Proc. of the 2001 Tyrrhenian International Workshop on Digital Communications (IWDC 2001)
Evolutionary Trends of the Internet, September 17-20, 2001, Taormina, Italy.

Differentiated Services in the GPRS Wireless Access Environment

Sergios Soursos, Costas Courcoubetis, and George C. Polyzos

{sns, courcou, polyzos}@aueb.gr

Department of Informatics
Athens University of Economics and Business
Athens 10434, Greece

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. Even though Quality-of-Service notions and parameters are included in the GPRS specification, no realization path for Quality-of-Service support has been proposed. In this paper we adapt the Differentiated Services framework and apply it over the GPRS air interface in order to provide various levels of service differentiation and a true end-to-end application of the Internet Differentiated Services architecture.

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.

An 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) [11] using the standard Internet QoS frameworks (i.e., Integrated Services or Differentiated Services).

On the other hand, 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" [12]. As described later in this paper, GPRS operators can easily implement this proposal, with no need for radical changes to their existing GPRS network architecture.

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. Finally, we discuss some open issues and present our conclusions.

2 The GPRS Environment

GPRS [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 the Figure 1.

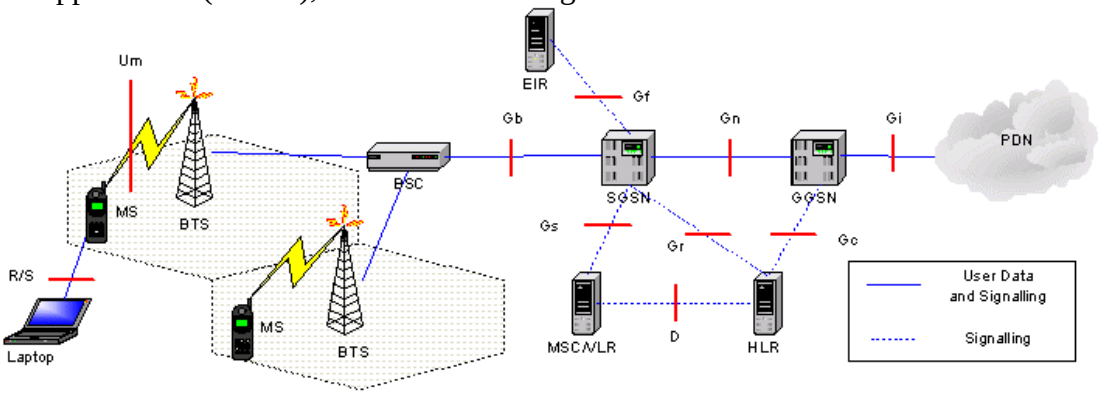

Figure 1 – The GPRS network

The SGSN keeps track of the location of mobile users, along with other information concerning the subscriber and its mobile equipment. This information is used to accomplish the tasks of the SGSN, such as packet routing and switching, session management, logical link management, mobility management, ciphering, authentication and charging functions. The GGSN, on the other hand, connects the GPRS core network to one or more external Packet Data Networks (PDNs). Among its tasks, is to convert the incoming packets to the appropriate protocol in order to forward them to the PDN. Also, the GGSN is responsible for the GPRS session management and the correct assignment of a SGSN to a Mobile Station (MS), depending on the MS's location. The GGSN also contributes to the gathering of useful information for the GPRS charging subsystem.

The core GPRS network is IP based. Among the various GSNs (SGSN and GGSN) the GPRS Tunnel Protocol (GTP) protocol is used. The GTP constructs tunnels between two GSNs that want to communicate [1]. GTP is based on IP. 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 ìs and is able to transfer 156.25 bits (both data and control). The recurrence of one particular timeslot defines a Packet Data CHannel (PDCH). 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 GSM, when a call is established, a channel is permanently allocated for the entire period. In other words, one timeslot is reserved for the whole duration of the call, even when there is no activity on the channel. This results in a significant waste of radio resources in the case of bursty traffic. In GPRS the radio channels, i.e. the timeslots, are allocated on a demand basis. 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. 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.

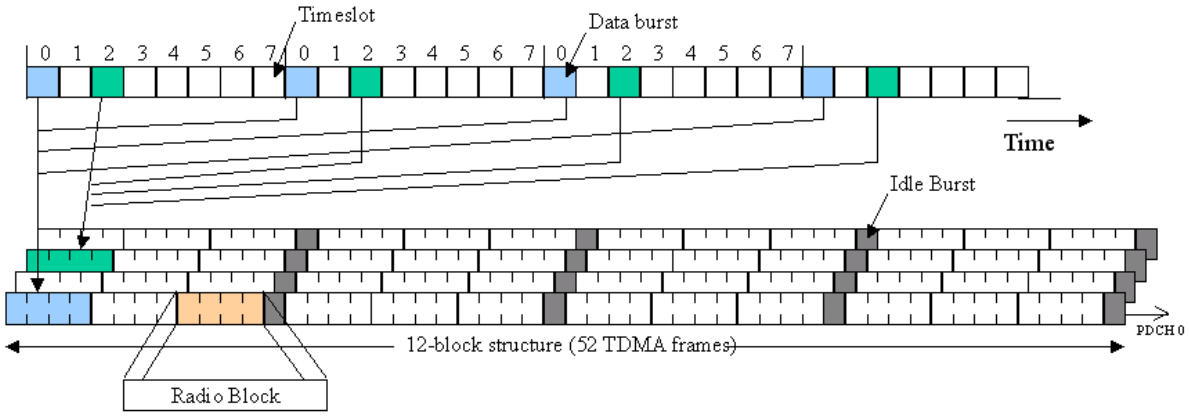

Figure 2 – Radio Channels

One or more (multi-slot capability) timeslots per TDMA frame may be assigned to a MS for the transfer of its data. 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]. During the activation of such a context, an address (compatible with the external network, i.e. IP or X.25) 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. The PDP context concerns the end-to-end path in the GPRS environment (MS ↔ GGSN).

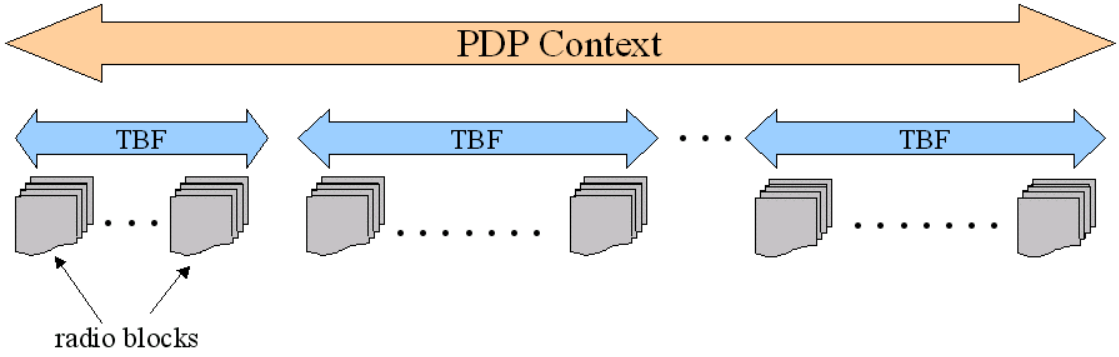

Figure 3 – PDP context & TBF

At the (lower) radio link level, when the MS starts receiving/sending data, a Temporary Block Flow (TBF) is created [5]. 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. The termination of a TBF 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.

ETSI has also specified a set of QoS parameters and the corresponding profiles that a user can choose. These parameters are precedence, reliability, delay, and peak and mean throughput [3]. Precedence (priority) defines three classes (high, medium and low). Three classes are also defined for reliability. Four classes for delay, nine classes for peak throughput and thirty-one classes for mean throughput (including best-effort). A user's profile may

require that the level of all (or some) parameters is defined. This profile is stored in the HLR and 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 will not guarantee that a user will conform to the agreed profile. Also, the QoS profiles are not taken into consideration by the resource allocation procedures. 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

The Internet is experiencing increased publicity lately and great success. Multimedia and business applications have increased the volume of data traveling 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.

Integrated Services [6], [7] was proposed as a first solution to the problem of ensuring QoS guarantees to a specific flow across a network domain, by reserving the needed resources at all the nodes from which the specific flow goes through. This is achieved through the Resource Reservation Protocol (RSVP) [6], which provides the necessary signaling in order to reserve network resources at each node. Although the Integrated Services solution works well in small networks, attempts to expand it to wider (inter-)networks, such as the Internet, has revealed many scalability problems.

An alternative architecture, Differentiated Services [7], was designed to address these scalability problems by providing QoS support on aggregate flows. In a domain where Differentiated Services are used, i.e. a DS domain, the user keeps a contract with the service provider. This contract, the Service Level Agreement (SLA) will characterize the user's flow passing through this DS domain, so as to include it in an aggregate of flows. The SLA also defines behavior of the domain's nodes to the specific type of flow, i.e. the Per-Hop Behavior (PHB). SLAs are also arranged between adjacent DS domains, so as to specify how flows directed from one domain to another will be treated.

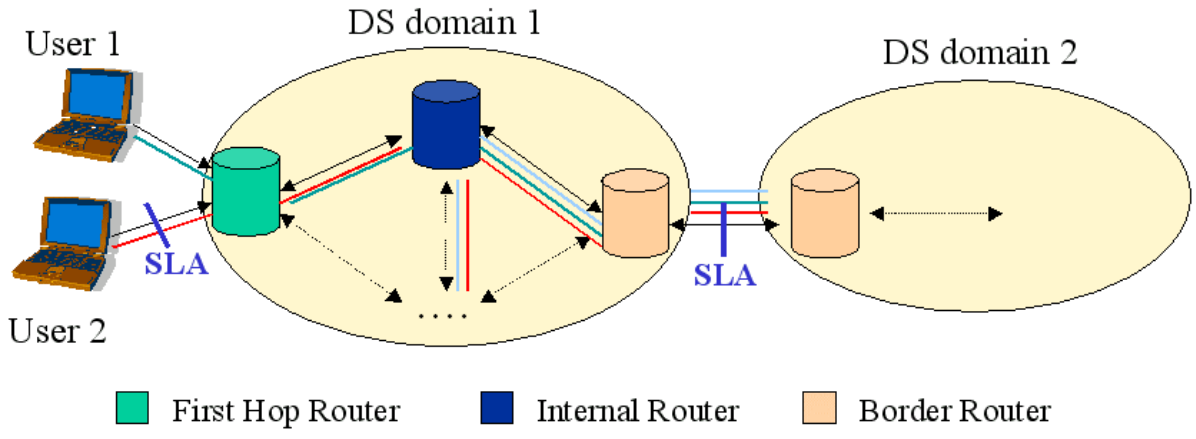

Figure 4 – The Differentiated Services Architecture

The DS field in an IP packet defines the PHB that each packet of a particular flow type shall have. This field uses reserved bits in the IP header-the "Type Of Service" field in

IPv4 and the “Traffic Class” field in IPv6. In Figure 4 we depict 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 Services [8]. In Premium Service, the key idea is that the user negotiates with the ISP a minimum bandwidth that will be available to the user no matter what the load of the link will be. Also, the ISP sets a maximum bandwidth allowed for this type of flow, so as to prevent the starvation of other flows. In most cases these two limits are equal, making Premium Service to act like a virtual leased line or, better, like the CBR service of ATM. 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).

It has been realized that there are many benefits from the deployment of both Premium and Assured services in a single DS domain. Premium service is thought of as a conservative assignment, while Assured service gives a user the opportunity to transmit additional traffic without penalty. Nowadays, Differentiated Services are known as the combination of these two services. This new architecture uses a two-bit field to distinguish the various types of 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 ISP has previously defined the constant rate that Premium Service should guarantee. Also, exceeding packets that belong to a Premium flow are dropped or delayed, while exceeding packets of Assured Service are forwarded as best effort. In Figure 5 we depict the tasks accomplished by the first hop router of the two-bit DiffServ architecture.

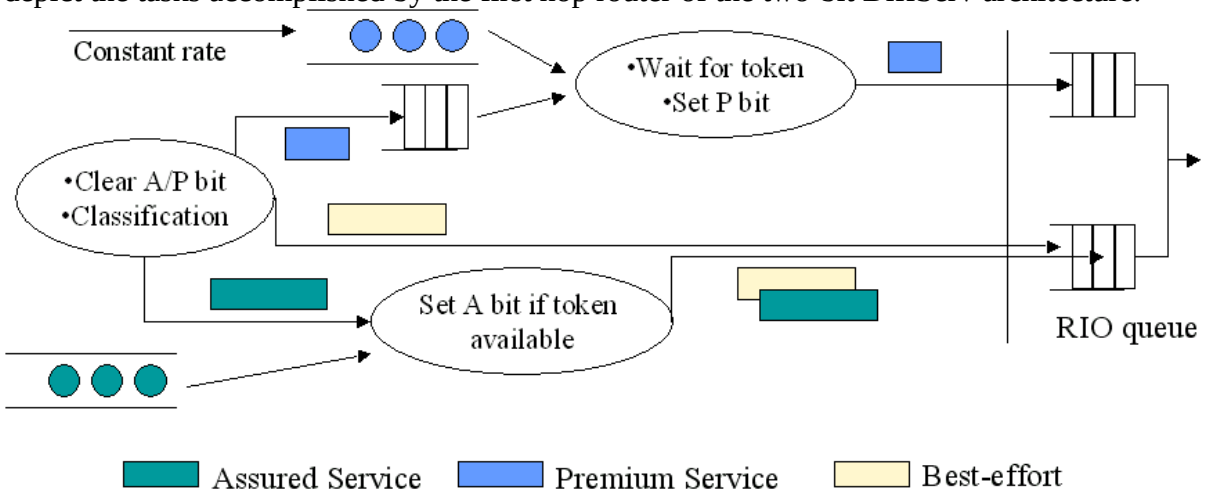

Figure 5 – First Hop Router

In the first hop router, 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 wait in the first queue until there is a token available in the token pool. When a token becomes available, the packets are forwarded to the output queue. 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. The queue that is used by both Assured Service and best-effort packets is a RIO (RED with In and Out) queue. RIO queues are RED (Random Early Detection) type queues with two thresholds instead of one, one for in-profile packets and one for out-of-profile packets. In this case, in-profile are the packets marked with the A-bit, while the rest (best-effort packets) are assumed to be out-of-profile. The threshold for in-profile packets is higher than the threshold for the out-of-profile packets, so that the later are discarded more often than the former. With this technique, a "better than best-effort" service is given to the packets using Assured Service.

Note that in the above figure, only the architecture concerning flows from one user is depicted. This is because the first hop router is the first, and only, router that controls and shapes individual flows. Therefore, we can assume that for each user there are two pools of tokens and a queue. The output queues are the same for all users and their characteristics depend on the outbound transfer rate of the router. The output queues can be served either by a simple priority scheme or by a more complex algorithm, such as the Weighted Fair Queuing (WFQ) algorithm.

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 agreements. Thus, the packets tagged with the P-bit are not queued, as in the first hop router, but they are discarded, as shown in Figure 6.

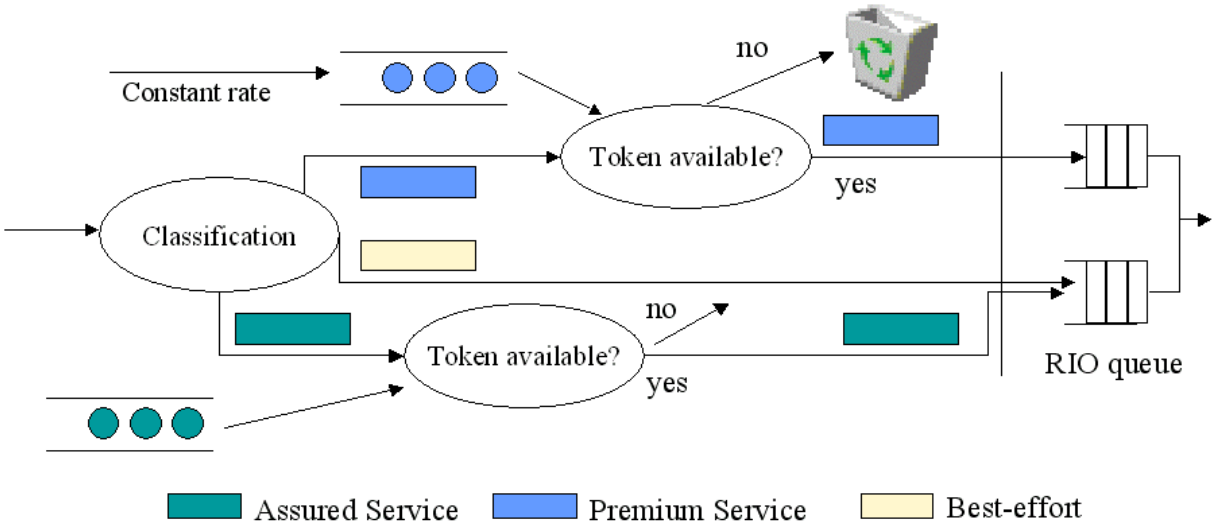

Figure 6 – Border Router

4 Differentiated Services 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.

Before we proceed to the application of the Two-bit DiffServ architecture in the GPRS environment, we must make some assumptions. First, we assume that the core GPRS network has sufficient resources for all traffic. In other words, the point of congestion is not the GPRS backbone, but the radio link, i.e. the access link that connects the MS with the appropriate BSS. This is an important but reasonable assumption given that the scarce resource in the GPRS network is the radio spectrum. Also, we assume that the size of the frames transferred over the radio link is fixed and equal to the size of a GPRS RLC/MAC packet (frame).

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. 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. Non-conforming traffic should be either discarded or degraded, as depicted in Figure 6. 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).

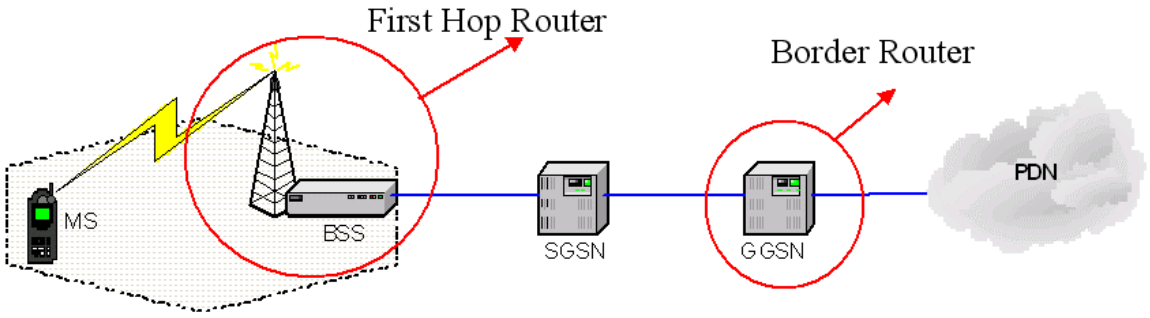

Figure 7 – Two-bit Differentiation Architecture in GPRS

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 5. 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 5, 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.

When a pair is inserted into the system, three possible actions may occur:

- the pair is forwarded to the appropriate output queue, if the counters of the Premium or Assured Service’s pools are bigger than zero, or if the priority chosen is equal to “3”
- the pair is inserted into the waiting queue of Premium Service, if the corresponding counter is equal to zero, or
- the pair is forwarded to the corresponding output queue with its priority set to “3”, if the Assured Service’s pool counter is equal to zero.

If the priority chosen is zero, then the corresponding value in the pair inserted into the system will not be zero. Instead, the real value from the default PDP context is used.

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 in the next section.

5 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. Since the radio priority is defined only during the establishment of the TBF, when the MS requests permission to transfer its data, the result is to limit the duration of a TBF to the transmission of one radio block. 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 is the charging and pricing of such services. Although it is outside the scope of this paper, we should mention that the architecture described here enables charging using congestion pricing techniques. A first step in this direction is described in [12], where the existing congestion pricing theory is extended to the DiffServ environment described here.

6 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 Differentiated Services 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.

Acknowledgments

This research was supported by the European Union's Fifth Framework Project M3I (Market-Managed Multiservice Internet - RTD No IST-1999-11429).

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+); General Packet Radio Service (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+); General Packet Radio Service (GPRS); Mobile Station (MS) – Base Station (BSS) Interface; Radio Link Control/Medium Access Control (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," Proceedings of Eighth International Conference on High Performance Networking, Vienna, Austria, 21-25 Sept. 1998. Edited by: Van As, H.R., Norwell, MA, USA: Kluwer Academic Publishers, 1998. p. 255-73.

[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, Sept.-Oct. 2000, vol.14, (no.5), p. 8-17.

[12] S. Soursos, "Enhancing the GPRS Environment with Differentiated Services and Applying Congestion Pricing," M.Sc. thesis, Dept. of Informatics, Athens University of Economics and Business, February 2001.