Eruption of Policy in the Charging Arena

Marc Cheboldaeff
Payment & Charging Solutions
Alcatel-Lucent
Ratingen, Germany
Marc.Cheboldaeff@alcatel-lucent.com

***Abstract*—In the early days of mobile Internet, bandwidth was not an issue, so price plans were quite simple, very often expressed in the form of a flat rate or "all you can eat" pricing. As long as capacity was greatly available, this was a convenient and simple way to define a tariff, both for the end-user and the service provider. With the tremendous growth of data traffic observed recently, bandwidth becomes more and more a scarce resource. Consequently, flat rate pricing leads to a minority of heavy users cannibalizing the whole resource, while being subsidized by low users. This is of course not acceptable! It is neither fair for the majority of end-users, nor profitable for the service provider. The goal of this paper is to study how service providers can grant the required quality of service to the "right" users, in other words to users who will generate revenue for their use. This will improve overall customer experience in the end, while service providers can see the return on their investment in network infrastructure by somehow "monetizing the bandwidth".**

Keywords- Rating; Charging; IMS; OCS; Policy; PCRF; PCC; QoS; QoE

I. INTRODUCTION

The Charging and Policy topics in telecommunications networks cannot be considered as two distinct topics anymore. The days where policy management was considered as a pure network internal mechanism are over. Determining the right Quality of Service (QoS) is not only a network management topic like congestion control or call gapping. Main reason is that policy decisions do not depend only on the network traffic or load at a certain point in time.

Of course, policy decisions depend also on the kind of contents that is being transmitted: high-definition videos obviously require a better QoS than poor-quality videos. Similarly, progressive downloads do not require the same policy as live streaming.

Furthermore, policy decisions depend on the type of device as well: sessions triggered by older handsets do not require the same quality as sessions triggered by latest smart phones. In addition, they depend on the underlying technology too: it might not be necessary to grant the same QoS for a data session running on a General Packet Radio Service (GPRS) network or Universal Mobile Telecommunication System (UMTS) network, than on a Long Term Evolution (LTE) network. The GPRS architecture, sometimes called 2.5G (intermediate stage between the second and third network generation) is described in [2], the UMTS or 3G architecture is described in [3], while the LTE architecture is described in [4]. The reader might refer as well to the Terminology section at the end of this paper to get the meaning of the various acronyms used.

Independently of these technical aspects, policy decisions should most importantly depend on subscriber's personal information, which encompasses business information including, but not limited to, the price plan. This is the aspect that we are going to tackle in this paper.

We shall present first the evolution of pricing schemes from fixed tariff plans to flexible offers taking into account QoS, emphasizing the importance of real-time policy and charging decisions. We shall then investigate which subscriber data is relevant in this context. Afterwards, we analyse the technical impacts in order to achieve real-time policy and charging control on an individual basis. We then design a solution, and describe its implementation. In the subsequent sections, we review other possible solutions and the position of standard bodies in this area. Finally, we address a framework aiming at changing policy in a more user-friendly way.

II. HIGH QOS AS A TARIFF OPTION ?

The old days of fixed price plans, i.e., "one size fits all", are definitely over. Nowadays, service providers tend to target specific market segments with dedicated offers in order to increase customer satisfaction and avoid subscribers' churn.

Subscribers are not expected to just accept generic tariff plans anymore, but instead they are invited to take actively part in the definition of their own "tailor-made" tariff. Often, subscribers can choose a base or default tariff, on top of which they can combine various options, each being applicable to a certain usage; for example a bucket of 100 roaming voice minutes valid 30 days, or a renewable monthly bucket of 1 Giga Byte (GB) for data traffic from the home network, etc. This is illustrated in Figure 1. Additionally, the reader who wishes to get more insight on the increased diversity of tariff options and their technological impact can refer to [5].

The left part of Figure 1 represents old tariff schemes, for example in Public Switched Telephony Networks (PSTN), where a fixed rate per time interval is usually defined, while the right part of Figure 1 depicts newer tariff schemes, where various pricing components can be combined freely.

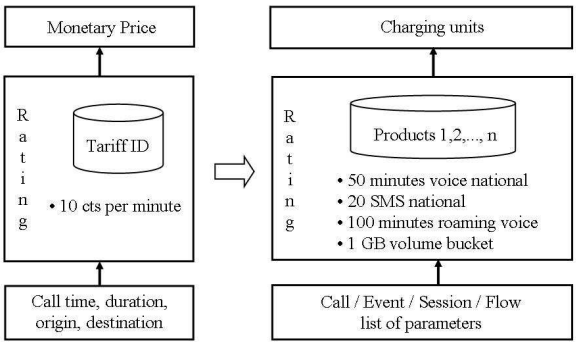

Figure 1. Evolution of tariff schemes

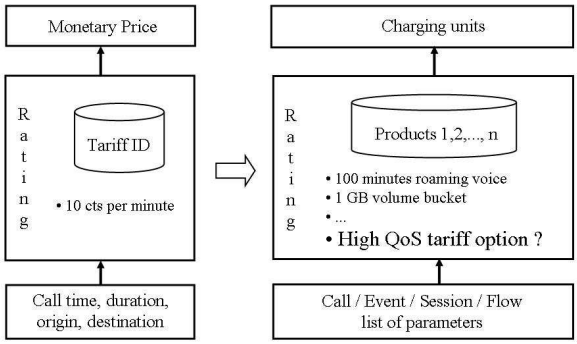

Figure 2. High QoS as a tariff option

These tariff options can be considered as different products, that the end-user may want to buy or not. For these tariff options, the charging unit is not necessarily money: once the user buys for example a bucket of data volume, he/she has a certain amount of Kilo Byte (KB) or Mega Byte (MB) on his/her account, so that a rating engine does not necessarily need to calculate a price at each session or event, but a volume amount.

Of course, these tariff options presented as products to end-customers should be easy to understand by the latter. If increased flexibility leads only to confusing complexity, there is no added-value! If a customer can easily represent for himself/herself what a number of minutes or Short Message Service (SMS) texts means, it might be more difficult to understand what it means for mega bytes! What can an end-user do with 100MB for example? How many pictures, how many mails can be retrieved? Not so easy to determine… So the options may be presented in a more user-friendly way, like an unlimited bucket applicable to Facebook or YouTube. Buckets could mix multiple traffic types too, like a Twitter bucket including Data and SMS.

In order to preserve the customer Quality of Experience (QoE), a data bucket should go hand in hand with a minimal QoS. Indeed, it would be frustrating for a subscriber to pay a certain fee to get 1GB for data traffic, and then be confronted to low speed and delays when surfing on the Internet from a mobile device! Here we see an initial correlation between the subscriber's charging profile and the subscriber's policy profile.

Furthermore, certain subscribers might be willing to pay, on top of their base tariff, which might include standard data traffic, a certain fee for having a high QoS guaranteed, independently of usage, whether cumulated or not. They just want to be sure that whenever they are going to access mobile Internet, high-speed will be guaranteed. Such a "High QoS" tariff option is represented in Figure 2.

On the right part of the picture, a "high-QoS tariff option" means that the subscriber pays a certain fee, and gets a guaranteed QoS in return. In fact, guaranteeing QoS in Internet Protocol (IP) networks, which typically work in "best-effort" mode, may not be technically achievable, but at least *prioritization* of premium users could be an option [6].

Of course, the customer should not have the feeling that in order to get a decent normal QoS, he/she has to pay more. If QoS is sold as a tariff option, it means that the obtained QoS will be beyond normal, or that this user will be prioritized over standard users.

III. IMPORTANCE OF REAL-TIME POLICY & CHARGING DECISIONS

In the previous section, we mentioned the possibility for a subscriber to buy a bucket or certain amount of units for a defined data usage. In a simple tariff offering, it might happen that the subscriber's default tariff does not cover data traffic, so that data traffic is allowed only when a data bucket is purchased by the subscriber on top of the default tariff.

In other words, when the data bucket is exhausted, data traffic should be blocked. However, the exhaustion event and thus the blocking effect might occur in the middle of an on-going data session. Such a behavior is not so user-friendly, even if notifications may be sent out for example when 80% and 90% of the bucket has been consumed already. This scenario illustrates though a basic interaction framework between charging and policy: if a data option is valid in the subscriber's profile, then traffic is allowed; if no data option is available, then traffic must be blocked.

Since a data option relates to a data usage, the charging system should track, preferably in real-time, the value of a corresponding usage counter for the subscriber. Consequently, the rule above could be expressed in the following way: if the counter value is lower than a pre-defined limit, e.g., 1GB, meaning that 1GB have not yet been consumed in the current period, then data traffic is allowed; if the counter value exceeds the limit, then data traffic must be blocked.

This is the kind of behavior that was required to be implemented as mandatory by European regulation authorities in order to control the cost of roaming data traffic and avoid "bill shocks" to subscribers. According to this regulation [7], from July 1st 2010 onward, the user should be notified when reaching 50€ of international data consumption. The rule is applicable both to prepaid and post-paid subscribers. The subscriber should be able to define a different limit if the possibility is offered by the service

provider, or opt out of this bill shock safeguard entirely. Here, we see the importance of real-time or online charging, i.e., "charging information can affect, in real-time, the service rendered and therefore a direct interaction of the charging mechanism with session/service control is required" as defined in [8].

Online charging is opposed to off-line charging, where charging takes place after usage is reported, with a certain delay, usually based on Call Detail Records (CDR) or Session Detail Records. During this delay, some chargeable traffic may occur. The cost might be quite high, even for a short time interval, in the case for example of roaming traffic. If a service provider had committed to block traffic when a certain limit is reached - eventually temporarily awaiting customer's willingness to continue - and the consumption is actually blocked when the consumption is already over this limit, then the delta cannot be legally charged to the end-customer, so it means a revenue leakage for the service provider in the end.

It should be noted that the distinction between online and off-line charging is not the same as the distinction between prepaid and post-paid. The second distinction refers to when the payment is made, whether prior to usage or afterward. However, online charging makes sense both to prepaid and post-paid subscribers, for example if end-users want to know exactly at a certain point in time how much they have spent in a billing cycle, or again, if traffic should be blocked exactly when a certain usage limit is reached.

IV. POLICY & CHARGING DECISIONS INFLUENCED BY USAGE COUNTERS

In the previous section, the decision on policy is only "allow" or "block", so actually a dedicated Policy Function is not mandatory as such in this scenario, because an online charging system is already able to cut-off a data session when a usage threshold is reached, in the same way as it can cut-off a data session when a prepaid subscriber's balance is exhausted.

Assuming now that the subscriber is entitled to make data traffic in his/her default tariff, and not only if he/she buys a data bucket on top of it, a smarter scenario would consist in throttling the data traffic when the limit is reached instead of blocking it. In other words, the subscriber would enjoy a high data speed as long as usage is charged from the data bucket, and a reduced speed when data usage is charged from the subscriber's main balance. In this case, the charging system is not the only one impacted; policy control is needed too, because QoS needs to be changed when some usage threshold is reached, and this change should happen again preferably in real-time.

Looking into the real-time aspect in more details, in fact, it would not be a big issue if QoS was not reduced in real-time, because it would be to subscriber's advantage: the subscriber could enjoy a higher QoS a bit longer than what he/she should. At the opposite, if QoS needs to be restored, or increased, it is important for the customer's quality of experience that it happens in real-time. Indeed, it would be frustrating for a user to book a new data option through an Interactive Voice Recognition (IVR) service menu, or by clicking on a pop-up window at the beginning of a download, and then have to wait some time till the QoS is actually increased.

For the sake of simplicity, we mention mainly traffic speed as attribute defining the Quality of Service in this paper, but of course QoS encompasses other attributes than speed, like delay, jitter, etc., so that a good QoS cannot be just reduced to high traffic speed. The reader, who wishes to have more insight on the definition, measurability and feasibility of a good QoS, QoE, etc., should refer to [9].

In this section, we presented use cases where the value of a volume counter should trigger a policy change. In fact, this is a good way to control heavy users, and make sure that their high usage is translated in terms of revenue for the service provider. However, we can think of other subscriber data which might trigger policy decisions too, independently of volume usage. Let us give a few examples in the next section.

V. OTHER SUBSCRIBER DATA INFLUENCING POLICY

Especially in the world of prepaid charging, subscribers are assigned a certain life cycle. A life cycle is a finite-state machine consisting of states, the transition from one state to another being triggered by the expiry of a certain time interval or by some action. For example, when a prepaid card is sent to a retailer's shop, its state might be "pre-active". When the subscriber is making the first call, after some welcome announcement is played, the card might move to a different state like "active", so that the subscriber will not hear again the welcome announcement at the next call. If the subscriber is not performing any recharges for six months, the state might move to "near expiry" and the subscriber can only enjoy limited functionality; for example, he/she might not be able to make international calls. If there is still not any recharge one month later, the card might become "inactive", etc.

In the context of policy control, we see that life cycle transitions could influence policy decisions too. For example, if a prepaid card is near expiry and the subscriber cannot make international calls anymore, maybe he/she should be throtted as well when doing mobile Internet? In this case, the transition from the life cycle's state "active" to the state "near expiry" should trigger a policy change in order to reduce the QoS. As soon as the subscriber performs the next recharge, the subscriber's state moves back to "active" and simultaneously the QoS should be set back to normal. In other words, the transition from the life cycle's state "near expiry" back to "active" should trigger another policy change, in order this time to restore the QoS.

Speaking about recharges, if a prepaid subscriber is performing lots of recharges in a short time frame, it means that he/she generates lots of revenue for the service provider. So maybe this subscriber should be paid special attention and be guaranteed a high QoS for any data session that he/she is attempting? Here again, we see that the criterion for the policy decision is not strictly usage, but the amount of recharges over a recent period.

Extending this framework, service providers can run some profiling tool on their subscribers' database or Data

Warehouse (DWH), and elaborate sophisticated policy rules based on the subscriber's charging history and behavior in order to grant the best QoS to what they consider the "best" customers. We see here a correlation between policy determination and loyalty management.

In this context, not only the network characteristics would decide how policies are granted, but subscriber profiles too. The approach would evolve from a *network-centric* approach to a *subscriber-centric* approach. The legal aspect of Network Neutrality should not be neglected here: in general, traffic should not be blocked if it originates from certain group of subscribers or from certain applications.

What are the technological impacts of this subscriber-centric approach in terms of network architecture? This is what we are going to study in the next section.

VI. IMPACTS IN TERMS OF NETWORK ARCHITECTURE

According to the 3rd Generation Partnership Project (3GPP), whatever the network access technology is, whether GPRS, UMTS, Wireless Fidelity (WiFi) [10] or LTE, data traffic transits through a packet gateway. This is represented in a simplified way in Figure 3.

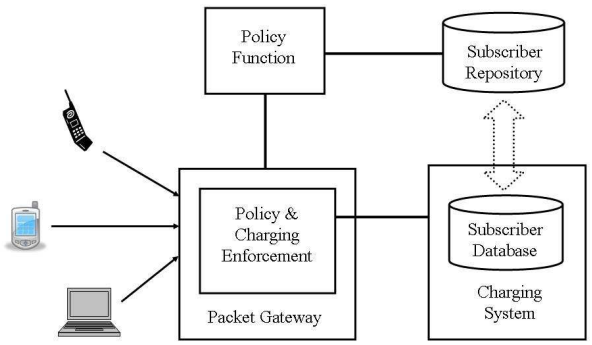

Figure 3. Data traffic from various access networks

In the case of GPRS or UMTS networks, the packet gateway might be a GPRS Gateway Support Node (GGSN); while in the case of an LTE network, it will be a Packet Data Network Gateway (PDN-GW). Besides, the latter acts as the anchor point between 3GPP and non-3GPP technologies such as WiFi or Worldwide Interoperability for Microwave Access (WiMAX) [11].

The PDN-GW provides to the User Equipment (UE) connectivity to external packet data networks by being the point of exit and entry of traffic for the UE. A single UE may have simultaneous connectivity with more than one PDN-GW for accessing multiple PDNs. The PDN-GW performs packet filtering, lawful interception and packet screening. Especially, the PDN-GW performs policy and charging enforcement, based on instruction from the policy function on one side, and from the charging system on the other side. The reader who wishes to have more insight on the Policy & Charging Control (PCC) architecture should refer to [12].

Studying in more details the reference architecture for data core network, i.e., the IP Multimedia Sub-system (IMS) standard architecture, and focusing on online charging [13], a so-called Online Charging System (OCS) relies on two databases:

- The database in the Rating Function (RF), which contains generic tariff information at service level;
- The database in the Account Balance Management Function (ABMF), which contains subscriber-specific information relevant for rating purposes.

Actually, searching the literature, an interaction between the policy decision function and external databases is mentioned in [14], but it does not relate specifically to the database of an OCS. And the dynamic mid-session interaction is not studied in detail either. A direct interaction between a so-called Policy & Control Resource Function (PCRF) and an OCS has already been studied in [15], but it restricts to an interaction of the PCRF with the Rating or Tariff Function of the OCS. It means that the policy decision might indeed depend on generic tariff rules, but it still does not depend on subscriber-specific information such as his/her current consumption or life cycle state. Moreover, reducing the subscriber's tariff information to a single tariff class ID might be restrictive given newer tariff schemes, where multiple charging options might be applied separately on top of a default tariff. The reader, who wishes to have more information about newer tariff schemes, might refer to [5]. Such charging options are amongst others usage-based discounts, subscriber bonus or individual buckets, e.g., free minutes, that the subscriber can book in addition to his/her default tariff, or that he/she gets as a reward for high consumption or recharge.

Basically, one of the functions of the OCS is to perform account balance management towards external systems through the ABMF. For this purpose, the OCS might store subscriber's pieces of information applicable for rating like usage counters. Furthermore, it might store additional information like his/her life-cycle state, e.g., validity dates, or the status of his/her valid tariff options.

According to [13], in order to support the online rating process, the Rating Function necessitates counters. The counters are maintained by the Rating Function through the Account Balance Management Function. Assuming that these counters are maintained at subscriber level, storing them together with other real-time subscriber information in the ABMF makes sense.

According to [16], in order to support the policy decision process, the PCRF may receive information about total allowed usage per user from a subscribers' repository called Subscription Profile Repository (SPR). Going further in this direction, some additional subscriber information might be relevant to the PCRF in order to determine the right policy: not only static data like an allowed usage threshold specific to a subscriber, but also subscriber's dynamic data like the value of specific counters at a certain point in time, his/her life-cycle state, or the status of his/her valid tariff options.

Storing such data in the SPR would be necessary to support scenarios like the following: as long as the subscriber consumption within one month does not exceed a

certain limit, he/she is eligible for a better QoS than once the threshold has been exceeded. Alternatively, a scenario might occur, in which a specific subscriber buys on top of his/her standard tariff an option for data traffic, so that he/she is eligible for a better policy than "normal" subscribers.

Consequently, the SPR would have to store such information as well. However, this information is still mandatory in the subscribers' database of the charging system because it might influence ratings. For example, having subscribed to a certain data option might lead to a reduced or negligible price for data traffic. Or taking the example mentioned earlier, once the subscriber consumption within one month exceeds a certain limit (not necessarily the same limit as for policy decision, but possibly tracked by the same counter!), the subscriber might enjoy cheaper rates for data traffic.

This shows that some subscriber data is meaningful both for the subscribers' repository (SPR) and the subscribers' database of the Charging System (ABMF). There could be here a kind of overlapping between the SPR and the ABMF, as the dotted arrow in the right part of Figure 3 suggests.

Replicating the information both in the SPR and in the ABMF would be an option. But this would assume efficient synchronization mechanisms between the two databases, since the number of subscribers respectively their data traffic in today's telecommunication networks might be substantial. Furthermore, the involved pieces of information consist of real-time data. If the policy should change when the subscriber's consumption reaches a certain limit, the change should happen in real-time and without delay as we saw earlier. In the same way, if the rating should change when a certain limit is reached, the change should happen in real-time too.

Duplication of databases, which store a great deal of real-time data, could increase the complexity of the implementation. If the relevant subscriber information is already present in the OCS, why should not the PCRF retrieve it directly from the OCS? This is represented by the dotted arrow in Figure 4.

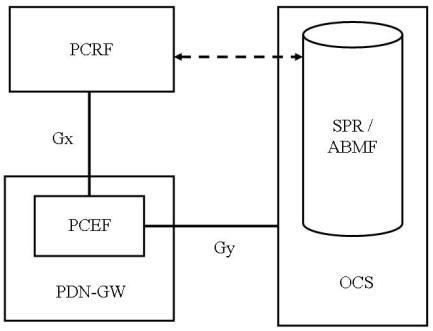

Figure 4. OCS acting as an SPR

VII. PROPOSED APPROACH

The proposed approach described for the first time in [1] consists of a framework where the PCRF and the OCS exchange in real-time subscriber information, which is necessary not only for charging, but also in order to determine the right policy. The goal is to support such scenarios where the policy might be changed in the middle of a session based on the value of some subscriber data volume counter.

The latter is stored in the OCS as master copy in any case because it is relevant for charging, in order to support offers like the following: after a subscriber has consumed 500MB within one week, he/she gets 10 free SMS, or he/she is eventually granted free-of-charge data traffic till the end of the week. Furthermore, these counters are relevant to the PCRF in order to support similar offers where, for example, the data speed is throttled once the subscriber has reached 1GB consumption within one month. In the context of the present contribution, we shall focus on volume counters. However, as mentioned earlier, it could be another piece of subscriber data, which would be relevant for the policy server, for example, the life-cycle state of the subscriber. For example, if a prepaid data card is near expiry, the surfing speed may diminish.

In the context of the implementation, which will be described in the next section, these are the values of subscriber volume counters, which should be reported in real-time from the OCS to the PCRF. More precisely, the counter values will be reported when they exceed some predefined threshold. The latter might be defined either for a certain subscribers' marketing category, or for all the subscribers in the same tariff, or individually at subscriber level. Since these thresholds might be reached in the middle of a session, the OCS might have to notify the PCRF in the middle of a data session too.

Nevertheless, the PCRF should retrieve latest subscriber information like the tariff plan information and the values of the volume counters at the beginning of the session as well, in order to determine correctly the initial policy. Alternatively, the PCRF could replicate this subscriber information, meaning again that some synchronization mechanism would have to be implemented.

In general, the message flow when a data session is established would resemble Figure 5.

In (1), the Policy & Control Enforcement Function (PCEF) asks the PCRF about the policy that should apply to the session, which is about to start for this subscriber. For this purpose, the PCRF retrieves latest subscriber information from the OCS in (2) and (3). Consequently, the PCRF can notify the initial policy to the PCEF in (4). This would happen through the Gx interface in accordance with [16].

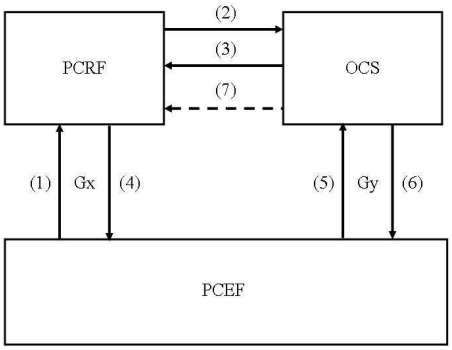

Figure 5. Message flow with PCRF/OCS interaction

Once the policy has been determined, the PCEF requests the OCS for a volume slice in (5). After checking the current subscriber's consumption, the subscriber's default tariff respectively his/her available options and current balance, the OCS allocates a slice in (6). This would happen through the Gy interface in accordance with [12]. In order to allocate the proper slice, the OCS takes into account charging-relevant thresholds, but it should take into account policy-relevant thresholds as well, because this will ensure a timely charging or policy change: as soon as the volume quota leading to the threshold will be consumed, the OCS is able to notify the PCRF. Depending on the duration of the session, there might be several volume slices requested, i.e., several messages like (5) and (6).

The arrow in (7) is represented in a dotted line because it may or may not occur during a session: the OCS would notify the PCRF only in the case that a policy-relevant threshold is exceeded during the on-going data session.

As stated above, the protocol for (1) & (4) respectively (5) & (6) is Gx respectively Gy. The protocol for (2) & (3) respectively (7) will be discussed in the next section. Since (2) & (3) respectively (7) were not fully covered by standard bodies at the time of the implementation, the most convenient protocol had to be assessed.

VIII. IMPLEMENTATION

Regarding the protocol for (7) in Figure 5, since Gx and Gy rely on Diameter [17], and Gy on Diameter Credit Control Application [18], it was decided to use Diameter Credit Control Request (CCR) Event. The reader might have noted that in (5) & (6), the OCS acts as a Diameter Server towards its client, i.e., the PCEF, while in (7) the OCS acts as a Diameter Client toward the Diameter Server, which is the PCRF in this case. As there might be several PCRF nodes, the OCS should support an N+K PCRF architecture in order to ensure a good scalability. The OCS should be able to send CCR Event messages to the PCRF nodes in round-robin way in order to ensure high-availability, meaning that the functionality can still be supported, even if one PCRF node is down.

Regarding (2) and (3), it is about the PCRF's retrieving subscriber profile data from the OCS database at the beginning of a session. Therefore, it is not really about Credit Control, nor Authentication/Accounting. Consequently, Diameter was not chosen, but Simple Object Access Protocol/eXtended Markup Language (SOAP/XML) instead, because it is a simple protocol to let applications exchange information over HyperText Transfer Protocol (HTTP) [19] in a platform-independent manner. For more information on SOAP/XML, the reader might refer to [20] and [21].

Within this framework, the following scenario has been implemented: let us assume that a subscriber is entitled a downlink/uplink speed of 768/384 Kilo bit per second (Kbps) as long as he/she has not exceeded 10MB within a month. Once he/she reaches 10MB, he/she should be throttled to 128/64 Kbps. Let us assume that at the beginning of a session, the subscriber has a consumption of 9.9MB in the current month.

Consequently, when the session is established, the PCRF communicates a QoS corresponding to 768/384 Kbps to the PCEF. In addition, the OCS allocates a quota of only 0.1MB (10-9.9) in the initial Credit Control Answer (CCA) message. That way, when the threshold of 10MB is reached, the PCRF can be notified in real-time. This is represented in Figure 6.

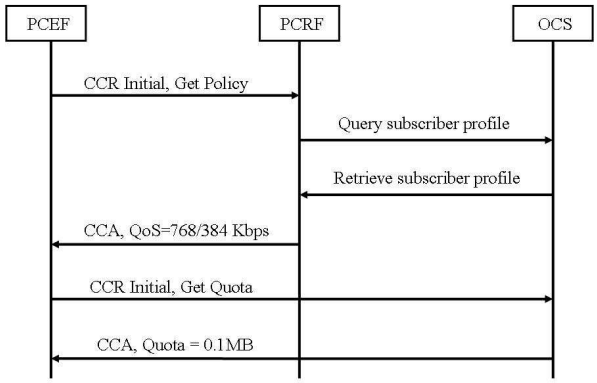

Figure 6. Initial slice granted by the OCS at session start

In case the PCRF has a local database duplicating the OCS database, and containing subscriber information that is not outdated, the query of the subscriber profile from the PCRF to the OCS may be skipped.

When the allocated quota of 0.1MB has been used up, the PCEF should request another volume quota. If the subscriber balance is sufficient, the OCS will allocate another quota so that the data session can carry on. The allocated quota might be bigger than 0.1MB this time, for example 0.5MB. Simultaneously, the OCS will notify through a Diameter CCR Event message as indicated previously that the volume threshold of 10MB has been reached for this subscriber, so that the PCRF can deduce the new QoS and notify it to the PCEF. This is represented in Figure 7.

In order to further notify the policy's change to the PCEF, the PCRF uses Diameter Re-Authentication Request / Answer messages (RAR/RAA) in accordance with [22].

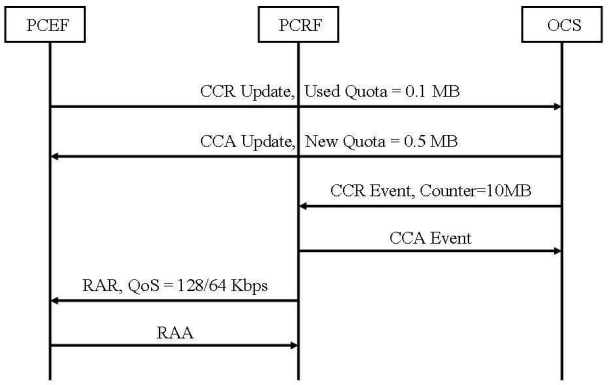

Figure 7. Mid-session notification from OCS to PCRF

In case of multiple parallel sessions, the policy change should apply to all on-going sessions. For example, let us assume that one session – Session 1 – starts when the counter value is 9.9MB. Given the threshold of 10MB, the OCS should allocate initially a slice of 0.1MB. Before the latter is used up, another session – Session 2 – starts. The OCS also allocates 0.1MB as initial slice because the counter value is still 9.9MB in the OCS database. This is represented in Figure 8.

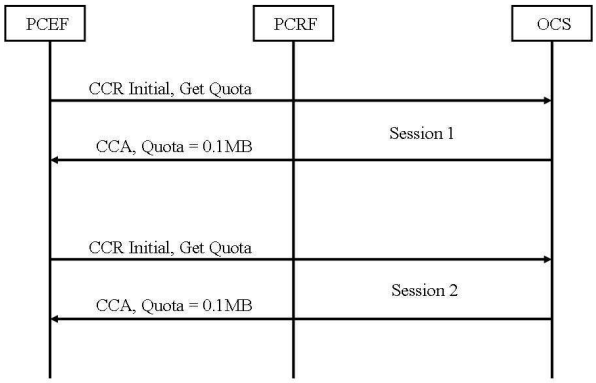

Figure 8. Initial slice for parallel sessions

As soon as the initial slice of 0.1MB of Session 1 or Session 2 is used up, the PCEF will request another slice. The OCS will grant a new slice, but it will update the volume counter value to 10MB, which should trigger the notification to the PCRF. This is represented in Figure 9, where the first session using up the 0.1MB quota is Session 1.

Consequently, the PCRF should notify the PCEF to change the QoS obviously for Session 1, but for Session 2 as well, because the volume threshold is applicable to both Session 1 and Session 2, even if the QoS change was triggered by Session 1 only.

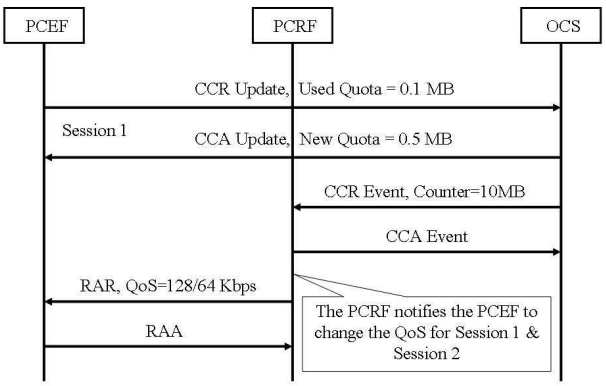

Figure 9. Mid-session notifications for parallel sessions

IX. ALTERNATIVE SOLUTIONS

We understand why PCRF and OCS should interact with each other, and we proposed a framework where they can exchange messages directly. However, interaction does not necessarily mean a direct interface between both components. The existing interfaces Gx [16] and Gy [12] could be extended to support this interaction. This is illustrated in Figure 10 (for one single session, not for two parallel sessions).

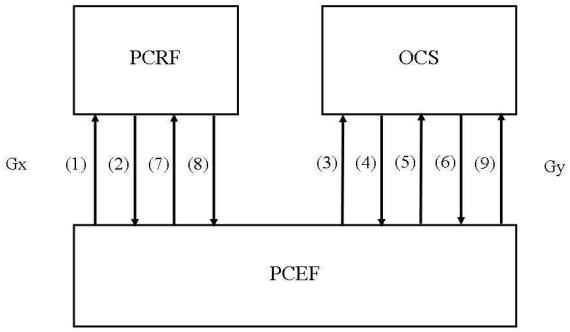

Figure 10. Flow without PCRF/OCS direct interface

The two different levels of QoS could be mapped to two different charging keys or *Rating Groups*. Assuming that the limit has not been reached yet for the subscriber in the current period, the PCRF would apply initially the first rating group together with the high QoS to the session being established in (1) & (2). In the Credit Control Request (CCR) Initial message in (3), the OCS would receive the first rating group and grant in (4) a volume quota equal or lower than the volume delta till the limit.

When the quota is used up, the PCEF notifies the OCS in a CCR Update message in (5). The Credit Control Answer (CCA) message from the OCS in (6) could indicate graceful service termination, and the Final-Unit-Action would be set to "Redirect". This way, the PCEF could forward in (7) the service termination message back to the PCRF, which could

react in (8) by returning the second charging key or Rating Group, in addition to the normal QoS information, instead of the high QoS. Upon receipt of this new Rating Group in (9), the OCS would continue granting credit to the service. Consequently, the session could continue, but not with the same QoS.

However, this dummy service termination and redirect action would have implied an extension of the existing protocols Gx and Gy, in order to support the exchange of limit-reached information from the OCS to the PCRF through Diameter Redirect.

Another alternative solution would have been to have one single system for the PCRF and OCS, as represented in Figure 11.

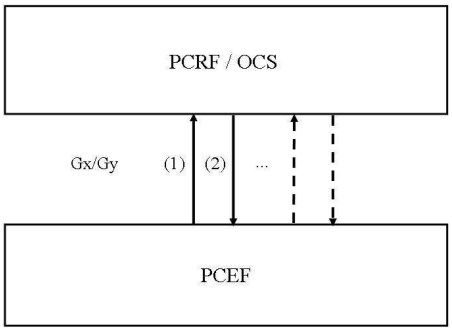

Figure 11. Integrated Policy & Charging function

As can be seen on the picture, the message flow would look quite simple. There would be successive messages like in (1) & (2), over a combined Gx and Gy interface, controlling both the policy rule to be applied and the volume quota to be granted. Such a solution may reduce the traffic between the PCEF and the integrated PCRF/OCS, and may reduce hardware and maintenance costs as well.

However, service providers, even if they are not managing policies yet on an individual basis, are already billing subscribers individually, thus they have a legacy charging system. It might be less risky to introduce a new policy function as a separate project than replacing completely the legacy charging system on top of adding a new policy control resource function.

Furthermore, keeping two separate systems for two different purposes can bring more flexibility in designing evolved policy and charging rules. Finally, it is also easier in terms of hardware and software upgrade, regarding maintenance windows, downtime, etc., which should be as short as possible, at least for a real-time charging system. A service provider might grant a high QoS temporarily for free to the complete subscriber base during a maintenance window in the night at low traffic hours, but if it grants free calls to everyone, the revenue impact is immediate, even if the number of calls is not huge.

Actually, it is a bit like having a TV and DVD player integrated in the same device. Some people may like it, but if the device is down, it means that both systems are down.

X. STANDARDIZATION

The time, when the prototype described in the 'Implementation' section was designed, goes back to the beginning of the year 2009. At this time, there was no standard regarding PCRF – OCS interaction.

Actually, IMS Release 7 introduced the concept of integrated Policy Charging & Control (PCC) architecture, with separate components for the policy function and the charging system. However, OCS – PCRF interaction was not covered at that point. Some discussion started during the course of 2009 in the context of 3GPP about "QoS and gating control based on spending limits".

A document issued end of 2009 discussed various options [23]. For the first time, a direct interface between the PCRF and the OCS was named: the 'Sy' interface. However, no conclusion was drawn at that time about which alternative would become the recommended solution.

Since then, a new version of the document [24] has been issued mid 2011, and the recommendation is now the following: "The Sy based solution where PCRF initiates Sy interaction shall be used". Main reason is that "it has the advantage of causing no increase in signaling load at the PCEF". Finally, a new technical specification dedicated to this Sy reference point was issued end of 2011 entitled "Spending Limit Reporting over Sy reference point" [25].

As the name of the specification suggests, this interface currently focuses on the exchange of counter information related to spending limits. It may be worth extending this interface in the future, in order to be able to exchange other subscriber's pieces of information, which might affect policy decisions too, like the subscriber's life cycle state, or his/her tariff options as we mentioned previously.

XI. NOTIFICATION FRAMEWORK

In the previous sections, we focused on a framework with the aim of real-time mid-session policy and charging control. In the context of the European regulation for roaming data traffic, we mentioned the possibility to notify the end-user when a certain threshold, or alternatively when a percentage of this threshold, is reached. The notification text would then explain when and why the policy is going to be changed. The possible message flow for an SMS notification is described in Figure 12.

When a session is initiated by the end-user in (1), the PDN-GW requests first from the PCRF in (2) & (3) information about the policy to be applied, and then it requests from the OCS in (4) & (5) information about the charging scheme to be applied. If it is just about notification, and not about actual policy change, the OCS can notify instead of the PCRF an SMS Center (SMS-C) like in (6), in order that the latter sends in (7) a notification to the end-user, e.g., "at this point in time, you have consumed 40€ in roaming data traffic, you are approaching the limit of 50€". The OCS provides all the information regarding the

subscriber identification and the text message that the SMS-C may need.

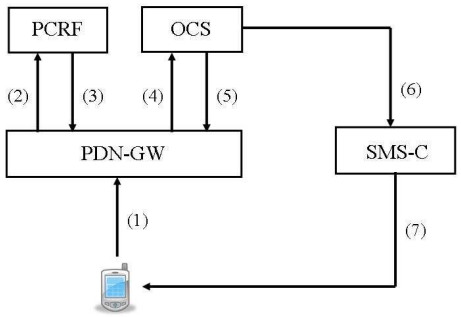

Figure 12. SMS notification in the context of Policy control

Of course, if the end-user is in the middle of a data session on his/her mobile phone, it might not be so convenient to read an SMS that is just incoming. Consequently, the user might instead be redirected to a landing page displaying notification contents. This scenario is described in Figure 13.

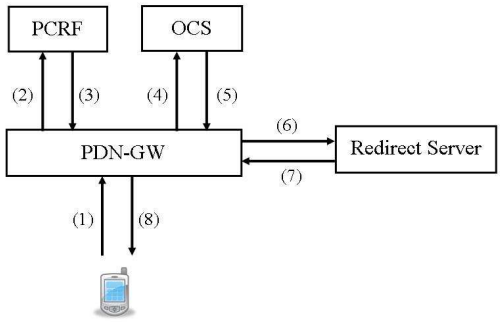

Figure 13. Redirection in the context of Policy Control

In comparison with the SMS notification, the OCS does not notify the SMS-C, but instead it sends back to the PDN-GW a Diameter Re-direct message in (5), providing the address of a Redirect Server, that the PDN-GW is able to contact in (6), in order to retrieve the landing page contents in (7), and send it back to the end-user in (8). The landing page might contain some links to invite the end-user to upgrade his price plan or his/her QoS, or to opt for a new attracting offer dedicated to data services.

Such a dialogue is definitely more user-friendly than a brutal QoS change, especially in the case of throttling. Before any policy change, and even if the ability to modify the QoS in real-time has been mutually agreed in advance between the service provider and the end-customer according to contractual terms, it might enhance the end-customer's experience to have a kind of interactive dialogue within a Web-based application.

Taking the example of throttled traffic, instead of just throttling the traffic, it might be more user-friendly to put temporarily the session on hold, and to trigger simultaneously via push mechanism the display of a pop-up window on the end-user's terminal, in order to ask him/her whether he/she agrees with additional expense or with booking a new data option. Such an interaction is illustrated in Figure 14.

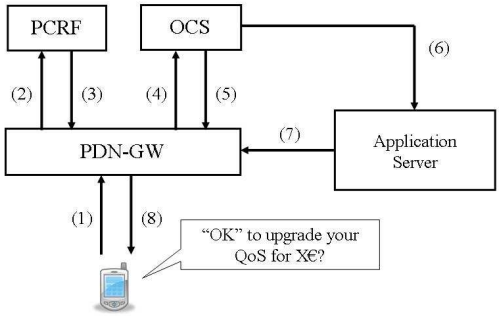

Figure 14. Pop-up window in the context of Policy Control

An application could be triggered by an Application Server (AS) on the end-user's mobile equipment. To make best use of the terminal specificities, it makes sense to have a dialogue between the AS and a specific client application running on the terminal; meaning that the AS might have to identify the type of end-user's device first, and subsequently trigger a corresponding notification interactive application on the end-user's terminal. The application on a Blackberry and on an iPhone might be different.

This application could conduct an intelligent dialogue with the end-user, asking him/her whether he/she agrees to upgrade his price and QoS plan, or accept a trial offer. Such an offer could depend on the session's context like URL. For example, if the subscriber tries to download a video from a certain video portal, he/she can be proposed a special bundle combining volume and bandwidth applicable to the video portal that the subscriber is just visiting.

In Figure 14, a user starts for example a video download in (1), and gets the standard QoS in (2) & (3). After the video download content type has been identified and authorized by the OCS in (4) & (5), the OCS notifies in parallel an AS in (6), which triggers a pop-up window on the end-user's terminal application in (7) and (8).

If the end-user answers positively, e.g., to a QoS upgrade against a certain fee of X€, eventually combined to a new "free" volume bucket, a message flow like the one represented in Figure 15 might occur. This would make an example of "in-application" smart charging.

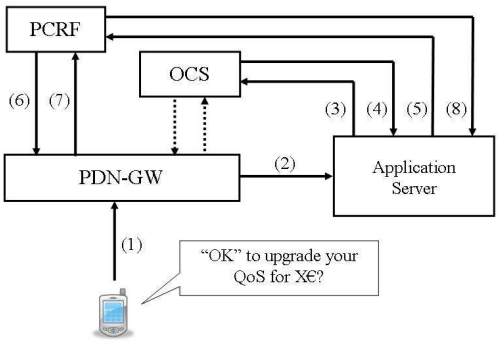

Figure 15. Application Server notifying directly a policy change

The positive answer is forwarded from the end-user's terminal to the AS in (1) & (2). Then the AS sends a debit request of X€ to the OCS in (3), and if the user's account has enough credit, the OCS will answer positively in (4). Consequently, the AS can trigger the QoS upgrade toward the PCRF in (5). The PCRF will enforce a new policy toward the PDN-GW in (6), and once the change is acknowledged by the PDN-GW in (7), the PCRF can notify the AS in (8). Since the balance of the subscriber's account has decreased, and the data tariff might have changed too, the OCS might grant a different volume quota than initially. This is represented in the dotted arrows in Figure 15. (5) and (8) would be implemented using the Rx interface in order to comply with [26].

Alternatively, the subscription for X€ to the QoS tariff option in (3) & (4) could lead the OCS's notifying itself the PCRF about the QoS change. This is represented in Figure 16.

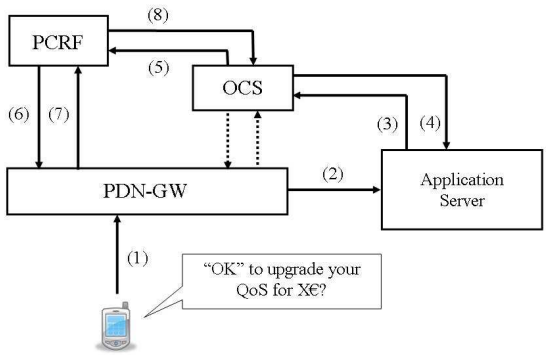

Figure 16. Application Server triggering a policy change through the OCS

The difference with Figure 15 is that the OCS is making use in (5) of the newly standardized Sy interface [25], in order to notify itself the PCRF.

XII. CONCLUSION

We studied in this paper several frameworks enabling better correlation between charging and policy control in today's telecommunication networks. This becomes a must given the tremendous increase of data traffic.

In this context, service providers face multiple challenges: the challenge to find the right balance between flexibility and complexity when proposing new tariff offers, which should combine competitive price and sufficient bandwidth for the end-user; the challenge to launch modern services consuming more network resources with the necessity to generate revenue from these services according to resource consumption; the challenge to empower new as well as existing customers to use innovative applications still preserving the overall end-user's quality of experience for the whole subscriber base; the challenge to design differentiated solutions for policy control, taking into account both standard architectures and the diversity of end-user terminals, especially smart phones and tablets.

TERMINOLOGY

3GPP	3rd Generation Partnership Project
ABMF	Account & Balance Management Function
AS	Application Server
CCA	Credit Control Answer
CCR	Credit Control Request
CDR	Call Detail Record
DVD	Digital Versatile Disc
DWH	Data Warehouse
GB	Giga Byte
GW	Gateway
GGSN	GPRS Gateway Support Node
GPRS	General Packet Radio Service
GW	Gateway
Gx	IMS reference point between PCEF & PCRF
Gy	IMS reference point between PCEF & OCS
HTTP	Hyper Text Transfer Protocol
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IVR	Interactive Voice Recognition
KB	Kilo Byte
Kbps	Kilo bit per second
LTE	Long Term Evolution
MB	Mega Byte
NN	Network Neutrality
OCS	Online Charging System
PCC	Policy & Charging Control
PCEF	Policy & Control Enforcement Function
PCRF	Policy & Control Resource Function
PDN	Packet Data Network
PSTN	Public Switched Telephony Network
QoE	Quality of Experience
QoS	Quality of Service
RAA	Re-Authentication Answer
RAR	Re-Authentication Request
RF	Rating Function
Rx	IMS reference point between AS & PCRF
SMS	Short Message Service
SMS-C	SMS Center
SOAP	Simple Object Access Protocol
SPR	Subscription Profile Repository

Sy IMS reference point between OCS & PCRF
TV Television
UE User Equipment
UMTS Universal Mobile Telecommunications System
WiFi Wireless Fidelity
WiMAX Worldwide Interoperability of Microwave Access
XML eXtended Markup Language

ACKNOWLEDGMENT

The author would like to thank Hongwei Li, Renée Fang, Jessica Han, Angelo Lattuada, Justin Bayley, Andy Wood, and Mark Bryant from Alcatel-Lucent, Michael Leahy from Openet, Dion Pirnaij, Guido Gillissen, and Ton van Boheemen from Vodafone, Steven Cotton from the TM Forum, Val Korolev from OpenCloud.

REFERENCES

[1] M. Cheboldaeff, "Interaction between an Online Charging System and a Policy Server", Tenth International Conference on Network (ICN), January 23-28, 2011. France. IARIA.

[2] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "General Packet Radio Service (GPRS); Service description", 3GPP TS TS 23.060, Release 10, June 2011.

[3] Y.-B. Lin, A.-C. Pang, Y.-R. Haung, I. Chlamtac, "An All-IP Approach for UMTS Third-Generation Mobile Networks", IEEE Network Magazine, September/October 2002

[4] 3rd Generation Partnership Project, 36 series, "LTE (Evolved UTRA) and LTE-Advanced radio technolgy", http://www.3gpp.org/ftp/Specs/html-info/36-series.htm [retrieved: January 3rd, 2012]

[5] M. Cheboldaeff, "Service Charging Challenges in Converged Networks", IEEE Communications Magazine, January 2011

[6] H. Zhou, K. Sparks, N. Gopalakrishnan, P. Monogioudis, F. Dominique, P. Busschbach, and J. Seymour, "Deprioritization of Heavy Users in Wireless Networks", IEEE Communications Magazine, October 2011

[7] Europe's information society, "Countering data roaming bill shocks", http://ec.europa.eu/information_society/activities/roaming/regulation/index_en.htm [retrieved: January 3rd, 2012]

[8] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "Charging management; Charging data description for the IP Multimedia Subsystem (IMS)", 3GPP TS TS 32.225, Release 5, March 2006.

[9] R. Stankiewicz, P. Cholda, and A. Jajszczyk, "QoX: What is It Really?", IEEE Communications Magazine, April 2011

[10] Institute of Electrical and Electronics Engineers (IEEE), Information technology, Part 11: Wireless LAN Medium Access Control and Physical Layer Specifications, IEEE Std 802.11

[11] Institute of Electrical and Electronics Engineers (IEEE), Information technology, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, IEEE Std 802.16

[12] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "Policy and Charging Control (PCC) architecture", 3GPP TS 23.203 v11.2.0, Release 11, June 2011.

[13] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "Online Charging System (OCS): Applications and Interfaces", 3GPP TS 32.296 v11.0.0, Release 11, June 2011

[14] R. Good, and N. Ventura, "Application driven Policy Based Resource Management for IP multimedia subsystems", 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities (TridentCom), 2009

[15] T. Grgic, K. Ivesic, M. Grbac, and M. Matijasevic, "Policy-based Charging in IMS for Multimedia Services with Negotiable QoS Requirements", 10th International Conference on Telecommunications (ConTEL), 2009

[16] 3rd Generation Partnership Project, Technical Specification Group Core Network and Terminals, "Policy and Charging over Gx/Sd Reference Point", 3GPP TS 29.212 v11.1.0, Release 11, June 2011

[17] P. Calhoun, J. Loughney, E. Guttman, G. Zorn, and J. Arkko, "Diameter Base Protocol", IETF RFC 3588

[18] H. Hakala, L. Mattila, J-P. Koskinen, M. Stura, and J. Loughney, "Diameter Credit Control Application", IETF RFC 4006.

[19] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", IETF RFC 2616

[20] SOAP Version 1.2, World Wide Web Consortium (W3C) Recommendation, http://www.w3.org/TR/soap12-part1/ [retrieved: January 3rd, 2012]

[21] World Wide Web Consortium (W3C), http://www.w3.org/standards/xml/ [retrieved: January 3rd, 2012]

[22] P. Calhoun, G. Zorn, D. Spence, and D. Mitton, "Diameter Network Access Server Application", IETF RFC 4005

[23] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "Study on Policy solutions and enhancements", 3GPP TR 23.813 v0.1.0, Release 10, November 2009

[24] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "Study on Policy solutions and enhancements", 3GPP TR 23.813 v11.0.0, Release 11, June 2011

[25] 3rd Generation Partnership Project, Technical Specification Group Core Network and Terminals, "Policy and Charging Control: Spending Limit Reporting over Sy reference point", 3GPP TS 29.219 v1.0.0, Release 11, November 2011

[26] 3rd Generation Partnership Project, Technical Specification Group Core Network and Terminals, "Policy and Charging Control over Rx reference point", 3GPP TS 29.214 v11.1.0, Release 11, June 2011