Real-time charging in mobile environment

BÁLINT DÁVID ARY
ary.balint@isolation.hu
DR. SÁNDOR IMRE
imre@hit.bme.hu

Key words: content provision, UMTS, charging and billing, network structure

Charging the services offered by the packet based UMTS environment is much more complex, than it is in the circuit based GSM systems. The situation is even more complicated, as services can be easily developed using standard APIs and can be offered by third party providers. Moreover, pre-paid users need a real-time approach for charging and billing, which limits the admissible charging solutions. In our study we give a short survey of the motivations for developing the UMTS system, we summarize the legal and technical difficulties, and we introduce our new concept which should ease the majority of the problems and lighten the network overhead caused by the real-time charging solution presented in the corresponding technical reports.

1. Introduction

On the turn of the 20th and 21th century, the mobile telecommunication equipment went through a powerful development. After the early analogue systems the GSM (Global System for Mobile Communications) appeared, and the UMTS (Universal Mobile Telecommunication System) is being introduced nowadays in many countries in Europe. The evolution includes the increase of bandwidth, the fulfilment of the All-IP concept and the appearance of multimedia capable devices, thus different multimedia services like on-line streaming, video conferencing, Internet browsing and several packet based, on-demand services could be available with a next generation mobile phone. These changes were mostly called forth by the information society, which spends more and more money on communication. Even if these technical evolutions are not always urged by the subscriber demand, the new possibilities and functions are getting used widely.

Temporarily, the services accessible with mobile devices are offered by the network provider (who is the content and application provider as well), but with the continuous growth of the number of accessible functions and media it is predictable, that the network operator won't have enough time and/or energy to invent and offer new services, although it could lead to significant superiority in the market competition. With these conditions the supply of infrastructure and content should decouple, so content and applications should be served by 3rd party providers.

International telecommunication companies have made huge investments in UMTS, although a fully operating 3rd generation mobile network does not exist yet. The reason is that the changes are so significant, that the existing management systems are unable to handle these new demands. Not only the services should be developed but also the managing part should be revised, extended, and new features must be added, to realize the functions defined in the standards.

One of the main parts of the management system is charging. The return of the invested funds can only be hoped by new "killer-applications" and their proper accounting. The charging system of GSM networks was not designed to handle the bandwidth and the data/media types of UMTS services. A new charging concept should be developed, capable of handling the existence of 3rd party providers, and to support all kind of charging methods, like pre-paid, paynow, or post-paid mode.

The new architecture must be compatible with the existing GSM network charging architecture, and must operate real-time. Because the subscribers want to pay to only one provider (one-stop-shop concept), this accentuated provider has to maintain a financial relationship with the other providers and has to settle the invoices. The business model must be flexible, in order to support all combinations of content and provider relationships, hence the charging system must be accurate, convenient, transparent, and must be able to cooperate with other autonomous systems [6].

2. Business models

The amount of money paid after a service should be shared among the network operator and the content provider. But because the subscriber dislikes paying to more than one supplier for one service, the two providers should maintain some financial relationship, and settle the liabilities periodically after validation and identification. The content provider and the network operator have to agree on the parameters of the provided services (e.g.: required transport quality, parameters to be measured). Both providers must authenticate the user, and must know his or her financial status to decide, whether to accept or reject the service request.

Nevertheless, the rendition of the full account due to the user's personal rights is not viable.

Taking the business scope of 3rd generation networks, several business roles exist. The network operator provides access and transport services. The role of the content provider is to provide services, contents or applications that add value to transport services. These applications or contents can be produced by the content provider itself or purchased from other providers. The key function of the content aggregator is to package and offer services from one or several content providers. In third generation networks the key solution is flexibility [6]. A subscriber can belong to the content or the network provider, and both of them can charge the user.

Although several combinations of the roles and relationships are possible, the UMTS Forum outlined the most presumable business models [7]. A charging system must be prepared to deal with the different combination of business roles, and the charging systems of the providers must be compatible with each other.

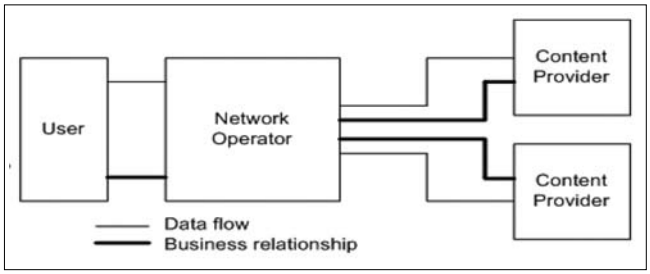

Figure 1. Network Operator Centric Business Model

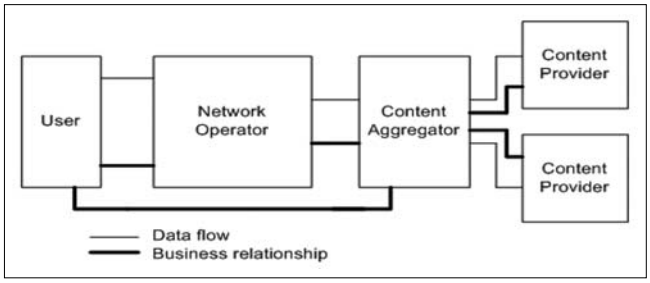

Figure 2. Content Aggregation Centric Business Model

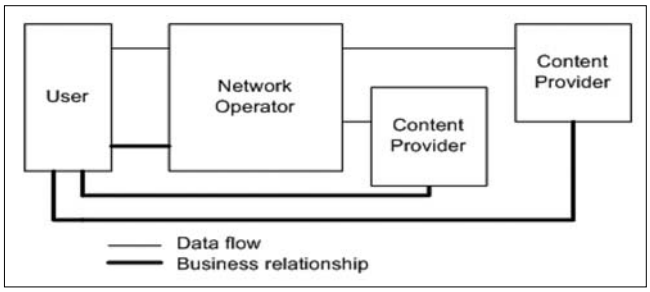

Figure 3. Content Provider Centric Business Model

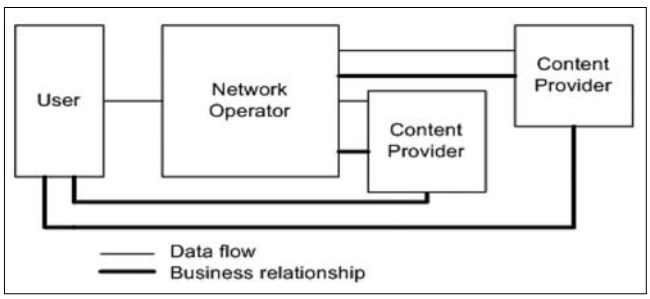

Figure 4. Hidden Network Operator Model

2.1. Network operator centric business model

In this model *(Figure 1.)* the network operator provides the content indirectly, charges the user and does the payoff to the 3rd parties. The subscriber can use the money on his/her infrastructure-account to pay for the content. In this way, content providing seems to the user like a value added network resource usage. The provider does not store the data and isn't responsible for it's content, therefore the Internet connection strongly determines the quality of the service (QoS). This model is the most convenient for the subscriber, although the operator has full control over the content providing.

2.2. Content Aggregation Centric Business Model

In the content aggregation centric business model *(Figure 2.)*, the content is accessed through a portal (which is not part of the mobile network). The service cost is split into two parts: the cost of the access to the aggregator (network resource usage) which is paid to the infrastructure provider, and the cost of the content accessed, which is paid to the aggregator. The fee of the content is defined by the content aggregator, who may be in connection with other content providers. In this way a chain of providers is involved in the transaction. To avoid the subscriber's chagrin, caused by the multiple payoffs, the content aggregator and the network operator should be in financial relationship and settle the bills among the 3rd party providers.

2.3. Content Provider Centric Business Model

The content provider centric business model *(Figure 3.)* is quite similar to the content aggregation centric business model, but the content provider plays the role of the content aggregator as well. Because of the huge number of 3rd party providers, the realization of the business relationship is much harder, than it was in the content aggregation centric business model. The main disadvantages of this solution are that the content/application providers must solve the accounting of services on their own, which can be more expensive than the service itself [10] and that the subscribers have to maintain an account with every content provider separately. This solution could lead to problems in case there are many providers. This model brings huge freedom to the services offered, but it means enormous administrative overhead as well.

2.4. Hidden Network Operator Model

In this model *(Figure 4.)*, the network operator stays unrevealed for the subscriber, as the content provider provides the mobile equipment, services and applications for the user and pays the necessary fees for the network operator. The model is suitable for companies with a very strong brand and with a tough customer base.

3. Technical challenges

In wired communication and in the circuit based GSM systems accounting was much easier. Because of the permanent and reserved bandwidth, the price of the service depends only on the length of the connection. The GPRS (General Packet Radio Service) and the UMTS are packet based, so the measurement of value and quality of the service are more complex.

In order to compute the quantity of the service, we should count the bits that have gone through the system. This method would require huge computing capabilities from the network elements (because of the high speed transport) and would impose a big overhead on the system (for N transmitted bits, the exact size can only be written in $\log_2 N$ bits). Counting of packets isn't the perfect solution either, inasmuch as the packet lengths in IP network vary in considerable range. Because of the defective quality of the Internet, a correct method should be aware of the lost and doubled packets. The additive cost because of these failures should not be charged to the subscriber.

In non-circuit switched systems the measurement of quality means problems as well, because in best-effort services a fixed reserved bandwidth is absent. Without a permanent connection a guarantee for capacity and delay can only be given with heavy signalling. In multimedia services the measurement of QoS is especially hard, because (f.e.) in case of video-streaming, the actual content affects the minimal requirements for the quality.

We showed that measuring quantity and quality is a quite complex task, which imposes notable overhead on the system. Furthermore, in a pre-paid environment, it must be done in real-time. In the present solutions, most providers solve it by combining data measuring with some easily measurable unit (time based accounting or constant bit rate accounting), or the measurement is done with greater scale (more kilobytes for instance) [9].

Another problem can be derived from the mobility. Every equipment has an IP address in the UMTS environment. If we use a fix IP address, and update the router tables in the network, the movement would be transparent for the charging mechanisms, but the router updates would cause overhead and signalling problems in the network. If the IP address changes continuously the network elements (which supply the charging information) must be informed respectively.

In the UMTS system, several media are accessible *(Table 1.)* [1]. Standards give possibilities to subscribers to possess separate accounts for all media available in the system. In case of 3rd party providers, if the accounting is done by the network provider, the operator should be aware of the exact method of charging and the measurable parameters of the service.

In 3G mobile networks it is possible for the users, to gain information about the price of the services, before the actual requisition. This supplementary service is ensured by the AoC (Advice of Charge) function. The mobile equipment retrieves charging-related information from the network, which contains the dependence of the price on service-time or service-data, or in case of an event based service (MMS for instance), on the entire, exact price of the service. This information is obviously sent by the same network element, which manages the charging and billing of the given service (the network operator or the 3rd party provider, depends on the business model used); nevertheless, if the information is sent by a 3rd party, the information should be validated and supervised [5].

4. Charging in UMTS networks

Besides the service and quality measurement the charging process also includes the settlement of invoices among the serving parties (subscriber, network and content provider). The price of network usage must also be settled between the network providers in case of roaming. This procedure is standardized, and uses the transfer account procedure (TAP) and a specific TAP format [1]. The construction of the bill presented to the user is also important; it must be simple and easily understandable [6]. The real money transaction between the parties (including the user) is usually obliged by contracts. These problems, solutions and mechanisms are beyond the scope of this article.

Standards define two different paying modes (prepaid and post-paid) and two charging methods: the offline and the online charging. The paying mode indicates the time, when the user has to pay for the service: afore or after the service require. The charging method indicates whether the subscriber's account is managed real-time (online) or not (offline). In offline charging, charging information is gathered after the requisition, and so, the subscriber account is debated after the service. Since this information is collected after the event/service, and sent trough a widespread network, real-time charging is not possible. The online charging mode assures that services are applied only if the subscriber has the necessary amount of money for them.

In offline mode the gateway (GGSN – Gateway GPRS Support Node) and the inner-nodes (SGSN – Serving GPRS Support Node) are sending charging in-

Table 1.

- speech,
- voice (real-time / streaming),
- video (real-time / streaming),
- data (download / upload / interactive content),
- messages (SMS / E-mail),
- data-flow (unspecified content),
- accessed web-pages, portals, etc.

Table 2.

- determinate data amount,
- determinate time-interval,
- the change of charging conditions,
- the change of QoS,
- the change of tariff,
- the change of position or cell,
- and the closure of voice, data or multimedia sessions

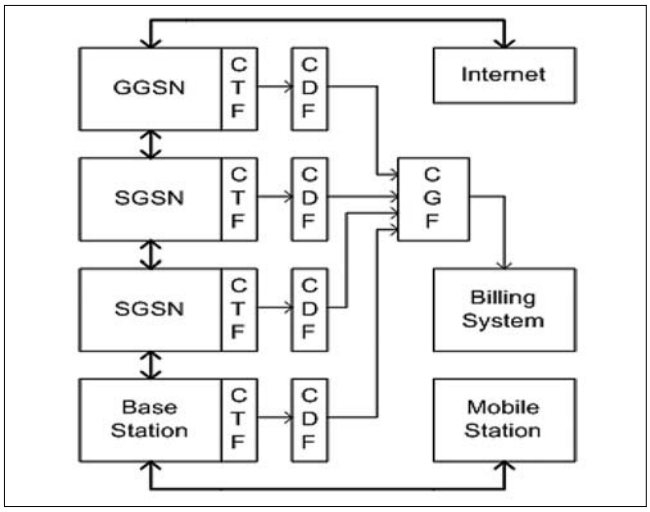

Figure 5. The offline charging architecture

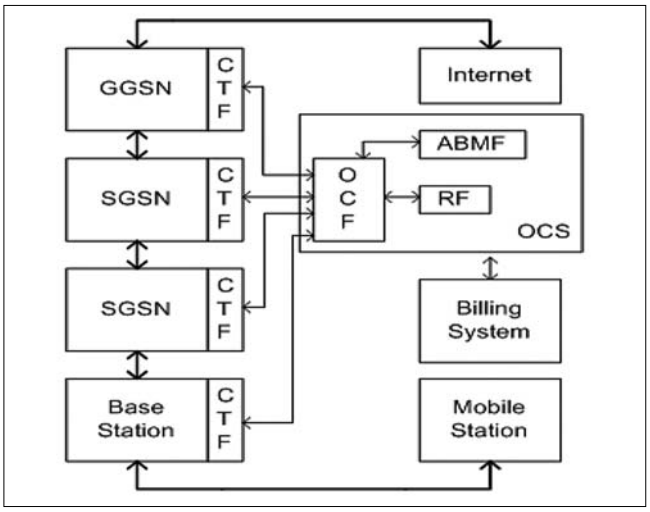

Figure 6. The online charging architecture

formation to the Billing System (BS). This charging information must be in standardized format, called Charging Data Record (CDR). The Charging Trigger Function (CTF) of the network elements generates charging events *(Table 2.)* based on the observation of network resource usage. The Charging Data Function (CDF) receives charging events from the Charging Trigger Function (CTF), and then uses the information contained in the charging events to construct CDRs. These records are sent to the Charging Gateway Function (CGF), which acts like a storage buffer, cleans, and preprocesses the CDRs. Finally, the CGF sends these processed CDRs to the Billing System *(Figure 5.)*. Because these charging records carry every information about the services required, the functionality of the CDRs extends beyond charging. With CDRs it's possible to analyze service-utilization, and gain statistical information about the services and content. By archiving the CDRs, the user-complaints can also be easily settled [1].

According to the standards [2], post-paid users can limit their account for a specific service; in light of this, a real-time charging method should be used for prepaid users and for post-paid users with credit limit. To ensure this, online charging should be applied.

In online mode *(Figure 6.)* the Online Charging System (OCS) is responsible for proper charging. The main task of this function is to realize real-time charging by continuously delegating certain amounts of credit to the serving network elements. These credits are deducted from the user's account. This method is called: unit reservation. If the service terminates before all credits are consumed, the network elements are retransfering the remaining credits to the OCS. To assure continuous service delivery, if the users do not terminate the service, a new amount of granted credit should be sent to the serving network element before the previous one runs out [3,4]. This unit-granting function is represented by the Online Charging Function (OCF) inside the OCS. The CTF generates charging events for the OCF as well, but this communication is bidirectional, as the OCF has to grant credits for the service. The OCS also includes the Account Balance Management Function (ABMF) and the Rating Function (RF). The Account Balance Management Function is the location of the subscriber's account balance within the OCS, and the Rating Function is used to determinate the value of the network resource usage and responsible for the

- rating of data volume (e.g. based on charging initiated by an access network entity),
- rating of session / connection time (e.g. based on charging initiated by a SIP application),
- and for rating of service events (e.g. based on charging of web content or MMS).

5. Mode-switching model

For a correct modelling it is obligatory to suit the related standards. The optimal model can be developed using the proper determination of the free parameters. Such variable parameters are the amount of data and/or time that triggers the CDR generation and the amount of granted credit during unit reservation. The smaller data/time amount we use, the more accurate the charging, and the larger the network overhead will be. It can be seen, that some trade-offs are necessary. Other variable parameters are the physical realization of the charging functions, inasmuch as these functions are not attached to hardware entities. The third variable parameter or method is the measurement of the services. The standards don't deal with the measuring methods, therefore for data transfer, the estimation of bandwidth or the exact bit count are possible solutions.

Since the unit reservation message should contain more or less the same information as the CDRs, we assume that they have the same size. In online charging, if the service reserves a large amount of credit from the user's account, access to additional, parallel resources could be denied, because there is no credit left on the account for another resource usage request; even if some service terminates afterwards, and the unused credits are returned to the users. In light of this, a more frequent unit reservation, with a smaller amount of cre-

dit should be applied. Because CDRs indicate the used services/data, this problem doesn't occur during offline charging. As follows, online charging causes bigger network overhead, than offline charging.

Our idea was not to glue the charging mode to the type of the payment (pre-paid, post-paid), but to dynamically switch between offline and online charging (if online charging is required) considering the user's account as well. Moreover, the overhead of the continuous unit reservation can also be reduced, by granting units only once. The quality of service should also be supervised, in order to charge services properly.

In our model, we assign a service specific limit to every service offered. If the user's account is above this limit, then charging is done in offline mode. If the subscriber's account drops below this limit, the online charging mechanism is applied (if required), and we grant all the consumable credit to the serving network element. In multi-task systems, it is possible to access more than one service at a time. In such cases, when the account drops below the limit, we shall delegate the credits to multiple network elements. A good solution is to distribute the account among the services with statistical methods, considering the money-consumption and properties of the services, and the behaviour of the user.

The UMTS services are based on a packet switched network, so we have to count with the packet-loss. The majority of these failures occur on the wireless part of the network but, of course (like on the regular Internet), some packet-loss or fault happens on the backbone as well. Statistical methods can be used to deal with these failures. Considering the quality of the operator's network, we can send more packets to the user, than it would be necessary with a perfect, flawless network, so the user presumably gets the proper amount of packets. It's practical to include a buffering mechanism between the wired and wireless part of the network, to cause the packets to be resent only from the base station in case of any failure, so the backbone isn't loaded with this traffic. The loss or fault occurring on the backbone can be solved with the error correction mechanism of the TCP – if necessary.

In order to measure the packet-loss and packet based QoS, the presence of trusted equipment is needed at the end of the connection. This could be the base station, or we can implement the protocol in the low level layer of the mobile phone. The main idea of this solution is that the element has to send some kind of information to the billing system, in order to inform it about the quality. The quality measurement of the data sequence is done with a sliding-window algorithm. After the arrival of a proper amount of packets, the delay (average, maximum, minimum and jitter), packet-loss, bandwidth and other QoS parameters can be calculated. The retransmission of the lost packets and signalling is done by higher protocols. The measurement of quality can be eliminated with the usage of pre-calculated statistical information for the network, but in this case the results won't mach the exact situation.

6. Analytical supplementation

Our model needs further refinement on order to determine the mode-switching limit, the handling of lost packets, and the measurement of QoS.

6.1. Mode-switching limit

Let us define a function called unit consumption speed

$$C(T), \tag{1}$$

having the measure of [*unit/sec*], which represents the consumed units in one second. The consumption rate depends on time to give the possibility to the operators to assign different prices to different time of the day and week for traffic shaping reasons. The consumed unit and money can be calculated from the consumption rate by means of the following equations

$$unit = C(T) \cdot \mathrm{t} \tag{2}$$

$$money = unit \cdot R(T), \tag{3}$$

where $R(T)$ represents the relation [2] between unit and money. The time-dependence of this function can be used to change the price of the units in case of inflation or discounts, or to apply different prices for different groups of users. Although the time dependence of the price can be divided into consumption speed and rating, it is not necessary, and it depends on the needs of the network operator.

Let T_c represent the time needed to query the user's proper account. The network elements are sending the CDRs usually in bigger time-intervals and the billing system debit the user's account periodically. T_c represents these intervals. With these notations and definitions the limit for mode-switch can be calculated. In ideal case it is

$$L = C(T) \cdot T_c. \tag{4}$$

If we own more units on our account than L, the charging is done offline with small network overhead; otherwise accounting is done online, with unit reservation. If we require more than one service at a time, the limit can be calculated by the sum of the limits of the services:

$$L = \sum L_i. \tag{5}$$

To reduce the network overhead, all credits below this threshold can be reserved. In case of multiple service demands, the units can be distributed to the serving network elements with the rate of the service's consumption speed. A re-sharing should be done every time a service ends, a new service started, or when an event based service occurs (SMS – for example). In order to ensure this, new functionality is required. The online charging function (OCF) should be able to force the network elements to retransfer the currently unused credits. After the transfer, the online charging function could re-share the credits among the services considering the new circumstances. When a fix consumption speed can not be assigned to the service (browsing or interactive content), the average consumption speed should be determined using various statistical models.

6.2. Propagation Delay

The events occurring in a distributed, wide network (signalling, queries) have propagation delay, which is not

constant in general. If we want to determine the mode-switching threshold properly, we have to consider the time needed the query the account (T_c) and to switch between modes (T_d), together with the variation of these values (T_{cj} and T_{dj}):

$$L = C(T)\cdot(T_c+T_{cj}+T_d+T_{dj}). \quad (6)$$

To ensure accurate charging, we should count with the maximum values of the jitters (T_{ci} and T_{di}). If we want to reduce the values of the mode-switching limits (in order to reduce the network overhead), we shall count with smaller values (with the expected value for example). In this case the possibility of users gaining more service than they paid for can be calculated from the distributions of the jitters. In case of re-sharing the control messages should be labelled with proper time-stamps to be able to charge the services gained during the retransfer and mode switching process.

The mode-switching thresholds can be calculated off-line for every service offered, and the system can use these pre-calculated values to switch between the charging modes. However, the actual limit can be dependant on the time of the day and on the user profile (discounts for group of users, statistical behaviour for interactive content).

6.3. Measurement of QoS

Performance can be defined using a sliding-window algorithm; always using the last N packets arrived to the user. With this method, the measured and experienced performance should be close to each other. Let t_j be the transmission starting time and a_j the arrival time of packet j. If the size of the sliding-window is N, the delay (average, minimum, maximum) can be calculated:

$$D_{average} = \sum(a_i - t_i) / N, \quad (7)$$

$$D_{min} = \min(a_i - t_i), \quad (8)$$

$$D_{max} = \max(a_i - t_i). \quad (9)$$

The jitter of the delay is the difference of the maximal and minimal delay:

$$D_{jitter} = D_{max} - D_{min}. \quad (10)$$

The packet-loss in case of N arrived, and M sent packets is:

$$Loss = N/M. \quad (11)$$

7. Conclusions and future plans

In our study, we enumerated the motivations for the appearance of 3rd party providers. We have showed the legal and technical issues of this new concept. Most of the technical problems come from the real-time nature and mobility in the packet based network. We also gave a small summary about the current state of the 3GPP standards, and finally, we gave a model to solve these problems. The model operates in such a way, that charging is made in the network offline, without a need for a real-time approach, to a large volume of users (who have more money on their account than the critical amount). This method invokes low CDR transfer, and low network overhead. Billing to critical users is more complicated, but also supported by 3GPP standards. With this idea the necessary network overhead can be decreased. Moreover, with a small function extension and statistical estimation, the overhead can be further reduced.

In the future, in order to develop the complete charging method, it is required to work out the exact method of measuring the data flow and the method to derive the quality of service from the IP based quality. For this, it is crucial to determine the statistic parameters of the services and users. The model is not complete unless the protocols and algorithms are fully developed.

References

[1] ETSI TS 122 115 V5.3.0 (2003-06). "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Service Aspects Charging and billing (3GPP TS 22.115, ver. 5.3.0, Rel. 5)." Technical report, 3GPP, 2003.

[2] ETSI TS 132 200 V5.7.0 (2004-06). "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Telecommunication management; Charging management; Charging principles (3GPP TS 32.200, ver. 5.7.0, Rel. 5)." Technical report, 3GPP, 2004.

[3] 3GPP TS 32.240 V6.0.0 (2004-09). "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Telecommunication management; Charging management; Charging architecture and principles (Rel. 6)." Technical report, 3GPP, 2004.

[4] 3GPP TS 32.260 V2.0.0 (2004-12). "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Telecommunication management; Charging management; IP Multimedia Subsystem (IMS) charging; (Rel. 6)." Technical report, 3GPP, 2004.

[5] ETSI TS 122 086 V5.0.0 (2002-06). "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Advice of Charge (AoC) supplementary services; Stage 1 (3GPP TS 22.086 version 5.0.0 Release 5)." Technical report, 3GPP, 2002

[6] UMTS Forum Report No.11. "Enabling UMTS 3rd Generation Services and Applications." Technical report, UMTS Forum, October 2000.

[7] UMTS Forum Report No.21. "Charging, Billing and Payment Views on 3G Business Models." Technical report, UMTS Forum, 2002.

[8] Maria Koutsopoulou, Alexandros Kaloxylos, Athanassia Alonistioti, Lazaros Merakos: "Charging, Accounting and Billing Management Schemes in Mobile Telecommunication Networks and the Internet." IEEE Communications Surveys, First Quarter 2004, 6(1), 2004, pp.50–58.

[9] Susana Schwartz: "Next-Gen Rating: It Will Be Only As Good as the Network." Billing World & OSS Today, February 2003, pp.16–22.

[10] John Cushnie: Charging and Billing for Future Mobile Internet Services, First Year PhD Research Report, September 2000.