Pricing Differentiated Services in the GPRS Environment

Sergios Soursos
Dept. of Informatics
Athens University of
Economics and Business
Athens 10434, Greece
sns@aueb.gr

Costas Courcoubetis
Dept. of Informatics
Athens University of
Economics and Business
Athens 10434, Greece
courcou@aueb.gr

George Polyzos
Dept. of Informatics
Athens University of
Economics and Business
Athens 10434, Greece
polyzos@aueb.gr

ABSTRACT

The General Packet Radio Service extends the existing GSM mobile communications technology by providing packet switching and higher data rates in order to efficiently access IP-based services in the Internet. Since no realization path for Quality-of-Service support has been proposed yet, we adapt the Differentiated Services framework and apply it over the GPRS air interface in order to provide various levels of service differentiation. We also focus on applying a charging technique so as to publish a unit price for each service class. These prices are designed to lead to the maximization of Social Welfare and the users' net benefit.

Categories and Subject Descriptors

C.2 [**Computer Systems Organization**]: Computer Communication Networks; C.2.1 [**Computer Communication Networks**]: Network Architecture and Design—*Wireless communication*; C.2.m [**Computer Communication Networks**]: Miscellaneous—*Network Economics*

General Terms

Congestion Pricing of Differentiated Services

Keywords

GPRS, Differentiated Sevices, Two-bit Differentiation, congestion pricing, tatonnement

1. INTRODUCTION

The convergence of mobile technologies with the technologies of the Internet was of great importance this last decade. One step towards this direction was made by the introduction of the General Packet Radio Service (GPRS) over the Global System for Mobile communications (GSM). GPRS is a packet-switched service offered as an extension of GSM. In contrast to the classic circuit-switched service provided by GSM, GPRS offers the efficiency of packet-switching desirable for bursty traffic, higher transfer speeds than the ones available today to a single end-terminal (theoretically up to 115 kbps) and instantaneous connectivity with any IP-based external packet network.

One important issue in this context is the Quality-of-Service (QoS) provided by GPRS. Even though GPRS specifications define QoS parameters and profiles, we are unaware of specific implementation plans and strategies in order to support specific QoS models, particularly over the wireless access network. Recent proposals in the area of GPRS QoS focus on providing QoS support in the core GPRS network (which is typically non-wireless and IP based) using the standard Internet QoS frameworks (i.e., Integrated Services or Differentiated Services) [11]. We believe that the critical part for the support of QoS to the applications and the end users is the access network where, because of the scarcity of the radio spectrum, greater congestion problems can result. Therefore, we have developed an architecture that provides QoS in the form of support for Differentiated Services over the radio link and integration with the Internet DiffServ architecture, thus providing end-to-end QoS "guarantees" [13].

Another important issue that derives from the development of a Differentiated Services architecture, is the charging for providing such services to the end users. The lack of pricing mechanisms may result in over-utilization of the network resources, leading to a degradation of the network's performance. Users must be given the right incentives to choose the service that is the most appropriate to satisfy their needs (in QoS levels). Pricing prevents users from getting tempted to request higher than needed services. Our charging method is closely related to the DiffServ model.

The structure of the remainder of this paper is as follows. First we provide a short overview of the GPRS technology and architecture. We then review briefly the Internet Differentiated Services architecture and we focus particularly on the description of the two-bit DiffServ scheme. In the following section we adapt the two-bit DiffServ scheme in the GPRS environment, describing all the new tasks that are required to be performed by the GPRS Serving Nodes (GSNs), the key new elements in the GSM architecture introduced to support GPRS. Next, we use congestion pricing techniques to determine the unit price for each service class and we prove that these prices maximize the Social Welfare and the users' net benefit. Finally, we discuss some open issues and present our conclusions.

2. THE GPRS ENVIRONMENT

GPRS [1, 2] is a new service offered by the GSM network. In order for the operators to be able to offer such services two new types of nodes must be added to the existing GSM architecture. These two nodes are the serving GPRS support node (SGSN) and the gateway GPRS support node (GGSN), as shown in Figure 1. The SGSN keeps track of the location of mobile users, along with other information concerning the subscriber and its mobile equipment, so as to accomplish tasks, such as packet routing and switching, session management, logical link management, mobility management, ciphering, authentication and charging functions. The GGSN connects the GPRS core network to one or more external Packet Data Networks (PDNs). Among its tasks are protocol translation, session management, assigning correct SGSNs to the Mobile Stations (MS) and collecting information for charging.

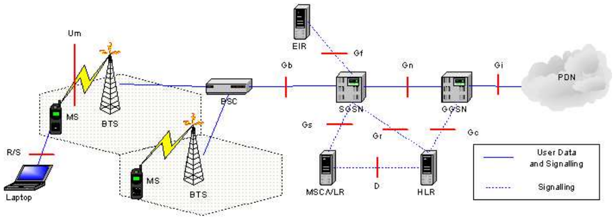

Figure 1: The GPRS network

The core GPRS network is IP based. At the radio link, the existing GSM structure is used, making it easier for operators to offer GRPS services. The uplink and downlink bands are divided through FDMA into 124 frequency carriers each. Each frequency is further divided through TDMA into eight timeslots, which form a TDMA frame. Each timeslot lasts 576.9 μs and is able to transfer 156.25 bits (both data and control). The recurrence of one particular timeslot defines a Packet Data Channel. Depending on the type of data transferred, a variety of logical channels are defined, which carry either data traffic or traffic for channel control, transmission control or other signaling purposes.

The major difference between GPRS and GSM concerning the radio interface is the way radio resources are allocated. In GPRS the radio channels, i.e. the timeslots, are allocated on a demand basis, in contrast to GSM where one timeslot is reserved for the entire duration of the call, even if there is no activity on the channel. This means that when a MS is not using a timeslot that has been allocated to it in the past, this timeslot can be re-allocated to another MS, so that no waste of radio resources is observed in the case of bursty traffic. The minimum allocation unit is a radio block, i.e. four timeslots in four consecutive TDMA frames. One RLC/MAC packet can be transferred in a radio block. During the transfer, the Base Station Subsystem (BSS) may decrease (or increase in some cases) the number of timeslots assigned to that particular MS, depending on the current demand for timeslots. This is accomplished by the use of flags (Uplink State Flag) and counters (Countdown Value) in the headers of the packets transferred on the radio link.

In order to make an exchange of data with external networks, a session must be established between the MS and the appropriate GGSN. This session is called Packet Data Protocol (PDP) context [4] and concerns the end-to-end path in the GPRS environment (MS $\leftrightarrow$ GGSN). During the activation of such a context, an IP address is assigned to the MS and is mapped to its IMSI and a path from the MS to the GGSN is built. The MS is now visible from the external network and is ready to send or receive packets. At the (lower) radio link level, when the MS starts receiving/sending data, a Temporary Block Flow (TBF) [5] is created. During this flow a MS can receive and send radio blocks uninterrupted. For a TBF establishment, the MS requests radio resources and the network replies indicating the timeslots available to the MS for data transfer. A TBF may be terminated even if the session has not ended yet since it depends on the demand for radio resources and the congestion of the link. After the termination, the MS must re-establish a new TBF to continue its data transfer.

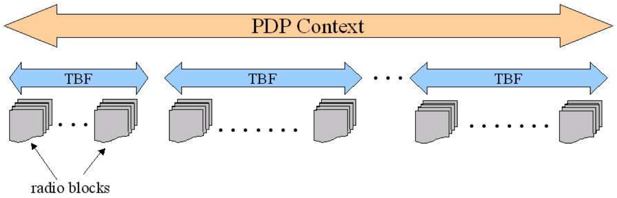

Figure 2: PDP Context and TBF

ETSI has also specified a set of QoS parameters and the corresponding profiles that a user can choose. These parameters are precedence (high, medium an low), reliability, delay, and peak and mean throughput [3] and thy form the user's profile which is stored in the HLR. Upon activation of a PDP context the mobile station is responsible for the required uplink traffic shaping. On the downlink, the GGSN is responsible to perform traffic shaping. It is obvious that such an implementation does not examine whether a user conforms to the agreed profile and the resource allocation procedures do not take into consideration the QoS profiles. Thus, it is up to the GPRS operator to use techniques that provide QoS "guarantees" and to police user traffic.

A first step in this direction is to use only the precedence parameter to define QoS classes and link allocation techniques. Precedence was chosen because of its simplicity and effectiveness and because it can be directly implemented in the GPRS architecture, as we will see in the following sections. Also, precedence can introduce very easily the idea of Differentiates Services, which is the preferred (realistic) approach for QoS in the Internet, gaining wide acceptance.

3. DIFFERENTIATED SERVICES

Multimedia and business applications have increased the volume of data travelling across the Internet, causing congestion and degradation of service quality. An important issue of practical and theoretical value is the efficient provision of appropriate QoS support.

The Differentiated Services architecture [7] was designed to address the scalability problems observed when applying the Integrated Services framework [6] on wide (inter-)networks, like the Internet. DiffServ's solution provides QoS support on aggregate flows. In a DS domain, the service provider and its users maintain contracts, the Service Level Agreement (SLA). The SLAs characterize the user's flow passing through the DS domain and include it in an aggregate of flows. They define the behavior of the domain's nodes to specific types of flow, i.e. the Per-Hop Behavior (PHB). The SLAs are also arranged between adjacent DS domains, so as to specify how flows directed from one domain to another will be treated. The DS field in an IP packet that defines the PHB uses reserved bits in the IP header - the "Type Of Service" field in IPv4 and the "Traffic Class" field in IPv6.

In the DS architecture, the first-hop router is the only DS node that handles individual flows. It has the task to check whether a flow originated from a user conforms to the contract that this user has signed and to shape it, if found to be out of bounds. This is achieved by using traffic conditioners. The internal routers handle aggregates of flows and treat them according to the PHB that characterizes them. The border router checks whether the incoming (or outgoing) flows conform to the contract that has been agreed to between the neighbor DS domains. All the traffic that exceeds the conditions of the contract is (typically) discarded.

Currently, there are no standardized PHBs, but two of the basic PHBs are widely accepted. These are the Premium (or Expedited) Service [9] and the Assured Service [8]. In Premium Service, the key idea is that the ISP provides the user with a constant available throughput, like the ATM CBR service . The exceeding packets are discarded while the remaining ones are forwarded to the next node. The Assured Service does not provide any strict guarantees to the users. It defines four independent classes. Within each class, packets are tagged with one of three different levels of drop precedence. So, whether a packet will be forwarded or not depends on the resources assigned to the class it belongs, the congestion level of that class and the drop precedence with which it is tagged. In other words, Assured Service provides a high probability that the ISP will transfer the high-priority-tagged packets reliably. Exceeding packets are not discarded, but they are transmitted with a lower priority (higher drop precedence).

The benefits from the deployment of both Premium and Assured services in a single DS domain are many. So nowadays, the Differentiated Services architecture is known as the combination of these two services and is called Two-bit Differentiated Service [10] . Each packet is tagged with the appropriate bit (A-bit and P-bit, with null for best-effort). The exceeding packets that belong to a Premium flow are dropped or delayed, while exceeding packets of Assured Service are forwarded as best effort.

In the first hop router (see Figure 3), packets that are tagged by users are checked for their conformity with the agreed SLA. In the case of Premium Service, all packets tagged with the P-bit are forwarded with a constant rate, defined by the traffic conditioner . In the case of Assured Service, the packets for which there is no token available are forwarded to the output queue as best-effort packets, with a null tag. At the border router the same basic tasks are performed, with a small variation. Since the border router manages and controls flow aggregates, it cannot buffer the packets that exceed the SLAs. Thus, the packets tagged with the P-bit are not queued, as in the first hop router, but they are discarded.

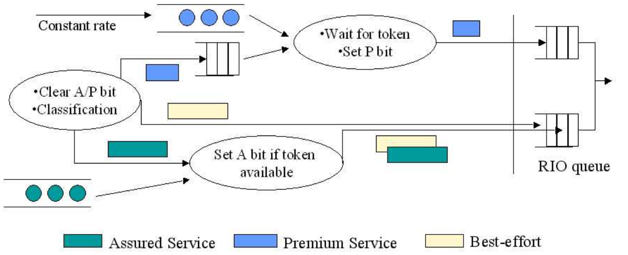

Figure 3: First Hop Router

4. DIFFSERV OVER THE GPRS AIR INTERFACE

In this section, we apply the Differentiated Services framework to the existing GPRS architecture. Specifically, we will see how the two-bit DiffServ architecture fits in GPRS, what changes must be made, and how it will be implemented.

We will give a simple example in order to make clear the reasons why we want to apply the Differentiated Services framework in the GPRS environment. Let us suppose that the GPRS network is attached to an external IP data network that uses Differentiated Services to provide QoS. The MS sends its IP packets to the GGSN, over the air interface where they are fragmented into RLC/MAC packets (frames). When these packets arrive at the GGSN, they are reassembled to IP packets and they are forwarded to the external network. Each IP packet is tagged according to the service that the user wants to receive. Thus, the GGSN acts like the first hop router in the Internet context, since there is only one IP hop from the MS to the GGSN, and checks whether the user flow conforms to the existing SLA. The next task of the GGSN is to forward the packets to the external network, where its nodes behave towards the packets as specified by the tag. We can easily conclude that any mobile user can use the Differentiated Services, as long as the external PDN supports them, in order to specify the way these packets will be treated in the external network. However, it is obvious that with the present techniques, the mobile user cannot control the way these packets are treated within the GPRS network. Our purpose is to design such a mechanism.

As described in the previous section, the two-bit DiffServ architecture involves two types of nodes in a DS domain: the first hop and the border router. In the case of our design for GPRS, we decided to have the GPRS network act as an independent DS domain (see Figure 4). As far as the border router is concerned, it is obvious that the GGSN is the most appropriate node for this task. It is the node that connects two DS domains. The GGSN monitors the incoming and outgoing flow aggregates in order to check their consistency with the SLAs between the two DS domains. Nonconforming traffic should be either discarded or degraded.

No special changes need to be made to the GGSN in order for it to act as a border router since it communicates via the IP protocol with both sides (both the SGSN and the border router of the neighbor domain).

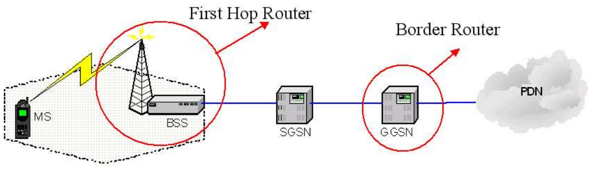

Figure 4: Two-Bit DiffServ into GPRS

When a PDP context is activated, the user can request a specific QoS level using the quality parameters mentioned earlier. In this case, the user sets the precedence parameter equal to one of the three available values. The highest priority makes use of the Premium Service, the medium priority of the Assured Service and the lowest priority of the best-effort service. This parameter is used to specify the behavior that the flow should receive in the GPRS core network, in the external network, if the later one uses Differentiated Services, and also the default radio priority used over the radio link.

As for the first hop router, this should be the BSS. Although its tasks will be the same with the ones described in Section 3, its structure will be totally different from the one depicted in Figure 3. This happens because of some differences in the architecture between an IP network and a GPRS network. Taking into account that the MSs send their data only when the BSS instructs them to and that they use the timeslot(s) defined by the USF field, we can assume that the traffic conditioner does not reside on the BSS, but it is distributed. The queues are realized in the MS (or in the notebook connected to the MS) and the tokens come from the BSS. Actually, the USF values are the tokens transferred over the radio link.

Another important difference in having the BSS as a first hop router is that within the BSS there is just an emulation of the system depicted in Figure 3, as described later in this section. Therefore, the BSS only needs a software upgrade in order to act as a first hop router, which makes it easier for implementation. No complex data structures are required. For queue implementation, linked lists can be used. Timers, counters and constants are all that is needed to realize the constant fill rate of the token pools and the thresholds of the RIO queues.

In the system described above, no packets do actually circulate, just requests for transfer. To be more precise, for each packet that the MS wants to transfer over the air, a pair (MS identity, service class) enters the above system. When the request exits the system then the BSS instructs the corresponding MS to transfer its packet by transmitting in a specified timeslot. The service class that a MS desires is declared with the use of the radio priority field at the TBF establishment request message. This field is two bits long, resulting into four values. We decided to have the following encoding: “1” for Premium Service, “2” for Assured Service and “3” for best-effort service. “0” specifies that the priority chosen at the PDP context activation will be used. The default value of the radio priority field is zero.

After the transmission of a packet (i.e., after four TDMA frames, since the packet is a radio block) the MS must make a new request to the BSS to transfer another packet. This makes clear that a TBF lasts for the transmission of only one radio block, after which the TBF is terminated and another one must be established to continue the transfer.

The architecture described above provides good results in both directions of the radio link. On the downlink, when data enter the GPRS network in order to reach a mobile user, the traffic is either characterized with, or translated to, one of the available service classes (Premium, Assured, best-effort). This is done at the GGSN. If the neighbor PDN does not support Differentiated Services, then the GGSN tags the incoming packets according to the profile of the user they are directed to. If, on the other hand, the neighbor PDN supports Differentiated Services, then the GGSN translates the incoming tags according to the SLA between the two DS domains.

On the uplink, the mobile user is able to tag his IP packets, activate a service class during PDP context activation or request a service class during the TBF establishment phase. The decision of which method to use depends on the user and on the network and is discussed later on this paper.

5. CHARGING DIFFERENTIATED SERVICES

The main objective of service differentiation, as discussed in previous sections, is to provide network users with a variety of services and let them decide which is the most suitable for their needs. It is obvious that a user who wants to participate in a video conference will choose a different service class than another user who wants simply to make an FTP connection and transfer some files. Thus, the service class is directly connected to the QoS requirements of the application used. However, there are no limitations in what service a user can choose. This means that the user who wants to transfer files may choose the class that was designed for video services. In that way the user takes advantage of the high transfer rates provided by this class. However, by doing so, the user increases the load of that class and the delay experienced by the users of that class. In other words, the QoS level of that class is degraded.

In the DiffServ framework, the available classes do not provide any strict guarantees on minimum performance levels. This fact leads users to demand the most they can from the network, in order to be sure that their requests will be served. Thus, one may expect that the higher priority class will be over-utilized, leading to a degradation of its performance level, an increase of the average delay and the congestion level and a misuse of network resources. The above example implies that users must have the right incentives to use the most suitable to them service class. To do so, the network operators must introduce some charging techniques in order to limit the uncontrolled use of their network resources. The charging techniques should be related to the QoS level that each service class offers. One indicator of this is the congestion level of each class. We should note that when a user enters a high priority class, he increases the congestion of that class and he decreases the network resources available to the lower classes. As we will see later in this section, the charging scheme must take into consideration these relations.

In the charging model that we propose (see Figure 5), we assume that we have three priority classes, which are related to the Differentiated Service classes described in Section 4, and n users ($i \in I, I = \{1, ..., n\}$). The highest priority class (Class 1) is dedicated to the Premium Service, the medium priority class (Class 2) to the Assured Service and the lower priority class (Class 3) to the best-effort traffic. Users' preferences are considered not to be able to have any effect on the prices or the delays. In other words, users take as granted the published prices and the experienced delays. We define as x_i^j the quantity (i.e. the number of packets) that user j sends to priority class i. The sum of flows in class i (i.e. the load of class i) is defined as $y_i = \sum_j x_i^j$. y_i also defines the demand for that class.

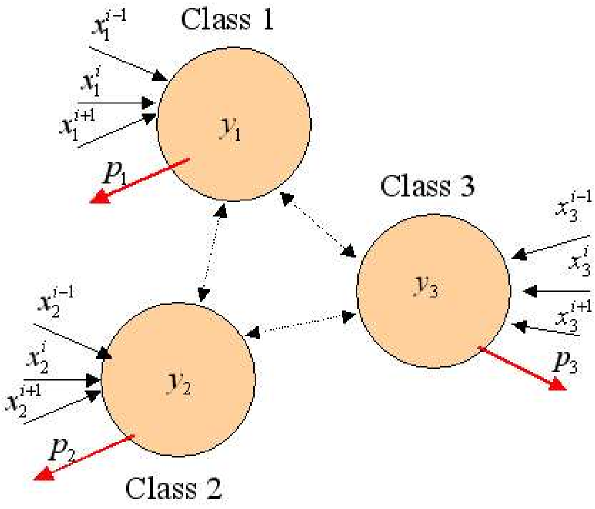

Figure 5: The charging model

Additionally, we define γ_i^j the delay cost experienced by user j for sending one unit of data in class i. We observe that for a single user the delay cost he experiences in different service classes is not the same. This is reasonable since the same delay for different services has different impact on the user's utility. To give an example, a delay of 2-3 seconds may not have any impact on a user that uses FTP, but it has an important impact on a user who uses teleconferencing software. In the same sense, the delay cost experienced on a packet using the higher priority class is greater than using a lower priority class.

Parameter d_i denotes the delay that one unit of data experiences in class i. As mentioned earlier, the delay in each class depends on the congestion level of her own and of the higher classes. Therefore, we use the definitions $d_1(y_1)$, $d_2(y_1, y_2)$ and $d_3(y_1, y_2, y_3)$ for expressing the relation of the congestion level with the delay experienced. The utility that a user i has from sending x units of data in the network is given by the function $u_i(x)$. From all the above, we can conclude that the total utility that a user i has from sending his data over the three priority classes, taking into consideration the delay in each class and the cost the user experiences, is

$$V_i(x_1^i, x_2^i, x_3^i) = u_i(x_1^i, x_2^i, x_3^i) - \gamma_1^i d_1(y_1) x_1^i - \gamma_2^i d_2(y_1, y_2) x_2^i - \gamma_3^i d_3(y_1, y_2, y_3) x_3^i$$

We assume that the network operator's objective is to maximize the Social Welfare. By maximizing Social Welfare we accomplish the maximization of the sum of total utilities of all the users that use the network services, i.e.

$$\max_{\{x_1^i, x_2^i, x_3^i\}} \sum_{i=1}^{n} V_i(x_1^i, x_2^i, x_3^i)$$

which equals to

$$\max_{\{x_1^i, x_2^i, x_3^i\}} \sum_{i=1}^{n} [u_i(x_1^i, x_2^i, x_3^i) - \gamma_1^i d_1(y_1) x_1^i - \gamma_2^i d_2(y_1, y_2) x_2^i - \gamma_3^i d_3(y_1, y_2, y_3) x_3^i] \tag{1}$$

At the optimal point, the sum of the users' total utilities is maximized, meaning that all the users are pleased with the QoS offered by the network.

The partial derivative of the above maximization function for x_1^i is

$$\frac{\partial u_i}{\partial x_1^i} - \gamma_1^i d_1(y_1) - \frac{\partial d_1(y_1)}{\partial y_1} \sum_{i=1}^{n} \gamma_1^i x_1^i - \frac{\partial d_2(y_1, y_2)}{\partial y_1} \sum_{i=1}^{n} \gamma_2^i x_2^i - \frac{\partial d_3(y_1, y_2, y_3)}{\partial y_1} \sum_{i=1}^{n} \gamma_3^i x_3^i = 0, \tag{2}$$

for x_2^i is

$$\frac{\partial u_i}{\partial x_2^i} - \gamma_2^i d_2(y_1, y_2) - \frac{\partial d_2(y_1, y_2)}{\partial y_2} \sum_{i=1}^{n} \gamma_2^i x_2^i - \frac{\partial d_3(y_1, y_2, y_3)}{\partial y_2} \sum_{i=1}^{n} \gamma_3^i x_3^i = 0 \tag{3}$$

and for x_3^i is

$$\frac{\partial u_i}{\partial x_3^i} - \gamma_3^i d_3(y_1, y_2, y_3) - \frac{\partial d_3(y_1, y_2, y_3)}{\partial y_3} \sum_{i=1}^{n} \gamma_3^i x_3^i = 0 \tag{4}$$

The above system of equations provides the socially optimal demands $\{x_1^{i*}, x_2^{i*}, x_3^{i*}\}$. The prices for each priority class are given below. For the first class, the unit price is

$$p_1 = \frac{\partial d_1(y_1)}{\partial y_1} \sum_{i=1}^{n} \gamma_1^i x_1^i|_{x=x_1^*} + \frac{\partial d_2(y_1, y_2)}{\partial y_1} \sum_{i=1}^{n} \gamma_2^i x_2^i|_{x=x_2^*} + \frac{\partial d_3(y_1, y_2, y_3)}{\partial y_1} \sum_{i=1}^{n} \gamma_3^i x_3^i|_{x=x_3^*}. \tag{5}$$

For the second class, we have

$$p_2 = \frac{\partial d_2(y_1, y_2)}{\partial y_2} \sum_{i=1}^{n} \gamma_2^i x_2^i|_{x=x_2^*} + \frac{\partial d_3(y_1, y_2, y_3)}{\partial y_2} \sum_{i=1}^{n} \gamma_3^i x_3^i|_{x=x_3^*} \qquad (6)$$

For the third and lower class, the unit price equals to

$$p_3 = \frac{\partial d_3(y_1, y_2, y_3)}{\partial y_3} \sum_{i=1}^{n} \gamma_3^i x_3^i|_{x=x_3^*} \qquad (7)$$

We observe that the unit price for each priority class equals to the extra (marginal) delay cost suffered by all the users of the specific class and the lower ones due to the marginal increase in demand for the services provided by the specific class.

Now, we must prove that the above prices, when published, will urge network users to buy those quantities that will maximize their net benefit. The net benefit's maximization problem, for every user i, is defined as follows:

$$\max_{\{x_1^i, x_2^i, x_3^i \geq 0\}} \left[V_i(x_1^i, x_2^i, x_3^i) - p_1 x_1^i - p_2 x_2^i - p_3 x_3^i \right] \Rightarrow$$

$$\max_{\{x_1^i, x_2^i, x_3^i \geq 0\}} [u_i(x_1^i, x_2^i, x_3^i) - \gamma_1^i d_1({y_1}^*) x_1^i - \gamma_2^i d_2({y_1}^*, {y_2}^*) x_2^i - \gamma_3^i d_3({y_1}^*, {y_2}^*, {y_3}^*) x_3^i - p_1 x_1^i - p_2 x_2^i - p_3 x_3^i] \qquad (8)$$

We will prove that the maximization problems (1) and (8) are equivalent, thus, the solution of the first provides the solution for the second.

Taking the partial derivative of the maximization function (8) for x_1^i, we have

$$\frac{\partial u_i}{\partial x_1^i} - \gamma_1^i d_1(y_1) - \frac{\partial d_1(y_1)}{\partial y_1} \gamma_1^i x_1^i - \frac{\partial d_2(y_1, y_2)}{\partial y_1} \gamma_2^i x_2^i - \frac{\partial d_3(y_1, y_2, y_3)}{\partial y_1} \gamma_3^i x_3^i - \frac{\partial d_1(y_1)}{\partial y_1} \sum_{i=1}^{n} \gamma_1^i x_1^i|_{x=x_1^*} - \frac{\partial d_2(y_1, y_2)}{\partial y_1} \sum_{i=1}^{n} \gamma_2^i x_2^i|_{x=x_2^*} - \frac{\partial d_3(y_1, y_2, y_3)}{\partial y_1} \sum_{i=1}^{n} \gamma_3^i x_3^i|_{x=x_3^*} = 0 \qquad (9)$$

As mentioned earlier, since a change in user's preferences (i.e. submitted volume of data) does not affect the experienced delay, the fractions $\frac{\partial d_1(y_1)}{\partial y_1}$, $\frac{\partial d_2(y_1, y_2)}{\partial y_1}$ and $\frac{\partial d_3(y_1, y_2, y_3)}{\partial y_1}$ are equal to zero. Thus, the equations (2) and (9) are equal. The same happens if we take the partial derivatives of (8) for x_2^i and x_3^i. We can conclude that the maximization problems (1) and (8) are equivalent and they have the same solution, i.e.$\{x_1^i, x_2^i, x_3^i\} = \{{x_1^i}^*, {x_2^i}^*, {x_3^i}^*\}$. Therefore, the prices published by the network operator maximize both the Social Welfare and the users' net benefit.

6. DISCUSSION

In this section we discuss some issues concerning the proposals we made in this paper. One first issue concerns the transfer rate offered by the Premium Service. It is obvious that if the GPRS operator defines the Premium Service's constant rate, then he can calculate how many simultaneous users a BSS can handle, taking into consideration the number of channels that the BSS serves, the number of timeslots in each frequency carrier assigned to GPRS traffic, the size of radio blocks and, for statistical decisions, user profiles. Thus, the operator will be able to perform Call Admission Control on Premium Service requests, which is required since this type of service is the only that offers strict guarantees.

A second issue is the length of a TBF, in the case of adapting Differentiated Services to the GPRS environment. As described in Section 4, the length of a TBF is set equal to the time to transmit one radio block. This happens because it is necessary for the BSS to receive a request for every packet that must be transferred on the uplink. Furthermore, the BSS must know the radio priority of each packet. This makes the emulation system easier to implement and keeps the computational load to the BSS very low. However, it also results in an unnecessary use of extra TBFs (and TFIs) for the transfer of packets from the same MS. On the downlink things are simpler since the BSS is the one that does all the scheduling and buffering.

Another important issue is which service class should be assigned to the IP packets that are reassembled at the GGSN and forwarded to the external network, in the case where Differentiated Services are also supported by the external PDN. There are many possibilities. The user's application may use the "Type of Service" or the "Traffic Class" field of the IP packet to define what service should be used to the external network. Another solution is to use the default priority class defined at the PDP Context activation phase. The first solution gives the user the ability to have his packets treated differently inside and outside the GPRS network. The second solution allows the user to have his packets treated uniformly in both networks. It is desirable that the user should be able to make the final choice, so the GPRS network should probably implement both solutions.

One last issue, concerning the pricing of such services is the exact determination of the optimal prices that have to be published. As it was mentioned in Section 5, the prices depend on the effect that a change in the demand of a priority class has on the delay experienced by the users of the specific class and the lower ones. But since this is rather complex to calculate, it is quite improbable for the network provider to know the exact function d_i. Thus, there is a question of how should the partial derivatives of the function d_i be calculated. The only solution, since the exact d_i function is not known, is measurement and estimation. For this purpose, the tatonnement process [12] can be used. Initially, the prices are set equal to zero, and for a period of time, the network operator observes the behavior of the system and measures the levels of congestion and delay for different time instances. By doing so, the network operator succeeds in constructing an approximate plot of the function d_i and is able to find its tangent, i.e. its derivatives. The new approximate prices can be caclulated from equations (5), (6) and (7). These new prices are published to the market and the system adapts. The network operator measures again the experienced delay in all classes and finds, using the same technique, the new prices. This iterative procedure, called tatonnement, stops when the old prices differ slightly from the new ones. It is not necessary that the published price is exactly equal to the estimated price because extreme conditions may occur during measurements. Instead, if we consider p_i^t the unit price of class i at time t and $\hat{p_i^t}$ the estimated price, a way of determining the new price at time $t+1$ is $p_i^{t+1} = \alpha \hat{p_i^t} + (1-\alpha)p_i^t$, where $\alpha \in (0,1)$ is an adjustment parameter used for stability purposes.

7. CONCLUSIONS

We have presented a way to apply the Differentiated Services framework to the GPRS wireless access environment. Our purpose was to enhance the GPRS network with QoS support that will be taken into consideration by the radio resource allocation procedures. For this purpose, the precedence QoS parameter and the radio-priority field were used, in combination with an adapted Two-bit DiffServ architecture. Note that the wireless access part is expected to be the most congested part of the GPRS network because of the scarcity of the wireless spectrum and therefore the part of the system where QoS support is most critical. At the same time, dynamic charging techniques can be combined with the service differentiation in order to make the resource allocation decisions efficient.

With the proposed architecture, GPRS operators will be able to provide end-to-end service differentiation fully compatible with the rest of the Internet and in cooperation with content providers. Mobile users will be able to select what service they want to be used for the transfer of their data and they will be charged accordingly. Even if the external networks do not provide service differentiation, GPRS operators will manage to offer a first level of differentiation to the wireless access network that they own.

Furthermore, we have introduced a pricing technique for charging the offered services. Taking into account that the three DiffServ classes are not independent from each other and that the QoS level each one offers depends on the congestion level of the specific class and the higher ones, we have provided an arithmetic model of constructing socially optimal prices that also maximize the users' net benefit. The exact determination of the prices to be published is achieved by using the tatonnement process.

With this pricing scheme, network users are urged to decide which service class is the most appropriate for them, since the amount they pay depends on the class they choose. By doing so, we provide a first level of assurance that the network resources will be utilized in the optimal way for both society and users.

8. REFERENCES

[1] R. Kalden, I. Meirick and M. Meyer, "Wireless Internet Access Based on GPRS," IEEE Personal Communications, vol. 7, no. 2, pp. 8-18, April 2000.

[2] C. Bettstetter, H.-J. Vogel, and J. Eberspacher, "GSM Phase 2+, General Packet Radio Service GPRS: Architecture, Protocols and Air Interface," IEEE Communications Surveys, vol. 2, no. 3, 1999, (http://www.comsoc.org/pubs/surveys/).

[3] GSM 02.60: "Digital cellular telecommunications system (Phase 2+); GPRS; Service Description; Stage 1"

[4] GSM 03.60: "Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Service Description; Stage 2"

[5] GSM 04.60: "Digital cellular telecommunications system (Phase 2+); GPRS; MS - BSS Interface; RLC/MAC protocol."

[6] P.F. Chimento, "Tutorial on QoS support for IP," CTIT Technical Report 23, 1998.

[7] F. Baumgartner, T. Braun, P. Habegger, "Differentiated Services: A new approach for Quality of Service in the Internet," Proc. 8th Int. Conference on High Performance Networking, Vienna, Austria, Sept. 1998.

[8] J. Heinane, F. Baker, W. Weiss, J. Wroclawski, "Assured Forwarding PHB Group," RFC 2597, June 1999.

[9] V. Jacobson, K. Nichols, K. Poduri, "An Expedited Forwarding PHB," RFC 2598, February 1999.

[10] K. Nichols, V. Jacobson, L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet," RFC 2638, July 1999.

[11] G. Priggouris, S. Hadjiefthymiades, L. Merakos, "Supporting IP QoS in the General Packet Radio Service," IEEE Network, vol.14, (no.5), p. 8-17, 2000.

[12] A. Gupta, D. Stahl, A. Whinston, "A Stochastic Equilibrium Model of Internet Pricing," 7th World Congress of the Econometrica Society, Tokyo, Japan, August 1995.

[13] S. Soursos, "Enhancing the GPRS Environment with Differentiated Services and Applying Congestion Pricing," M.Sc. thesis, Dept. of Informatics, AUEB, February 2001.