharing the *S* units among them, but the point is how monetary units are mapped smoothly into *S*. Neither diameter CCA, nor the radius extensions for prepaid specify this.

C. Tariff Modeling

A tariff function is associated with each service s_i, labeled $f_i(.)$. Depending on the provider's policy, these functions can be the same per market segment, or individually defined per customer. The tariff function for each service s_i depends on one or more variables. To allow a correct formal modelling, *m* denotes the total number of different variables per bundle for all services. If these *m* variables are arranged in a vector, tariff functions for each service can be defined in the form:

$$f_i(\underline{x}) \qquad \underline{x} \in \mathbb{R}^m \qquad f_i : \mathbb{R}^m \curvearrowright \mathbb{R} \tag{1}$$

The input to Eqn. (1) is accounting data representing the amount of resources that have been consumed by using service s_i and the output is the charge to be paid. Input variables stand for any accountable unit—reaching from network up to application information. Typical network-related accounting information includes volume, flow identifiers, QoS, or number of packets. Accounting data from applications may include duration of sessions (*e.g.*, a VoIP call), number of messages sent, user context related data and so forth.

With the definition of tariff functions shown in Eqn. (1), providers have much more flexibility to define user incentive tariffs than with the simple tariff schemes foreseen by diameter CCA and the radius extensions for prepaid. At least, the latter [33] dedicates its chapter 3.4 to the "support for complex rating functions". However, the chapter lacks a formal definition of tariff functions and therefore, it's not clear how they will be supported.

Mapping accounting data into monetary units is done by the tariff function $f_i(.)$ with the following mathematical properties: (1) $f_i(.)$ is represented by a smooth or piecewise smooth functions and (2) $f_i(.)$ is monotonically increasing. These properties are needed for the calculation of time intervals, cf. Section V. These few requirements on $f_i(.)$ enable the provider any freedom in defining tariffs meeting its own and customers' demands.

Whenever possible, accounting data representing resource consumptions should be related unambiguously to a service. This separation of accounting data is also the basis for a clear cost modeling so that the portion of direct costs can be maximized. Sharing accounting data between different services can yield to a multitude of problems. First, user incentive pricing may not be possible. Second, it leads to a significant portion of general costs, which cannot be properly allocated to the cost units.

The conclusion from this unambiguously relating accounting data to services, is that all services in the bundle do not share input variables of their $f_i(\underline{x})$ amongst them. Therefore, any two subsets out of the total *m* variables belonging to different services are disjoint and, hence, elements of $\underline{x}$ can be ordered in a way such that successive elements belong to the same service, cf. Fig. 2.

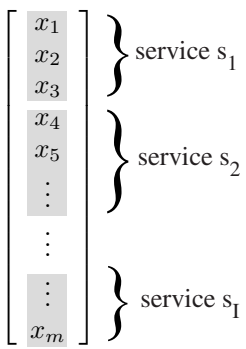

Fig. 2. Ordering and assigning elements of vector x to different services.

Every service bundle consists of *I* services $s_1, s_2, ..., s_I$. If a customer consumes all *I* services with prepaid charging option at the same time, then the charge to be paid, is defined by the tariff function $F(\underline{x})$, Eqn. (2):

$$F(\underline{x}) := \sum_{i=1}^{I} f_i(\underline{x}) \qquad F : \mathbb{R}^m \curvearrowright \mathbb{R} \tag{2}$$

Using all services at the same time for a given service bundle also means that Eqn. (2) represents the maximum charge to be applied from a providers perspective.

V. Time Intervals

The tariff functions derived so far depend on *m* variables. However, those tariff functions can be simplified in a way that the number of variables is reduced to only one, i.e. time.

A. Relation between Time Intervals and Service Bundles

As stated in Section II, alarm-based prepaid systems monitor in real-time resource consumptions for services used by the customer. For a given service bundle and a consumption of all *I* services at the same time, this would mean that an alarm-based prepaid system has to monitor *m* variables in real-time. To overcome all problems associated with that approach, the concept of so-called time interval on the basis of service bundles has been developed, yielding to (1) a drastically reduced number of variables needed to be monitored in real-time, (2) a simplified charging, accounting system, and (3) an increased scalability of the charging system.

The core concept is to calculate a so-called *time interval*, wherein the customer can use any combination of all *I* services without exceeding a given charge. The prepaid system has to monitor only this time interval and some view variables belonging to event-based services, which is much less than *m* variables.

B. Transformation into Time-based Tariff Functions

To each service s_i corresponds a tariff function $f_i(\underline{x})$, where $\underline{x}$ consists of *m* elements. For each of these tariff functions a transformation has to be found separately. With the partitioning of vector $\underline{x}$ as presented in Fig. 2, the subset of variables needed for a service s_i is bounded by *p* and *q*, cf. Eqn. (3).

All other variables of $\underline{x}$ outside the boundaries [*p*,*q*] are not

$$\begin{aligned} \underline{x} &= [x_1 \cdots x_p \cdots x_k \cdots x_q \cdots x_m]^T \\ &\forall\ k,p,q \in [1,m] :\ p \le k \le q \end{aligned} \tag{3}$$

needed by $f_i(\underline{x})$. They represent resource consumptions by other services than s_i and will be set to zero for the ongoing modeling. To do so, a helping vector $\underline{v}$ is defined, cf. Eqn. (4).

$$\begin{aligned} \underline{v} &= [0\ \cdots\ 0\ v_p\ \cdots\ v_k\ \cdots\ v_q\ 0\ \cdots\ 0]^T\ \ \underline{v} \in \mathbb{R}^m \\ &\forall k \in [p,q] :\ v_k = 1 \end{aligned} \tag{4}$$

With the definition of Eqn. (4) $\underline{x}$ can be transformed in a way that all other variables become zero that are not needed by $f_i(\underline{x})$, cf. Eqn. (5)

$$\begin{aligned} \underline{x} &= \underline{x}^T \underline{v} = \\ &[0\ \cdots\ 0\ x_p\ \cdots\ x_k\ \cdots\ x_q\ 0\ \cdots\ 0]^T \end{aligned} \tag{5}$$

Now, whenever the absolute of $\underline{x}$ increases, it can only be due to resource consumptions by service s_i.

Since $f_i(\underline{x})$ is a function depending on m variables, $f_i(\underline{x})$ is a curve on a surface in an *m*-dimensional space. Instead of looking at $f_i(\underline{x})$ in an *m*-dimensional space, one can study the behavior of $f_i(\underline{x})$ in the time domain: The time is divided into discrete time intervals, wherein the behavior of Eqn. (1) and Eqn. (2) is studied:

$$[t_j, t_{j+1}] \qquad \forall j \in \mathbb{N} : \quad t_j < t_{j+1} \tag{6}$$

The following values for $\underline{x}$ are assigned with the borders of the time interval:

$$t_j \rightsquigarrow \underline{x}_j, \quad t_{j+1} \rightsquigarrow \underline{x}_{j+1} \qquad \underline{x}_j, \underline{x}_{j+1} \in \mathbb{R}^m \tag{7}$$

Fig. 3 shows the behavior of a hypothetical tariff function $f_i(\underline{x})$ in the time domain, with the current time interval shown on the x axis.

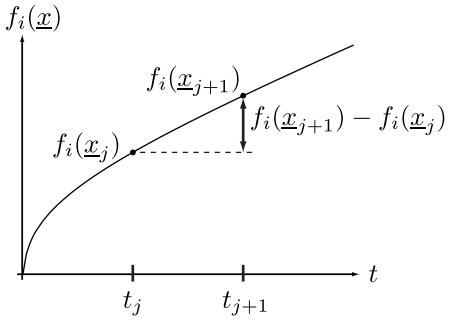

Fig. 3. Values of tariff function with respect to the time interval.

At the beginning of the interval $\underline{x}_j$ represents the amount of resources that have been consumed so far—i.e. up to the current time interval—by service s_i. If $\underline{x}_j > 0$, then $f_i(\underline{x}_j)$ is the charge, which has already been subtracted from the customers' credit. There are only two cases, where $\underline{x}_j = 0$: First, the

provider has decided to reset resource consumptions accumulated so far for service s_i, or second, the current time interval is the very first time that a customer uses service s_i. In these two cases also $f_i(\underline{0})$ is zero.

The charge to be paid at t_{j+1}—i.e. at the end of the time interval—is given by the delta between the values of $f_i(.)$ at the interval borders.

On one hand, if service s_i is not used within the current time interval, no resources are consumed and, hence, the charge to be paid at t_{j+1} is zero as well. On the other hand, Fig. 3 also implies that any usage of service s_i within the interval according to Eqn. (6) must lead to resource consumptions,

$$\exists \underline{\Delta} \in]\underline{0}, \underline{\Delta}_{max}] : \quad \underline{x}_{j+1} = \underline{x}_j + \underline{\Delta} \qquad (\underline{\Delta} \in \mathbb{R}^m) \tag{8}$$

The upper limit of $\underline{\Delta}$ is represented by $\underline{\Delta}_{max}$ which is the maximum of resources that could be consumed by service s_i in the interval Eqn. (6).

An important property of Eqn. (1) is that an increase in resource consumptions must always lead to an increase of $f_i(\underline{x})$—i.e. $f_i(\underline{x})$ is monotonically increasing, hence, it follows Eqn. (9).

$$f_i(\underline{x}_j) < f_i(\underline{x}_{j+1}) = f_i(\underline{x}_j + \underline{\Delta}) \tag{9}$$

The logical conclusion from Eqn. (8) and Eqn. (9) is that it exists no smaller $\underline{\Delta}$ than $\underline{\Delta}_{max}$ within the interval that could lead to a higher charge, cf. Eqn. (10):

$$f_i(\underline{x}_j + \underline{\Delta}_{max}) > f_i(\underline{x}_j + \underline{\Delta}) \qquad \forall \underline{\Delta} \in]\underline{0}, \underline{\Delta}_{max}[\tag{10}$$

The charge to be paid at t_{j+1}, i.e. at the end of the time interval, is given by Eqn. (11).

$$c_i(\underline{x}_j, \underline{x}_{j+1}) = f_i(\underline{x}_{j+1}) - f_i(\underline{x}_j) \quad c_i : \mathbb{R}^m \curvearrowright \mathbb{R} \tag{11}$$

Note that $c_i(.)$ can only be evaluated at the end of the time interval, i.e. when all input values are known. Before t_{j+1} is reached $\underline{x}_{j+1}$ is still unknown. But with properties derived so far, the upper bound of $\underline{x}_{j+1}$ is given by Eqn. (12).

$$\underline{x}_{j+1} \leq \underline{x}_j + \underline{\Delta}_{max} \tag{12}$$

The maximum charge to be paid for resource consumptions of the service s_i in the time interval Eqn. (6) is defined as $cmx_i(.)$, cf. Eqn. (13):

$$\begin{aligned} cmx_i(\underline{\Delta}_{max}) &= f_i\left(\underline{x}_j + \underline{\Delta}_{max}\right) - f_i(\underline{x}_j) \\ &\geq c_i(\underline{x}_j, \underline{x}_{j+1}) \\ & \qquad cmx_i : \mathbb{R}^m \curvearrowright \mathbb{R} \end{aligned} \tag{13}$$

Note that the variable $\underline{x}_j$ and the corresponding value of $f_i(.)$ is known from the last interval. The length of the time interval is given by Δt in Eqn. (14):

Now, two important analogies are needed to achieve the mapping from $f_i(.)$ into the time domain: First, between Δt and

$$\Delta t := t_{j+1} - t_j \tag{14}$$

$\underline{\Delta}_{max}$: $\underline{\Delta}_{max}$ represents the maximum of resources that a customer can consume within the time Δt. Second, t_j corresponds to the known value $\underline{x}_j$.

The value of $\underline{x}_{j+1}$ will not be known before t_{j+1} is reached. But, this is not true for the upper bound $\underline{x}_{j+1}$: The right hand side of Eqn. (12) and the left hand side of Eqn. (13) can be calculated, if $\underline{\Delta}_{max}$ can be represented by functions depending on t only.

To achieve this, every variable x_k is being substituted by functions depending on time only. Finding such time dependent functions is only possible if all requirements as listed under Section III are met.

The maximum amount of resources that can be consumed within the current time interval is independent of those resources that have been consumed in the last time interval. Therefore, t is only relative to borders of the current time interval as shown in Eqn. (15).

$$t \in [0, \Delta t] \tag{15}$$

Every x_k is now being substituted by functions $h_k(t)$, cf. Eqn. (16).

$$\forall k \in [p, q] \ : \ x_k = h_k(t) \tag{16}$$

Note that $h_k(t)$ models the maximum consumption of the resource x_k over within a time t for the given time interval in Eqn. (15). This process of substitutions is repeated for all x_k of $\underline{\Delta}_{max}$, which can be calculated now for a given t. To show clearly, the transformed $\underline{\Delta}_{max}$—depending on t only—is rewritten in the notation $\underline{\Delta}_{max}(t)$ as shwon in Eqn. (17).

$$\begin{aligned} \underline{\Delta}_{max}(t) := \ & \underline{\Delta}_{max} \begin{pmatrix} [0 \ \cdots \ 0 \ h_p(t) \ \cdots \\ \cdots \ h_k(t) \ \cdots \\ \cdots \ h_q(t) \ 0 \ \cdots \ 0]^T \end{pmatrix} \\ & \underline{\Delta}_{max}(t) : \mathbb{R}^m \curvearrowright \mathbb{R}^m \end{aligned} \tag{17}$$

Thus, the maximum charge for service s_i is represented by the function $cm_i(t)$ as within Eqn. (18).

$$\begin{aligned} cm_i(t) &= f_i\left(\underline{x}_j + \underline{\Delta}_{max}(t)\right) - f_i(\underline{x}_j) \\ &\geq c_i(\underline{x}_j, \underline{x}_{j+1}) \\ & \qquad cm_i : \mathbb{R} \curvearrowright \mathbb{R} \end{aligned} \tag{18}$$

With Eqn. (18) the maximum charge to be paid in the time interval of Eqn. (6) can be calculated with functions depending on t and on the preceding value of $f_i(\underline{x}_j)$ from the beginning of the interval at t_j. Comparing the left hand side of Eqn. (18) with the left hand side of Eqn. (13) reveals that the maximum charge can now be calculated for a given duration of the time interval t:

$$cmx_i(\underline{\Delta}_{max}) \ \widehat{=} \ cm_i(t) \tag{19}$$

The process of substitutions can be repeated for all services s_i in the bundle. As for the service charge shown in Eqn. (18), the charge for all I services in the bundle $C(\underline{x}_j, \underline{x}_{j+1})$ is bound by $CM(t)$. Thus, the maximum charge to be paid for all I services in the bundle for an interval of length t is given by $CM(.)$ as shown in Eqn. (20).

$$\begin{aligned} CM(t) &= \sum_{i=1}^{I} cm_i(t) \qquad CM : \mathbb{R} \curvearrowright \mathbb{R} \\ &\geq C(\underline{x}_j, \underline{x}_{j+1}) = \sum_{i=1}^{I} c_i(\underline{x}_j, \underline{x}_{j+1}) \end{aligned} \tag{20}$$

As the sample application will show in detail below, the basic use of Eqn. (20) is as follows: For a given amount of c credits from customers' balance, the charging system will calculate a time interval t, such that Eqn. (20) is exactly equal to c. I.e., the left hand side of Eqn. (20) is given and the length of the time interval will be calculated. This time interval t is the shortest interval, wherein the customer can use up c credits, i.e., it exists no shorter t that could lead to this charge. If the provider calculates t in advance, it needs to monitor this time interval t—i.e. until t has not yet expired no re-check of the customer's credit is required. Another important fact is that during the calculated time interval t, it is impossible for the provider to lose money, due to overuse of credits by the customer, thus the economic efficiency is achieved.

VI. Applying Time Intervals

Based on the formal definition of time intervals, a gradual calculation has to be performed, to be applied within the charging system.

A. Lower Bound For Time Intervals

Metering, accounting, and charging systems are typically physically distributed within the network of a provider. The accounting system is responsible for monitoring time intervals and event-based services. When the end of a time interval is reached, the accounting system contacts the charging system to calculate a new time interval. Within this time, the user may continue using the services, i.e. additional resources will be consumed during that time. Thus, the maximal time needed to do the check is referred to as t_c. It consists mainly of the Round trip Time (RTT) between the charging and accounting systems, the time that is spent on the algorithm shown below, and of some headroom time for possible error handling. To allow an error free operation, t_c must be known by the accounting system, too. The accounting system will itself monitor t_c and, if it expires without having received an answer from the charging system, it takes necessary actions. It is left to the provider's policy how to handle such cases. Besides t_c, accounting systems have a minimum length of intervals that they are able support—this is referred to as t_{min}.

B. Algorithm for Calculating Time Intervals

Whenever a service s_i with a prepaid charging option is requested from the bundle, a time interval needs to be calculated. The Time Interval Calculation Algorithm (TICA) is a straightforward algorithm defined as follows:

```
   Preconditions: t_c, t_min known.

1  TICA(p: Real, x_u: Resources): Real
   {
2     x_j := DB.read("x_j");
3     charge(x_u, x_j);
4     c := DB.read("c");
5     if c = 0 then return 0 else
      {
6        c' := p*c;
7        t := 0;
         repeat
            t:= t + ε
8        until cm_i(t,x_j) < c';
9        if (t-t_c) < t_min then return 0 else
         return (t-t_c)
      }
   }
```

Initially, t_{min} and t_c have to be known. The first input to `TICA(…)` is the percentage of credits that should be used for the calculation of the time interval (line 1). If $p = 1$, it is impossible for the customer to use other services outside the current bundle with a prepaid charging option, since 100% of all credits have been already reserved. The second input to `TICA(…)` are resources that have been consumed in the last interval, but which have not yet been charged for. If the current interval is the first, $\underline{x}_u$ is $\underline{0}$. The return value of `TICA(…)` determines the calculated time interval.

Line 2 reads the old value ($\underline{x}_j$) of resources that have been consumed up to the end of the last interval. This value has already been used for the charge calculation in the previous step. But $\underline{x}_j$ is needed to evaluate the tariff function with the correct initial value in this step.

Before a new time interval can be calculated, the charge for the real usage from the last interval has to be deducted from remaining credits (line 3). Therefore, `charge(…)` is called, which evaluates the service tariff function and stores credits updated in the database. In the next step (line 4) credits are read out from the database. If all credits have been used up, the new time interval is 0 (line 5). Otherwise, a percentage of those credits c is allocated to c' (line 6).

Finding the smallest t yielding to the charge c' begins n line 7 with setting t initially to 0. Then, t is increased by a small ε until $cm_i(t, \underline{x}_j)$ equals c' (line 8). The proper size of the increment ε should be chosen with respect to the numerical precision of data types applied. $cm_i(t, \underline{x}_j)$ represents the numerical evaluation of Eqn. (20) for a given $\underline{x}_j$.

The guard of line 9 checks wether t is sufficiently long, if not, 0 will be returned. Otherwise, for save operations, t is reduced by t_c before being returned. However, it could be possible that the returned value ($t - t_c$) is very close to t_{min}. In such a scenario, it is left to the provider, which type of actions to be taken, *e.g.*, informing the customer to recharge his credits.

C. Message Flows

The key components involved in prepaid charging based on

time intervals, are the charging and the accounting system. The latter is represented by the AG, whereas the former is one component in the A4C servers, cf. Fig. 1. Concerning interactions needed for prepaid charging, the charging system plays the servers' role, whereas the AG is the client. The AG is also the first part of a Policy Enforcement Point (PEP). If a "Stop" message arrives from the charging system, some or all services are removed from the list of allowed services for this customer. Depending on the type of service, this updated list is then communicated from the AG to the ANQoS-Brokers or to the MMSPP, cf. Fig. 1. These two systems constitute the second part of the PEP, i.e. they are responsible for actually disconnecting the customer.

Thus, Fig. 4 shows the basic operation and messages exchanged between the two core components with respect to prepaid charging.

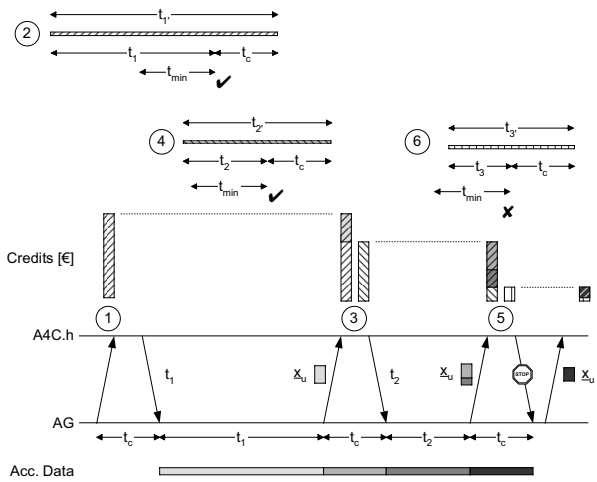

Fig. 4. Calculation of consecutive time intervals.

As shown in the TICA algorithm, a precondition is that t_c and t_{min} are known. The message sequence chart starts when a user is requesting a prepaid service. The AG contacts the A4C.h to check the user's credits and to ask for a first time interval, cf. ① in Fig. 4. User credits are shown as a hatched rectangle. The charging system calculates t_1', which corresponds to t in the algorithm shown before. Then, t_c is subtracted and the result is checked against t_{min}, cf. step ✍ in Fig. 4. In this case, t_1 is greater then t_{min} and, hence, the calculated interval is valid and is communicated to the AG. At this point, the customer can start using services belonging to the bundle. Within the time of t, accounting data is being accumulated in the AG. After t_1 is reached, the accounting data is forwarded to the charging system—shown as $\underline{x}_u$—combined with the request for a new time interval, cf ③

During the time needed to perform the credit checking in the charging system, customers may continue using services. Within t_c more accounting data will, thus, be accumulated at the AG. But this data can be forwarded only to the A4C.h at the end of the next interval.

At step ③, the charging system will do the rating for the actual resource consumption within t_1, i.e. the used resources $\underline{x}_u$. The resulting charge is subtracted from the initial credits, represented as a rectangle with a light-gray background. Remaining credits are used for the calculation of the second time interval, t_2. As shown at step ④, $t_2 > t_{min}$, and, therefore, it can be communicated to the AG.

After t_2 is reached, the AG will again ask for a renewal of the time interval. Combined with this request it sends the accounting data accumulated within t_2 and the preceding t_c—shown with different levels of gray, cf. step ⑤. Again, within the time needed to check credits, the user can continue using services. The calculation of the third interval leads to $t_3 < t_{min}$, cf. ⑥. In this scenario, the charging system sends a "Stop" message to the AG. Finally, the accumulated accounting data from the last t_c is sent to charging system, shown as a small rectangle with black background.

D. A Remark on Alarm-based Prepaid Systems

In purely alarm-based prepaid systems, the charging component would have to calculate a minimal set of $\underline{x}$, such that c credits, i.e. the left hand side of Eqn. (18) and Eqn. (20), can be used up. As already mentioned in Section II, it is not trivial to find thresholds for setting these alarms. One way to overcome these problems is to calculate a time interval as described before. This can either be done per service bundle, or just for the services being actually used in parallel with a prepaid charging option. After t is has been calculated, Eqn. (16) is applied inversely, i.e. for a the value t - t_{min}, all x_k are calculated by evaluating $h_k(.)$. Repeating that process for every x_k,will finally yield to a minimal set of $\underline{x}$. All values smaller than the calculated $\underline{x}$, would also yield to a smaller charge. Thus, alarms can be set for every x_k of $\underline{x}$. It has to be noted that all issues mentioned in Section II still remain valid for an alarm-based approach!

VII. Scenario And Example Application

To illustrate the developed concept of prepaid charging based on time intervals, a scenario-driven example is discussed in all relevant details.

A. Scenario Setting and Basics

The scenario is based on the all-IP architecture shown in Section III, where a hypothetical provider *A* is going to offer four services to its customers as listed in Table II. Provider *A* is maintaining a metering, accounting, and charging infrastructure that meets all requirements of Section III. Therefore, it is able to apply prepaid charging based on time intervals. The core algorithm TICA and tariff functions have been implemented in Maple [34].

TABLE II: Services Offered By Provider A

Service		DSCP		
ID	Description	Down-stream	Upstream	Use of Signaling DSCP?
s_1	Live Streaming TV	12	(5)	yes, 5

TABLE II: SERVICES OFFERED BY PROVIDER A

Service		DSCP		
ID	Description	Down-stream	Upstream	Use of Signaling DSCP?
s_2	Video Call	10	10	yes, 5
s_3	VoIP Call	9	9	yes, 5
s_4	Messaging	8	8	no
s_5	Signaling	5	5	—

After a careful marketing analysis of its customers, the provider A has decided to put services s_1 to s_5 into a bundle, to reduce the amount of real-time signalling.

All services from Table II use signaling. Some of them mark the signaling traffic with a special signaling DSCP. These services share a separate signaling channel, identified as service s_5. Other services simply mark signaling packets with the same DSCP as normal traffic, i.e. inband signaling. In this case, signaling traffic cannot be identified and is, therefore, charged for as normal traffic from that service. A special case is service s_1: It uses the upstream channel just to provide feedback on the quality of the received stream to the streaming server—hence there is no other upstream traffic than signaling.

TABLE III: QOS PARAMETERS

DSCP	QoS Parameters	
	Max. Bandwidth [kbps]	Relative Priority
5	16	1
9	32	1
10	64	1
12	256	2

Every DSCP value corresponds to a certain level of QoS, cf. Table III. In this scenario it is assumed that there are always sufficient resources in the network to deliver any combination of these services from that bundle.

After having defined the services with their QoS, providers needs to define tariff functions with their input variables. Table IV shows all 5 services and the signaling with their corresponding variables. It is assumed that metering and accounting systems are able to deliver all that information.

B. User-incentive Charging

With the prepaid charging option, the binding of customers to the provider A is not so tight as it is the case for a fixed postpaid contract, since in a multi-provider environment, customers can use services from other providers instead of provider A—but having the charges being subtracted from their credits that are actually kept at provider A.

TABLE IV: INPUT VARIABLES FOR TARIFF FUNCTIONS

Service	Tariff Parameters		
	Parame-ter Name	Description	Units
s_1	x_1	Volume with DSCP=12	KB
	x_2	Duration of TV Session	sec.
s_2	x_3	Duration of Video Call	sec.
	x_4	Number of Video Call Session Setups	—
s_3	x_5	Duration of VoIP Call	sec.
s_4	x_6	Number of Messages	—
s_5	x_7	Signaling Data, i.e. Volume with DSCP=5	KB

Therefore, the provider A has to apply a user-incentive charging, such that it is more beneficial for customers to stay at provider A than to switch to one if his competitors. A possible policy could consist of two parts.

First, provider A defines many of its tariffs in way that the charge per unit drops as the total amount of accumulated units increases. For postpaid contracts, many 2G operators have such tariff schemes in place today for GPRS (General Packet Radio Service) traffic, *e.g.*, sunrise [40] which has a piece-wise linear tariff function for the GPRS traffic volume.

Second, provider A introduces a so-called "bonus system" for some prepaid services: Charges for succeeding service sessions are relative to the amount of consumed resources from preceding sessions. Like in postpaid contracts, the charge per unit drops as the total amount of accumulated units increases. Typical today's "bonus systems" 2G operators are defined differently, cf. [40] and [37]. To prevent an infinite cumulating, a possible solution could be to clear consumed resources from preceding sessions when customers' credits have been under a certain amount of credits for more than, *e.g.*, 24 hours or when the end of the month is reached. This lower bound for the clearance of accumulated resource consumptions should not only be set due to marketing policies, moreover, technical considerations should be incorporated, too.

This bonus system determines a classical win-win situation: customers can benefit from lower tariffs and provider A has a way to bind customers even with prepaid charging.

C. Example Tariff Modeling

Service s_1 is a high quality streaming service whose tariff depends on the streamed volume and the session duration. Provider A has chosen a logarithmic tariff function for the former resource consumption and a piece-wise linear for the latter as shwon in Eqn. (21).

Note that the variable x_1 (volume) depends on x_2 (session duration), i.e., the longer the session the more volume will be streamed. The tariff is modeled in way that customers have incentives to have longer sessions than only very short ones.

$$f_1(\underline{x}) = \frac{1}{10}\log\left(x_1+1\right)^2 + \begin{cases} \frac{1}{480}x_2 & x_2 \in [0,600] \\ \frac{5}{8}+\frac{1}{960}x_2 & x_2 \in (600,\infty) \end{cases} \quad (21)$$

I.e., the charge per minutes drops after 10 minutes (600 seconds) and the charge for the volume follows a stretched logarithmic function.

Provider *A* wants to promote its new video call service s_2, therefore, he defines a tariff function making short sessions very cheap. But on the other hand, he wants to hinder people from making too short sessions, due to the overhead needed to setup a session. Therefore, he charges every session setup with a fixed rate of 20 cents. The complete tariff function for s_2 is defined in Eqn. (22).

$$f_2(\underline{x}) = \frac{1}{55000}x_3^2 + \frac{1}{5}x_4 \quad (22)$$

To compete with his competitors, provider A offers the classical VoIP service s_3 at a very cheap rate, where the charge per minute drops the longer the duration of the call lasts as defined in Eqn. (23).

$$f_3(\underline{x}) = \frac{1}{300}\log\left(x_5+1\right)^4 \quad (23)$$

Messaging is offered at a constant rate of 10 cents per message as shown in Eqn. (24).

$$f_4(\underline{x}) = \frac{1}{10}x_6 \quad (24)$$

Signaling traffic is charged at a very low and constant rate, 1 MB is available for only 5 cents as depicted in Eqn. (25).

$$f_5(\underline{x}) = \frac{1}{2048}x_7 \quad (25)$$

To prevent the misuse of signaling traffic or flooding attacks, all QoS brokers are configured in a way that customers cannot use this service for their own communications. It is assumed that rules and policies are properly defined and correctly initiated.

For every $f_i(.)$, the corresponding $c_i(.)$ have to be defined, $c_3(.)$ as shown in Eqn. (26).

$$c_3(\underline{x}_j, \underline{x}) = \frac{1}{300}\log\left(x_{(5)j}+x_5+1\right)^4 - f_3(x_{(5)j}) \quad (26)$$

Note that here $\underline{x}$ represents those resources consumed in this interval and that it is not relative to the old value $\underline{x}_j$ from the last interval. If there are no resources consumed within this interval, then $x_5 = 0$ and $c_3(.)$ equals 0, too.

Since customers are allowed to use any combination of services s_1 to s_4 including signalling (s_5), the maximum charge is given by Eqn. (27).

$$C(\underline{x}_j, \underline{x}) = \sum_{i=1}^{5} c_i(\underline{x}_j, \underline{x}) \quad (27)$$

D. Transformation of Tariff Functions

Since now all tariff functions have been defined properly, the next step is to transform Eqn. (27) into a function depending on time only. Note that an alarm-based prepaid system would have to monitor all those 7 variables in real-time under the condition that a customer would use all services in parallel. To reduce this large overhead, every input variable will be replaced by a function of time, *t*. It is assumed that *t* lies in the defined interval of Eqn. (6).

The first variable x_1 represents the volume, measured in kilo byte. Since it is known that the maximum bandwidth is 256 kbps (cf. Table III), the absolute maximum volume consumable can be calculated by the function $h_1(t)$, cf. Eqn. (28).

$$h_1(t) = \frac{256t}{8} = 32t \quad (28)$$

To be precise the above mentioned substitution should be rounded to integer numbers, since the smallest amount of volume is a byte. But the inaccuracy is at most one byte, for which the charge is negligible.

With Eqn. (28) the variable x_1 can be substituted with the content of Eqn. (29).

$$x_1 = 32t \quad (29)$$

For parameter x_2, the maximum charge would have to be paid, if a session of services s_1 starts at the beginning of the interval and would last until *t*. Similarly, considerations can be made for parameters x_3 and x_5, therefore, these three parameters can be substituted by *t*, as with Eqn. (30).

$$x_2, x_3, x_5 = t \quad (30)$$

Service s_2 uses a combined tariff of session duration and per session setup. The variable x_3 is substituted in similar way as x_2. From a theoretical point of view, a customer could have an unlimited number of sessions setups within the time interval. Since this is not realistic at all and very cumbersome to start a session every other second, it is sensible to limit the number of allowed session setups per time interval. The easiest way to achieve such a limitation is that provider A allows customers to setup a new session every 30 seconds on average. Thus, x_4 can be substituted in the following way as shown in Eqn. (31).

$$x_4 := \lfloor \frac{t}{30} \rfloor \leq \frac{t}{30} \quad (31)$$

Using the floor function is a valid mathematical way for this substitution. But it would introduce discontinuities in the resulting tariff function, which could lead to problems when

solving the equation for t afterwards. From a modeling point of view, the provider is interested in an upper bound for x_4, hence, the floor function can be omitted. Service s_2 depends on time only, so the transformation is similar x_2 from services s_1. For service s_4, provider A applies the same technique as for variable x_4 in Eqn. (31), but with an average interval of 15 seconds.

$$x_6 := \lfloor \frac{t}{15} \rfloor \leq \frac{t}{15} \tag{32}$$

Maximum charges for the signaling traffic can be calculated in a similar way as in Eqn. (28).

All variables x_k have been substituted by functions $h_k(t)$, thus, the corresponding $cm_i(t)$ can be defined and $cm_2(t)$ is shown in Eqn. (33).

$$\begin{aligned} cm_2(t) \quad = \quad & \frac{1}{55000}\left(x_{(3)j} + t + 1\right) + \frac{1}{5}\left(x_{(4)j} + \frac{t}{30}\right) \\ & - f_2(x_{(3)j}, x_{(4)j}) \end{aligned} \tag{33}$$

After all $cm_i(t)$ are defined, the maximum charge for using any combination of all services in the bundle, yields to $CM(t)$, cf. Eqn. (34).

$$CM(t) = \sum_{i=1}^{5} cm_i(t) \tag{34}$$

E. Example of Applying Time Intervals

As mentioned in Section VI, t_{min} and t_c have to be known before calculating exact time intervals. Based on the current implementation of the AG in the Daidalos project, t_{min} = 8 sec and t_c = 2 sec. Before using any services, the customer has to buy a certain amount of prepaid credits. With this initial value, the first time interval can be calculated. Thus, for a given amount credits, i.e. the left hand side of Eqn. (34), t is calculated such that it yields to this maximum charge. One of the main advantages of this approach is that intervals can be very long, cf. Fig. 5.

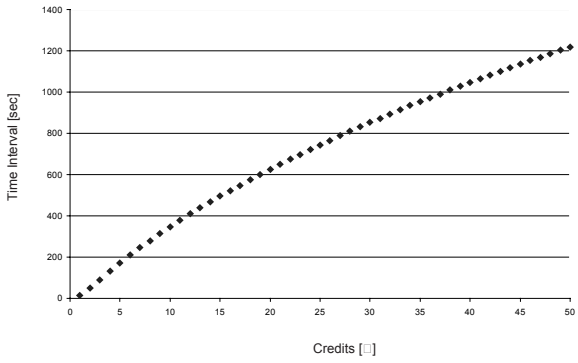

Fig. 5. Duration of initial time intervals for different credits.

The x axis shows initial credits that a user has bought, whereas the y axis shows the corresponding duration of the time interval. Note that within a time interval only very few parameters have to be monitored and, therefore, a huge amount of overhead and resources can be saved.

In the example scenario shown in Table V, the customer has bought a new prepaid credits with 20 €. For simplicity it is assumed that always 100% of these credits are used for the calculation of time intervals, i.e. p = 1 holds for the algorithm TICA. Rows in this table show the calculation for 6 consecutive time intervals. During a running interval, the AG has to monitor only time and parameters x_4 and x_6 belonging to event-based services. Thresholds for these parameters are calculated according to Eqn. (31) and Eqn. (32) for a given length t of the time interval. In this example scenario, it assumed that only the expiry of the time interval can trigger a request against the charging system for a new calculation. If thresholds for x_4 and x_6 are reached before the interval expires, the user will not be able to set up another video call session or to send another message.

Within a time interval, the customer may use an arbitrary combination of services, s_1 to s_4, whose resource consumptions are reflected in the vector $\underline{\Delta}$. The charges—column C in the Table V—are calculated on the basis of $\underline{\Delta}$ and $\underline{x}_j$, *i.e.*, resources used in preceding intervals. The used resources in Table IV are based on a realistic scenario

TABLE V: EXAMPLE APPLICATION OF TIME INTERVALS

Interval j =	Duration of Time Intervals (reduced by t_c), t =	Credits		Used Resources in Interval $\underline{\Delta}$ =
		Used for Calculation c' =	Consumed within Interval C =	
1	623	20.00	1.79	[1800, 60, 85, 1, 30, 2, 115]
2	560	18.21	8.11	[0, 0, 562, 0, 0, 3, 674.4]
3	210	10.10	0.59	[0, 0, 20, 0, 0, 1, 24]
4	195	9.51	1.41	[5580, 180, 0, 0, 120, 5, 90]
5	173	8.10	7.86	[5600, 175, 175, 5, 175, 11, 700]
6	2	0.24	—	—

With a starting credit of 20 €, the first time interval has a length of 623 seconds. Thus, within more than 10 minutes no additional signalling is required to re-check the customer's credit.

When the end of the first interval is reached, the charge C is calculated. In the example, the customer has spent 1.79 € which will be subtracted from the credit. Thus, 18.21 € can be used for the calculation of the next interval, resulting in 560 seconds possible service usage duration. Resource consumptions in Table V show an example behavior of a customer.

Of further interest is the fifth interval, where the customer uses all services in parallel. For t = 173 seconds, thresholds are

5 for x_4 and 11 for x_6 according to Eqn. (31) and Eqn. (32). For the remainder of those parameters, possible maximum resources to be used are bound by other means. Since Eqn. (34) is an upper bound for the theoretical maximum charge to be applied, the real maximum charge cannot exceed this value, i.e 8.10 €. Note that this fact is not just an numerical example, instead it is based on equation Eqn. (20). Therefore, from the provider's perspective it is impossible that a charge exceeds the reserved amount of credits within a time interval—there is never, note even a possible loss of money for the provider. This is a clear improvement compared to the hot billing approach, where it exists a possible loss of money, cf. [21].

The charge C for the fifth interval is 7.86 €, leaving only 24 cents on the credits. Calculation of the sixth interval gives a duration of 2 seconds which is lower than t_{min}. The charging system is now preparing a "Stop" message which will be sent to the AG. This message includes an updated list of services from the bundle the customer is still allowed to use. Since the sixth interval is smaller than t_{min}, this updated list must be a subset of all services in the bundle. It depends on the policy of the provider, which services are removed. One useful policy the provider can apply is that it removes all services from the bundle, except the VoIP service. Evaluating Eqn. (23) so that the charge is 24 cents, gives a maximum duration of a VoIP call of a little bit more than 20 minutes. Thus, the concept of time intervals will be applied for a bundle consisting of s_3 only.

If the customer continues using s_3 and if he does not recharge his credits, eventually the time interval will be smaller than t_{min} again. In this case, the only option for the provider is to disconnect definitively this customer. Although, the customer could, theoretically, continue using s_3 for the length of that last time interval that has just been calculated. Thus, a very small amount of credits will remain at the providers side: The charge that correspond to use of s_3 for that last interval, which is $< t_{min}$. But inspecting Eqn. (23) shows that the possible maximum charge using s_3 for even as long as the full t_{min}, yields to a charge << 1 cent, which is negligible.

With the above policy, the provider was able to minimize non-allocatable remaining credits to such a small amount, that a customer satisfaction is guaranteed.

VIII. Summary and Conclusions

First, critical issues for a prepaid charging approach of IP-based services have been identified. Second, an analysis of existing prepaid systems for IP-based networks has shown that no formal and proper solution exists yet to solve those issues in an integrated manner. Therefore third, the concept of time interval-based prepaid charging for service bundles has been introduced. Besides of this simplification of the charging process, the introduced tariff modeling allows for the definition of flexible and user-incentive tariff schemes.

To calculate time intervals for each bundle, it has been shown how these tariff functions are transformed into time-based functions. The TICA algorithm demonstrated how these time intervals are successively applied and calculated on the basis of remaining credits of the customer. Finally, the TICA algorithm has been applied in a real-world example scenario.

The approach derived within this paper is economically and technically reliable, flexible, and scalable, since the overall development considered those characteristics. First, reliability is achieved through formal properties, where it was shown that is impossible to overuse the reserved credits. Second, flexibility is given by supporting any kind of services based on IP and by allowing arbitrary tariff functions. Third, scalability is clearly improved compared to existing alarm-based prepaid system by reducing the overhead of real-time monitoring.

The basis for this time interval-based approach is an all-IP architecture, specifically the one utilized within the Daidalos project. However, all concepts shown in this paper are not restricted to this particular instance of an All-IP architecture. This is due to the fact that, if all technical requirements on this architecture and the accounting system are met, the developed time interval-based prepaid charging model is applicable. Even today's existing network architectures already meet partly these requirements and could, with some extensions, meet all of them.

Further work will extend the service bundle concept to support a dynamical adding and removing of services to and from the bundle. Besides this, calculating the upper bound of resource consumptions—i.e. $h_k(t)$—will be investigated further by incorporating sophisticated traffic modeling. Thus, the TICA algorithm proposed needs to be extended to support these improvements.

Acknowledgement

This work has been performed partially in the framework of the EU IST project Daidalos "Designing Advanced Interfaces for the Delivery and Administration of Location independent Optimized personal Services" (FP6-2002-IST-1-506997), where the ETH Zürich has been funded by the Swiss Bundesministerium für Bildung und Wissenschaft BBW, Bern, under Grant No. 03.0141. The authors would like to extend many thanks to their Daidalos partners.

References

[1] 3GGP, Third Generation Partnership Project, http://www.3gpp.org, June 2005.

[2] 3GPP2, Third Generation Partnership Project 2, http://www.3gpp2.org, June 2005.

[3] 3GPP2 C.S0024. CDMA2000 High Rate Packet Data Air Interface Specification, Release 0, Version 4.0, October 2002. 3rd Generation Partnership Project 2, 3GPP2.

[4] 3GPP2 C.S0001-0. Introduction to CDMA2000 Standards for Spread Spectrum Systems, Release 0, Version 1.0 July 1999. 3rd Generation Partnership Project 2, 3GPP2.

[5] 3GPP2 C.S0024-A. CDMA2000 High Rate Packet Data Air Interface Specification, Release A. Version 1.0, March 2004. 3rd Generation Partnership Project 2, 3GPP2.

[6] 3GPP2 X.S0013-000-0. All-IP Core Network Multimedia Domain. Version 1.0, December 2003. 3rd Generation Partnership Project 2, 3GPP2.

[7] ETSI TR 101 734 V1.1.1 (1999-09). Internet Protocol (IP) based networks; Parameters and mechanisms for charging. European Telecommunications Standards Institute, 1999.

[8] ETSI TS 123 002 V6.7.0 (2005-03). Digital cellular telecommunications system (Phase 2+), Universal Mobile Telecommunications System, Network architecture (3GPP TS 23.002 version 6.7.0 Release 6). European Telecommunications Standards Institute, 2005.

[9] ETSI TS 123 002 V4.8.0 (2003-06). Digital cellular telecommunications system (Phase 2+), Universal Mobile Telecommunications System, Network architecture (3GPP TS 23.002 version 4.8.0 Release 4). European Telecommunications Standards Institute, 2003.

[10] .ETSI TS 132 240 V6.1.0 (2005-03). Universal Mobile Telecommunications System (UMTS), Telecommunication management; Charging management; Charging architecture and principles (3GPP TS 32.240 version 6.1.0 Release 6). European Telecommunications Standards Institute, 2003.

[11] ETSI ES 203 915-12 V1.1.1 (2005-04). Open Service Access (OSA); Application Programming Interface (API); Part 12: Charging SCF (Parlay 5). European Telecommunications Standards Institute, 2005.

[12] IETF AAA (Authentication, Authorization, and Accounting) Working Group. http://www.ietf.org/html.charters/aaa-charter.html, June 2005.

[13] IETF RADEXT (RADIUS Extensions) Working Group. www.ietf.org/html.charters/radext-charter.html, June 2005.

[14] IMT-2000, International Mobile Telecommunication at 2000 MHz, http://www.itu.int/home/imt.html, June 2005.

[15] ITU-T, International Telecommunication Union. D-Series Recommendations, General tariff principles. ITU-T, Geneva, Switzerland, Various Dates.

[16] Alcatel. Innovative Payment and Billing Solution for Broadband Entertainment. Technology Whitepaper, Alcatel Telecommunications Review, 2003.

[17] Andres Arteta. Prepaid Billing Technologies-Which one is for you? Billing World (Billing World and OSS Today Magazine), Issue 2 (February 1998), pp. 54—61.

[18] P. Calhoun, J. Loughney, E. Guttman et al. Diameter Base Protocol. IETF RFC 3588, September 2003.

[19] G. Camarillo, M. A. García-Martín. The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds. Wiley, October 2004. ISBN: 0-470-87156-3.

[20] Jonathan P. Castro. All IP in 3G CDMA Networks: The UMTS Infrastructure and Service Platforms for Future Mobile Systems. Wiley, November 2004. ISBN: 0-470-85322-0.

[21] Ming-Feng Chang, Y-Bing Lin, Wei-Zu Yang. Performance of hot billing mobile prepaid service. Computer Networks, Volume 36, Issue 2-3 (July 2001), pp. 269—290.

[22] Ming-Feng Chang; Wei-Zu Yang; Yi-Bing Lin. Performance of Service-Node-Based Mobile Prepaid Service. IEEE Transactions on vehicular technology, Volume 51, No. 3 (May 2002), pp. 597—612.

[23] DAIDALOS, an EU Framework Programme 6 Integrated Project, http://www.ist-daidalos.org, June 2005.

[24] L.A. DaSilva. Pricing for QoS-enabled networks: A survey. IEEE Communications Surveys, SecondQuarter 2000.

[25] Ericsson. Prepaid Postpaid convergent Charging. Whitepaper, Ericsson, April 2005.

[26] M. Falkner, M. Devetsikiotis, I. Lambadaris. An overview of pricing concepts for broadband IP networks. IEEE Communications Surveys, SecondQuarter 2000.

[27] Harri Hakala, Leena Mattila et al. Diameter Credit-Control Application. draft-ietf-aaa-diameter-cc-06.txt, August 2004.

[28] F. Khan G. Foster, S. Vahid, N. Baker. Comparison of charging methods in 3G mobile networks (consumer behavior versus tariff models). Third International Conference on 3G Mobile Communication Technologies (2002). Conf. Publ. No. 489. pp. 382—387.

[29] M. Karsten, J. Schmitt, B. Stiller, L. Wolf. Charging for Packet-switched Network Communications—Motivation and Overview. Computer Communications, Vol. 23, No. 3 (February 2000), pp. 290—302.

[30] F.P. Kelly. Charging and accounting for bursty connections. In Internet Economics. MIT Press, 1997, pp. 253—278. ISBN 0-262-13336-9.

[31] M. Lilge. 2001. Evolution of Prepaid Service towards a Real-Time Payment System. IEEE Intelligent Network Workshop (May 2001), pp. 195—198.

[32] Yi-Bing Lin; Ming-Feng Chang; Herman Chung-Hwa Rao. Mobile prepaid phone services. IEEE Personal Communications, Volume 7, Issue 3 (June 2000), pp. 6—14.

[33] A. Lior, P. Yegani et al. PrePaid Extensions to Remote Authentication Dial-In User Service (RADIUS). draft-lior-radius-prepaid-extensions-07, February 2005.

[34] Maple. Maplesoft, Waterloo Maple Inc. http://www.maplesoft.com, june 2005.

[35] P. Newman. In search of the all-IP mobile network. IEEE Communications Magazine, Volume 42, Issue 12 (Dec. 2004), pp. S3—S8.

[36] NOKIA White paper. Service charging in the Intelligent Edge. Nokia Corporation, 2004.

[37] Orange Communications, website: http://www.orange.ch, June 2005.

[38] The Parlay Group Website. http://www.parlay.org, June 2005.

[39] D. Ranasinghe, J. Norgaard: Charging Generalized. A Generic Role Model for Charging and Billing Services in Telecommunications. “Intelligent networks”, Edt. J. Harju, T. Karttunen, O. Martikainen, Chapman Hall, London, England, 1995, pp. 159—172.

[40] Sunrise mobile communications, website: http://mobile.sunrise.ch, June 2005.