A Framework for QoS Support in the UMTS/GPRS Backbone Network Using DiffServ*

Farshid Agharebparast and Victor C.M. Leung
Department of Electrical and Computer Engineering
The University of British Columbia
2356 Main Mall, Vancouver, BC, Canada V6T 1Z4
Email: {farshid,vleung}@ece.ubc.ca

***Abstract*– A distinguishing feature of the Universal Mobile Telecommunications System (UMTS) is the support of different levels of quality of service (QoS) as required by subscribers and their applications. To provide QoS, the UMTS backbone network needs an efficient QoS mechanism to provide the demanded level of services on a UMTS core network. This paper presents a methodology of provisioning QoS in this backbone network based on the Differentiated Services (DiffServ) model. DiffServ is a relatively simple but scalable IP-based technology, which can efficiently provide QoS in networks of DiffServ supporting routers. This is accomplished by defining a framework for setting a DiffServ-based UMTS backbone router, as well as the requisite mapping function for interworking between a DiffServ domain and UMTS. Efficient schemes are presented for the scheduling and buffer management components of the backbone router supporting DiffServ. The performance of this system for provisioning UMTS primary QoS classes is evaluated by computer simulations. The results show that DiffServ can be an effective candidate for UMTS backbone bearer service.**

I. Introduction

GSM, the Global System for Mobile communications, is a very successful second generation cellular telecommunications system, which is used worldwide. Due to its dominance, it is desirable that the third generation (3G) system evolving from GSM utilizes as much GSM infrastructure as possible. General Packet Radio Service (GPRS) [1] [2] is an enhancement of GSM, which accommodates data connections with much higher bandwidth and is a true packet switched wireless system. GPRS utilizes the GSM infrastructure and provides packet data services by adding two new types of nodes to the network: gateway GPRS support node (GGSN) and serving GSN (SGSN), and by software upgrade of some other GSM nodes. Through GPRS, GSM is evolving into the 3G cellular network called the Universal Mobile Telecommunications System (UMTS) [3] [4]. UMTS, while reusing the GSM/GPRS core network, has a completely different radio access network, employing WCDMA [5][6][7] instead of the TDMA used in GSM/GPRS.

The prominent advantage of a 3G network is its ability to provision a wide range of diverse services with quality of service (QoS) guarantees. In these systems, end-to-end QoS is realized by the QoS performance of a number of bearer services along the datapath. UMTS specification defines that as part of the UMTS network, the backbone network is the portion of the datapath that provides the backbone bearer service and the associated QoS support. It is an IP network consisting of GGSN and SGSN nodes that uses tunneling to transfer the data packets. To realize its QoS support objective, an IP-based QoS technology should be deployed. The Internet Engineering Task Force (IETF) has developed a number of these QoS schemes, most notably, Differentiated Services (DiffServ) [8] and Integrated Services (IntServ) [9]. The basic characteristic of the core network of a wireless system is that it needs to accommodate a large number of connection sessions. The data rate of each connection, however, is usually well below the usual Internet backbone speeds and is not a technological problem. Due to these considerations, although IntServ has the advantage of providing fine-grained QoS control, its use will inevitably have scalability and complexity problems. On the other hand, DiffServ is an IP-based QoS mechanism, being standardized by IETF, that is scalable and can be implemented with little management complexity, especially for networks with well-defined traffic profiles. Since a backbone network is usually under the control of a wireless service provider, it is promising to deploy DiffServ as the QoS support mechanism in the backbone network.

A number of technical challenges need to be addressed

*Paper accepted for publication in Computing and Informatics. This work was supported by Motorola Canada Ltd. and the Natural Sciences and Engineering Research Council of Canada (NSERC).

regarding the implementation of DiffServ on a UMTS backbone network. To list a few, they include structure of the routers, definition of Per Hop Behaviors (PHBs), QoS mapping, and admission control. There is some prior work that proposes and discusses the deployment of DiffServ for the UMTS (or similarly GPRS) backbone network. Deployment of DiffServ and of IntServ for GPRS are compared to study the QoS reservations across the GPRS core network [10]. QoS issues in the core network of next generation mobile systems are also studied in [11] [12] [13].

However, to the best of our knowledge, there is no prior work in the open literature which specifically, and in detail, considers the structure of a DiffServ-based router and proposes algorithms for its components to satisfy the QoS requirements of a 3G core network (CN), or recommends QoS mapping between UMTS and DiffServ. Our paper fills these voids. Moreover, we provide a compact framework and basic methodology for further research on QoS support in the UMTS backbone bearer service using DiffServ.

This paper focuses on two of the important issues that need to be resolved for implementing DiffServ on a UMTS backbone network. First, a basic plan for QoS mapping from one QoS system to the other is presented. Second, the structure of a DiffServ router [14] [15] with the appropriate functionality for forwarding the UMTS aggregate traffic classes is proposed. Specifically, we integrate the novel algorithms, for service scheduling and buffer management which were presented in our earlier papers [16] [17] into a whole system and provide a performance evaluation.

However, this paper does not deal with dynamic resource allocation and admission control mechanisms. It is assumed that in the DiffServ network, based on the users' demands and gathered statistics, a set of PHBs with a defined resource allocated to each, is available. It is also assumed that the admission control has been done in the CN Iu Edge node. A framework for the admission control mechanism that is applicable to UMTS network is presented in [18]. Since the QoS classes and backbone network architectures of UMTS and GPRS release 99 have been defined in close agreements, the discussions provided in this paper are applicable to both systems [4] [19].

The paper is organized as follows. Sections II and III provide overviews of the UMTS infrastructure and QoS architecture, and DiffServ conceptual model, respectively. Section IV presents a methodology for configuring elements of a DiffServ-based UMTS backbone router and its QoS mapping function. Section V describes the simulation model of the backbone network implemented in OPNET and discusses the results. Finally, Section VI concludes the paper.

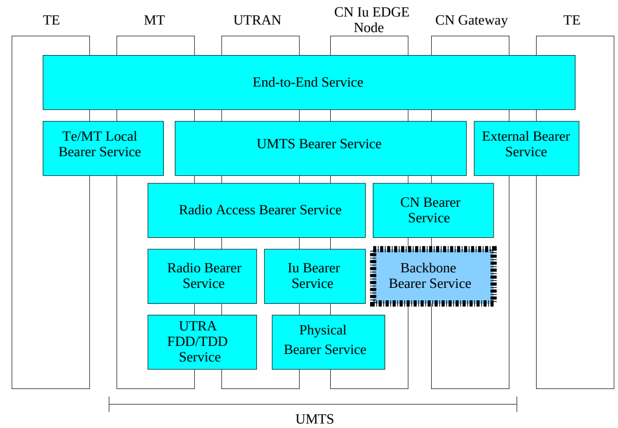

Fig. 1. UMTS QoS Architecture and Levels.

II. UMTS QoS Support

UMTS is a 3G standard developed by the European Telecommunications Standards Institute (ETSI). Due to the evolution from different second generation standards, there are a number of 3G alternatives to UMTS. For example, CDMA2000 is a 3G enhancement of IS-95, developed by the Telecommunications Industry Association. The International Telecommunications Union (ITU) is developing IMT-2000 which incorporates these diverse standards into a global mobile telecommunication standard. The world-wide 3G Partnership Project organizations, 3GPP and 3GPP-2, are industrial forums aimed at smoothing the implementation of 3G standards. In this paper, UMTS is considered and described.

UMTS is a mobile network providing wireless services for both data and voice. The quality of service defines the level of service this network provides to a subscriber on demand or based on the level of service subscription. It is the end-to-end QoS, from source node to destination, that counts in a user's viewpoint. The end-to-end QoS, with UMTS involved, depends on the performance of three distinct parts of the data path [4]: the MT/ET bearer service, the UMTS bearer service, and the external bearer service (Figure 1). A bearer service incorporates all the aspects needed to provide a set of pre-defined services. The UMTS bearer service is itself realized by the

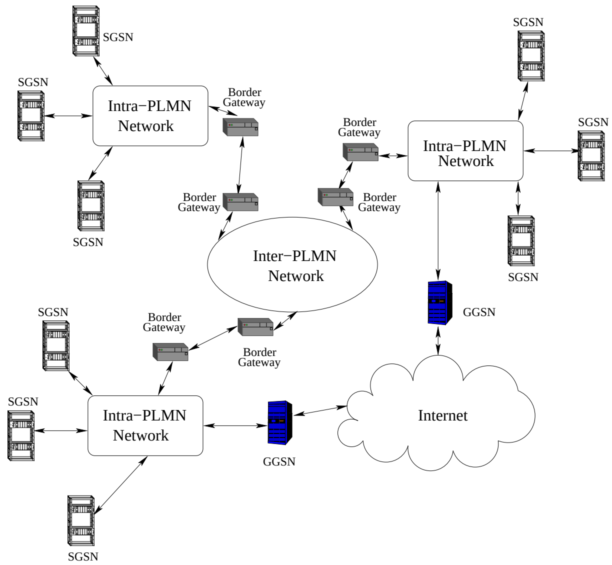

Fig. 2. Backbone Network.

radio interface and core network bearer services.

Figure 2 shows the structure of the GPRS backbone network. It is an IP network that consists of a collection of intra-PLMNs (public land mobile networks) and inter-PLMNs. An intra-PLMN is composed of the SGSN and GGSN nodes of a local PLMN, and an inter-PLMN connects a number of intra-PLMNs together.

In UMTS, QoS requirements are conveyed by QoS profiles, which have a single parameter with a number of attributes. A QoS profile is negotiated with the user as part of the PDP context establishment process. The QoS attributes define the grade of service requested by the user. They include: the traffic class, maximum and guaranteed bit rates, delivery order, transfer delay, maximum SDU size, and SDU error ratio. The most noteworthy of these attributes is the traffic class which defines the type of service requested. The rest of the attributes describe the demanded grade of service in more detail.

There are four UMTS traffic classes: conversational class, streaming class, interactive class and background class, listed according to their relative delay/jitter sensitivity, from highest to lowest. Conversational and streaming classes are meant to carry real-time traffic streams. The main distinctions between conversational and streaming classes are their delay and delay variation sensitivities. Interactive and background classes, however, are intended to cover all traditional Internet applications, such as WWW, email, and FTP.

Conversational Class: This class is used for conversational and multimedia applications, such as GSM telephony speech, voice over IP, and video conferencing. These applications are mainly used by two (or more) peers of human end-users. The fundamental QoS characteristics of this class are the preservation of the time variation between consecutive entities of the traffic, and stringent low delay.

Streaming Class: This class carries the traffic of a video or audio stream from a host machine to a human end-user, and is basically a one way transport. Its QoS characteristic is that it requires the time relation between the traffic consecutive entities be preserved. However, it has a much lower delay sensitivity when compared to the conversational class. An example of its application is MP3 video streaming or Internet radio.

Interactive Class: This class is for applications where an end-user (human or machine) is online, asking for some data from a host machine. The main characteristic of this class is its request-response pattern, with a lower delay sensitivity compared to the above traffic classes. It may only need payload content preservation.

Background Class: When an end-user (mainly a machine) sends or receives data in the background without any delay sensitivity, this class is applied. Examples of this kind of application are email and database downloading.

III. Differentiated Services

DiffServ is an IP-based QoS support framework that achieves simplicity and scalability by aggregating the traffic into a finite set of PHB (per hop behavior) groups [15]. The PHB group of a packet is identified by a codepoint in its IP header, and defines the specific forwarding

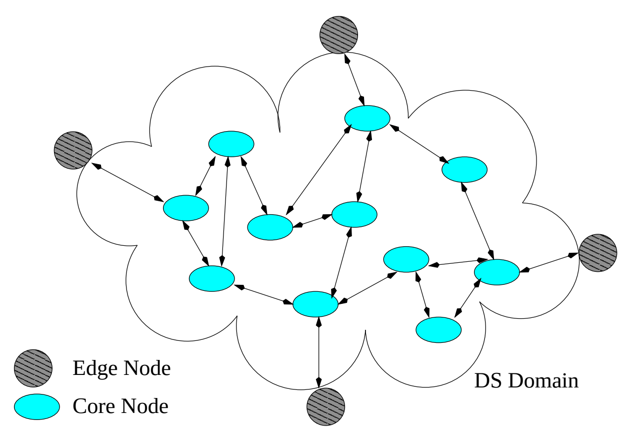

Fig. 3. A DiffServ Domain.

treatment that the packet will receive at each node.

A DiffServ network is partitioned into domains, each having two types of routers: edge or core. A DiffServ domain (Figure 3) is completely identified by its set of supported PHB, its edge and core routers, its policy in mapping QoS of an incoming packet to a pre-defined PHB, and the level of QoS associated with each PHB. Each packet belonging to an aggregate class is marked with a DiffServ codepoint (DSCP) in the IP header identifying a PHB group at an edge router, and then each node in that domain services the packet according to its DSCP.

A. *DiffServ Router*

This section describes the elements used in a DiffServ-aware node. Figure 4 gives a high-level view of the elaborate component interaction of a DS node. The routing core is an abstract of a router's normal routing functionality, and is not discussed in the DiffServ model. The configuration and management interface monitors service provisioning and gathers statistics regarding the traffic of each service level. These are used as network management bases from which DiffServ policies are assigned or altered. The network administrator interacts with this module through a network management protocol. The optional QoS agent is included to add per-flow and per-flow-aggregate signaling capability to the DiffServ model. This allows the node, for example, to snoop RSVP messages.

A DiffServ router consists of some classification, traffic conditioning and queuing blocks, such as a classifier, meter, marker, dropper, counter, multiplexer, queues, and scheduler, embodied into its ingress or egress interfaces. A simple router might have two egress/ingress interfaces.

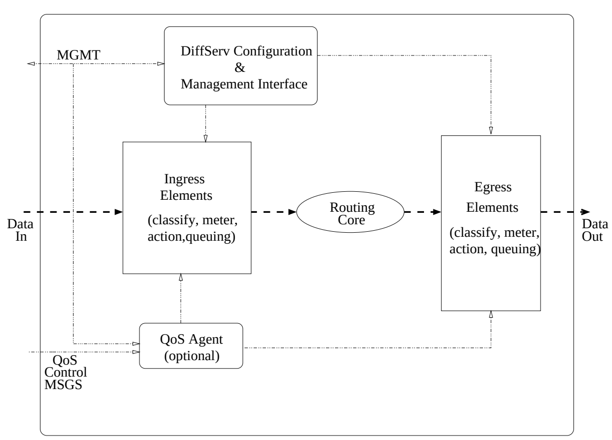

Fig. 4. High Level View of a DiffServ Router.

It is also possible to have an arbitrary number of these interfaces in an actual router. Besides, it is not mandatory to implement all of these blocks and their components on both ingress and egress points. The configuration of components depends on the service requirement of a router.

A data packet takes a datapath that is associated with its PHB. On arrival, a packet is first classified according to a set of rules. It is then checked by a metering block for being within its allocated rate. Afterward, the packet is passed along a set of traffic conditioning blocks such as a marker, dropper, and counter, and if accepted, is enqueued in the queuing block and then transmitted based on certain scheduler policy. There are four kinds of components in an ingress/egress interface: traffic classifiers, meters, actions, and queuing elements. Action elements are markers, droppers, counters, and multiplexers.

Mapping Function: Mapping is the function of translating the QoS of one system to the QoS parameters understood by another system. It is a necessary function for a UMTS packet entering the DiffServ-aware network, or when a packet coming from an external network enters the UMTS network through a GGSN. This function is usually integrated with the classifier into one functioning block.

Classifier: A Classifier is a functional element that selects packets, based on a policy. It might classify packets based on the content of the packet header, other data, and/or some implicit information related to that packet. In a DiffServ model, the packets are classified according to the content of a field in the IP header, originally called Type of Service (ToS) in the IPv4 protocol. ToS is an eight bit field and DiffServ uses 6 bits of it, also called a DS field, to convey the DSCP. A DSCP completely identifies the PHB associated to that packet. The remaining 2 bits, called CU (currently unused), are not exploited in DiffServ, and are proposed for use in the congestion notification function by the IETF Explicit Congestion Notification (ECN) group. The classifier has a single input and splits the incoming stream into a number of outgoing ports. The most common way of implementing classifiers is to use a number of filters. A filter is a set of match conditions based on a classification key. The contents of the DSCP is passed through the filter, and according to the matching results, it is grouped into a PHB aggregate.

Meter: A meter is a functional block responsible for measuring the rate characteristics of a packet stream. These gathered statistics are used by other blocks for their decision making. For example, a marker uses this data to re-mark the packet DSCP. IETF has two RFCs that proposes two types of meter implementation: the time sliding window meter/marker, and the two (three) token bucket meter/marker.

Marker: This module is responsible for setting the DS field of a DiffServ packet according to various rules. It may also use the statistics gathered by a meter to decide on the level of a packet's conformance to its allocated rate characteristics. For a non conformant stream, it takes action by re-marking the packet in either a class with lower priority, or lower drop precedence. It is also responsible for assigning a valid DSCP to a packet with an unknown DSCP.

Counter: A counter is responsible for updating the packet counter. It is also used to count the number of packets passing through it, as a form of traffic statistics.

Dropper: There are two major kinds of dropping elements in a DiffServ router: absolute and algorithmic. While absolute droppers are quite straightforward, research is still being undertaken on the viability, usability, and the schemes of algorithmic droppers. Random Early Detection (RED) [20] is a well-known algorithmic dropper, suggested for avoiding congestion in the network. It detects the incipient congestion and notifies the congestion manager of the source to decrease its flow rate. This way, congestion is avoided.

Queuing Block: This is the block that stores the packets while they are waiting to receive service by the scheduler and depart to the next hop or DiffServ module. There are two different types of buffers used for enqueuing packets: shared memory buffers and dedicated buffers. In a dedicated buffering method, each queue has its own separated memory space, while a shared buffering scheme uses a large memory which is shared among different traffic classes, according to the sharing rule.

Scheduler: The scheduling block has the most impact on the level of service a packet receives. It decides on which queue, among a set of queues, to service next.

B. *DiffServ PHBs*

The DiffServ working Group of IETF has defined a number of different PHB groups for different applications. While PHBs define the level of packet treatment on each node of a DS domain, Per Domain Behavior (PDB) groups define the level of treatment a packet receives edge-to-edge of a DiffServ domain. IETF has very few numbers of standard PDB. Due to their simplicity, we consider PHBs only. In this section an overview of all the up-to-date and major PHBs are presented. They are: Expedited Forwarding (EF), Assured Forwarding (AF), Alternative Best-Effort (ABE), Bulk Handling (BH) PDB, Lower than Best-Effort (LBE), Dynamic Real-time/non-real-time (RT/NRT), Assured Rate PDB (AR), and Network Control (NC) [8].

EF, AF, and BE are the most common PHBs of this list. EF PHB is aimed for a low loss, low latency, low jitter, assured bandwidth edge-to-edge service. It has also been called the Premium Service. Its recommended codepoint is 101110, and therefore, there can be only one instant of EF in a DiffServ domain with the recommended DSCP. AF PHB group provides four independently forwarded AF classes. A packet in each of these AF groups can be assigned to three dropping precedences. Therefore, 12 instances of AF with recommended DSCPs can exist in a DiffServ domain. Table I shows the recommended DSCP of the AF PHBs:

BE PHB is the traditional Internet traffic and its usage implies that there are some network resources available and there is a good faith commitment that the nodes in the path will do their best to forward the packet in a fair manner, however, there is no guarantee on its delivery or its level of service. Its recommended codepoint is 000000.

The other PHBs are less common, but can be used to provide special PHBs with different varieties of service classification. NC PHB is used for NC packets, which carry any signaling data needed in the DiffServ control plan, to exchange management information between DiffServ routers within a DiffServ domain. Its recommended codepoint is 11x000. The primary goal of defining the LBE PHB group is to provide an alternative to best effort traffic at the time of congestion. This PHB carries packets of an application that has a lower drop precedence, compared to best effort traffic. RT/NRT PHB aims at providing the delivery of real-time and non-real-time traffic.

BH PDB is used for traffic that may be starved even in a properly functioning network. This starvation is not a requirement of the PHB, but is used for comparison with

Drop Prec.	Class 1	Class 2	Class 3	Class 4
Low	001010	010010	011010	100010
Medium	001100	010100	011100	100100
High	001110	010110	011110	100110

TABLE I
RECOMMENDED CODEPOINTS OF AF PHB.

the best effort PHB. It can be used for carrying packets of an application that has a sufficiently low value. The ABE PHB group gives the best effort applications an option: to choose between receiving lower delay, or receiving higher throughput by defining two types of (green and blue) packets. Finally, AR PDB defines the overall rate allocation a packet received throughout a DiffServ domain. This PDB is suitable for carrying the traffic of applications that need rate assurance, but do not require delay bounds.

IV. Methodology

This section presents a basic but simple methodology that can be used for internetworking, interface, and deployment of DiffServ in the UMTS backbone network. Although the model is not comprehensive enough to cover all the requirements of an actual system, it includes sufficient details to facilitate modeling of more complex systems. The model presents the structure and policies needed to support QoS for typical UMTS traffic streams.

There are a number of steps that should be taken to implement DiffServ in the UMTS backbone network. First, the levels of traffic differentiation, including traffic classes and other QoS attributes that are provided by a UMTS provider, are defined. Accordingly, a set of PHBs need to be selected so that each has the characteristics required to provide the equivalent QoS of one of the traffic QoS levels. Then, a set of rule for mapping these two sets of QoS parameters from one to another is defined.

Having defined the group of PHBs and their characteristics, the DiffServ domains are created and the routers are configured so that they can support such PHBs. This includes the selection of proper DiffServ components and algorithms for each, and the setting and configuration for each PHB path in the router. Having defined the above traffic classes passing through a DiffServ core network, the related aspects of QoS mapping and router structure for provisioning a set of favorable PHB treatment to provide the required QoS is addressed. Accordingly, the building block components of a DiffServ router and their functionalities are described, and the architecture for each component is proposed.

In the following section, first the QoS mapping function required for internetworking between DiffServ and UMTS is discussed, and then the structure of a DiffServ-based UMTS backbone router and the scheme appropriate for each of its building blocks are addressed. The focus of the discussion is on four components: scheduler, algorithmic dropper, meter, and classifier. The other router's components such as: counter, marker, and multiplexer have not been discussed in our model because of their inherent structural simplicity.

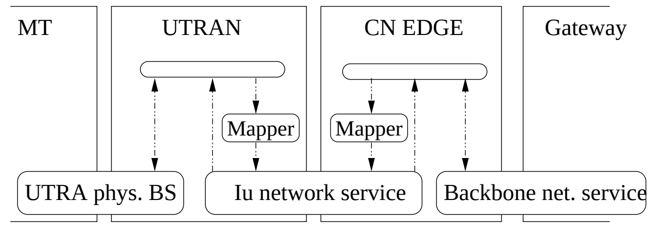

Fig. 5. The Position of the Mappers.

A. *Mapping and Classification*

In a UMTS DiffServ-based backbone network, there are two bearer interfaces where QoS mapping should be performed: between a backbone network and UTRAN (Figure 5), and between a backbone and external bearer [4]. We are mainly concerned about the former here, as the latter depends on the QoS structure of the external network and involves general DiffServ QoS mapping. When a packet passes to (from) the UMTS Iu bearer from (to) the backbone network, the QoS characteristics corresponding to it should be translated to ones understood by the new QoS system.

One of the most important responsibilities of a DiffServ domain ingress router is to map the QoS parameters of the incoming packet into the QoS PHB supported by that DiffServ domain, while providing the same grade of service. When a packet is marked by an appropriate DSCP, the other nodes (DiffServ domain core nodes) only need to perform a simple traffic classification based on the packets DSCP by using a set of filters. As discussed in section III-B, IETF has developed a number of standard PHB or PDB groups which can be used. A service provider then, should consider all the UMTS QoS classes that are defined in the network, taking into account all the supported UMTS traffic classes and QoS attributes. These classes are aggregated into a manageable set of new groups, based on their close QoS requirements. Correspondingly, a set of available DiffServ PHBs that have similar characteristics are chosen, and a one-to-one mapping is assigned. The bandwidth allocated to these PHBs can be assigned in a static manner or dynamically. In the static resource allocation, the decision is made based on the historical statistics of the resource used by each PHB. For a dynamic resource allocation, the QoS configuration of the routers are reset, when needed, through DiffServ management signaling and the routers' configuration and management interface.

We consider four traffic groups, each corresponding to one of the four UMTS traffic classes, that is, the traffic flows of each of these classes are aggregated into four groups, each represented by one DiffServ PHB. In addition, another PHB is considered for DiffServ network management purposes, resulting in five groups of traffic aggregates in the core network. The first group is called the DiffServ network control (NC) group, which is used for DiffServ administrative signaling in its control plane, and usually needs very little bandwidth, but has the most delay sensitive traffic. There is very little documentation regarding NC PHB in the DiffServ literature; nevertheless, in most cases such a PHB is necessary. Group one is an aggregate of traffic with a conversational class, having high delay and jitter sensitivity. Group two is an aggregate of traffic with a streaming class and is jitter sensitive. Group three is an aggregate of traffic with an interactive class, and finally, group four is an aggregate of background class traffic. Group three has moderate delay sensitivity and group four is sensitive to neither delay nor jitter.

One method of mapping applicable for a DiffServ-based backbone network, which is used for the traffic classes in this paper, is summarized in Table II. Nevertheless, depending on the services defined in the system, other mapping schemes might be used. In this prototype, only main UMTS traffic classes have been used; however, each of these groups can be divided into as many subgroups as needed. For example, background traffic class can be mapped to BE, LBE and BH PHBs. The proper performance of the QoS system mostly depends on the level of forwarding treatment received by each group, and therefore, the names and codepoints can be specific to the implementation, especially for an isolated DiffServ network.

Group	G0	G1	G2	G3	G4
PHB	NC	EF	AF1	AF2	BE

TABLE II
QOS MAPPING TABLE.

B. *Dropper Elements*

Since there are both real time and non real time traffic aggregates in the model, we make use of both types of DiffServ dropper blocks. An absolute dropper is simply used to drop the overloaded traffic of a real-time application. Because an overly delayed packet is no longer useful, and is better discarded. Therefore, when congestion happens, the real-time packets are forwarded to this block in the DiffServ node. In addition, when a real-time traffic misbehaves, the overloaded traffic is sent to the absolute dropper. Hence, for traffic group one and two, absolute droppers are used.

In contrast, an algorithmic dropper drops a packet according to a specific algorithm, usually well before the queue is full. There are a variety of RED algorithms. RED on shared-memory buffer schemes [16] show a desirable performance in terms of packet loss ratio when compared with dedicated queues with the same memory space. RED on shared memory with minimum allocation (R-SMA) is a favorable one among these schemes. In this scheme, a whole memory space is shared among a subgroup of traffic classes, and the decision to drop a packet is given based on the queue sizes of both the whole and individual subqueues. For non real-time traffic, especially a background traffic class, R-SMA is an effective congestion control mechanism. The following pseudo-code defines this scheme.

for each packet arrival with class i
calculate the average queue size avg
calculate average sub-queue i size avg(i)
if min-th(i) $\leq$ *avg(i) and avg* $<$ *max-th*
    calculate drop probability p_a
   with probability p_a *mark the arriving*
      packet
else if max-th $\leq$ *avg*
      mark/drop the packet

C. *Scheduler*

One of the basic tools of packet differentiation in networks is the scheduling element since its performance has the highest impact on the level of service a packet receives. There are a large number of different scheduling schemes presented in the literature (for a survey see [21]), each of which has some advantages and some disadvantages. In a DiffServ-based UMTS backbone network, the most desirable scenario is to have a fair, efficient, and simple scheme which supports link sharing and delay bounds. These are the fundamental characteristics that provide unbiased service with the desired delay or bandwidth allocation. It should also be able to assign different delay bound guarantees independently of the bandwidth allocations, and vice versa. This decoupled delay bound and bandwidth allocation capability enables the system to effectively and with minimum complexity, assign a diverse range of delay and resource allocations to each of the PHBs. Due to this, a DiffServ operator has a higher degree of control and does not need to tightly control bandwidth allocations for delay bound realizations. However, it is in general, not possible to concurrently realize complete link-sharing and guaranteed service, while having decoupled delay and bandwidth allocations [22].

Among the existent scheduling schemes, hierarchical fair service-curve (H-FSC) [22] has the desired properties mentioned above. However, its main drawback is that services are characterized by using service curves. Although theoretically just, it usually ends up with non-linear and complex service curves. We proposed a simpler mechanism in [17] called DDB-FFQ (Decoupled Delay-Bandwidth Frame-based Fair Queuing), which is a scheduling system having all the favorable features with less complexity. It is a link sharing [23] scheduling system consisting of a modified Frame-based Fair Queuing (FFQ) [24] scheduler and a number of rate estimators. FFQ is a rate-proportional scheduler [25] which closely approximates GPS while having a $O(1)$ timestamp computation complexity.

In our model, the NC traffic aggregate is given an absolute non-preemptive priority over other classes, that is, it is scheduled to be served anytime there is a packet in its queue. Since this traffic class is for management and signaling information, it only uses a very small fraction of the bandwidth with controlled packet sizes. Its priority scheduling will not have a significant impact on the service received by the other traffic aggregates, but might add a small constant (depending on the NC packet size and rate) to their delay bounds.

The rest of the traffic aggregates are serviced under the DDB-FFQ discipline [17]. For each traffic aggregate a pair of delay bounds and bandwidth is offered, depending on its requirements. The scheme tries to concurrently realize both delay and bandwidth demands, and in times of conflict between these requirements, link sharing takes effect.

D. Meter

The statistics estimated by a meter are used in a number of other components. It is used for subsequent traffic management decisions for the DiffServ control plane. It is also used for monitoring the traffic and detecting non-conformant traffic and congestion situations. The calculated rate is an indicator of the level of conformance a stream has according to certain rules. If the traffic is determined as non-conformant, the packet is forwarded to a dropper or a marker. The packet is then dropped or remarked to a lower level. For real-time applications, it is usually not useful to remark the packet and the packet is dropped. For other applications, the packet may be remarked to a lower class.

The meter is an essential part of the DDB-FFQ service discipline that is used for the scheduling block in our model [17]. It is used to evaluate the output link rate received by each traffic aggregate. It is of the time-sliding-window type, that is, an exponential weighted moving average (EWMA) rate estimator. A EWMA rate estimator is quite simple and requires some a priori traffic descriptors of the traffic. It has two key parameters: the time constant of the estimator and the frequency that estimation is performed. The time constant defines how long it would take before the estimated rate changes when traffic rate changes. The selection of its value depends on the traffic rate and the expected sensitivity of the estimator. The time constant roughly equals to $\frac{-s}{b*ln(1-\omega)}$, where s is the packet size, b is the allocated bandwidth and ω is the weight factor of EWMA [23]. Therefore, with fixed s and b, the weight factor controls the time constant. The estimation is performed at each packet arrival, to decide whether each packet belongs to a well-behaved aggregate. Thus, the misbehaving traffic is limited to its link sharing allocated rate within a time period proportional to the time contrast associated with its meter.

V. Simulation Model and Discussions

In this section, we present the simulation model and discuss its results. The simulation model was constructed in the OPNET environment to evaluate the performance of a system based on the paradigm we presented earlier. The objective of the computer simulation is to evaluate the performance of the model in terms of the QoS constraints: latency, bandwidth and packet loss. It is aimed at calculating the delay observed by each aggregate, to evaluate the bandwidth assigned to each group, relative

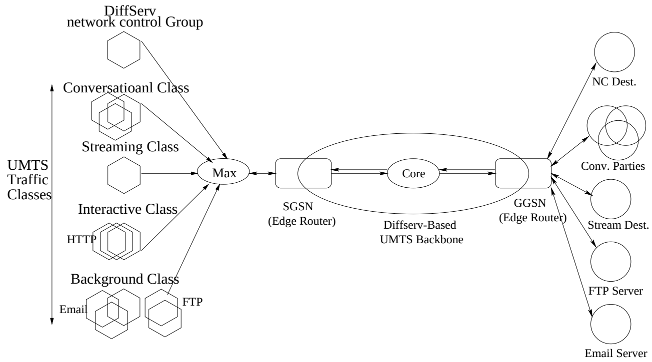

Fig. 6. The Simulation Network Schematic.

to its assigned rate, and to measure the packet loss ratio. The model only considers the backbone network portion of the datapath and consists of a set of traffic sources, a simple core network, and destination nodes (Figure 6). The core network consists of an SGSN, a core node and a GGSN. The SGSN and GGSN nodes also work as DiffServ edge routers. Therefore, they do IP encapsulation and perform the mapping function. The core node only applies the PHB associated with the codepoint marked in each packet header.

Traffic Model: The traffic sources chosen are a combination of all the UMTS traffic classes, thus the model considers UMTS conversational, streaming, interactive and background classes, named groups 1 to 4. The mapping is performed according to the discussion of section IV-A (Table II). The first group (G0) is the NC PHB group. Group one (G1) traffic is generated from the aggregation of three sources, demanding a delay requirement similar to conversational traffic class. Group two (G2) represents the streaming traffic. The interactive traffic (G3) comes from the aggregation of five instances of HTTP web browsing. Finally, the background traffic (G4) consists of two instances of heavy email and two FTP flows. The traffic models of G3 and G4 are based on OPNET built-in models and the traffic models of G1 and G2 are generated with exponential interarrival time and packet size.

Structure of the Routers: Figure 7 depicts the structure of every router in the model which performs packet forwarding according to the packet's DSCP. This router supports five PHB aggregates. Incoming packets are classified first. In an edge node, the QoS of a packet is mapped to a proper PHB and then the packet is encapsulated, having the associated DSCP in its header. In a core router, the packet only needs to be classified based on its DSCP. Then, it passes through a rate estimator. If accepted, the packet is enqueued, otherwise, it is forwarded to a dropper block. The scheduler then forwards the packet to the next hop using the DDB-FFQ service discipline.

There are five datapaths for the five PHBs supported by the router. The NC traffic is the first group and has a very simple datapath. The second group is an EF PHB. This traffic is metered first and non-conforming traffic is dropped, but the conforming traffic has a guaranteed delay bound. The third group is an AF11 group. It has a dedicated buffer that accepts traffic packets based on a tail dropping scheme. This traffic can be video streaming traffic. The fourth group is an AF21 group that first meters the traffic flow. If the traffic is non conforming it is remarked for an AF22 (with a lower drop precedence). The last group is best effort traffic. Any unrecognized packet is also put into this group. Groups three and four use a shared-memory buffer for their queue and the RED on a shared-memory buffer scheme is proposed as the dropping algorithm.

A bandwidth and a delay bound are assigned to each traffic group by using the corresponding assigned rate and the FFQ rate of the DDB-FFQ scheduler, respectively. Since the scheduler is a non-preemptive scheduler, the delay bounds are affected by the maximum packet size of even the lowest priority traffic [17]. Therefore, it is necessary for the edge routers to perform segmentation and re-assembly on large packets, depending on the system delay bounds. Here the delay bounds are set by considering a 90% packet size as the maximum size, which gives 90% packet delay bounds.

Simulation Results: The independency of the service received by each traffic group is demonstrated by observ-

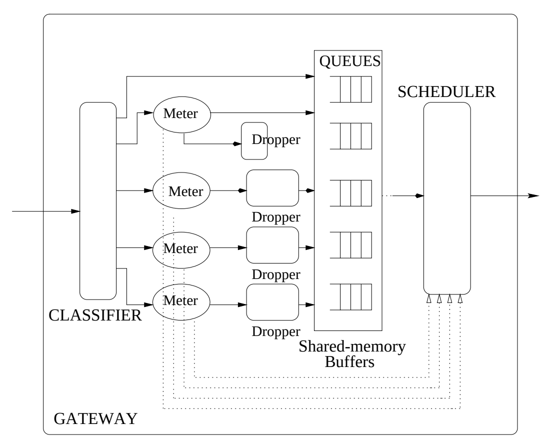

Fig. 7. The Structure of the DiffServ-based Backbone Router.

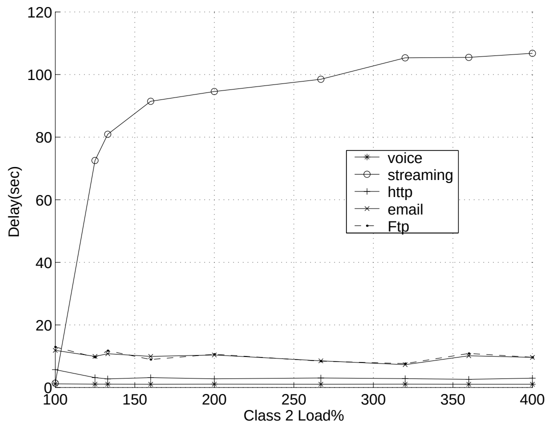

Fig. 8. Delay vs. G2 load%.

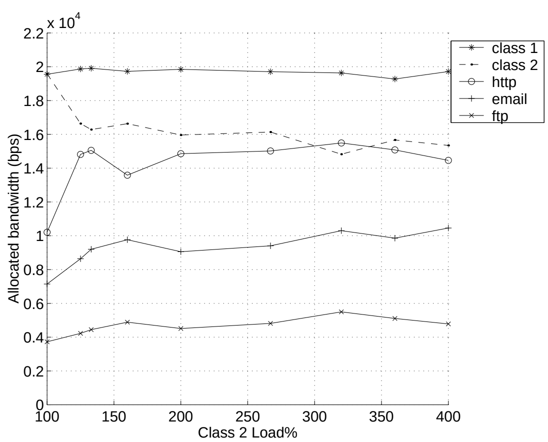

Fig. 9. Allocated bandwidth allocated vs. G2 load%.

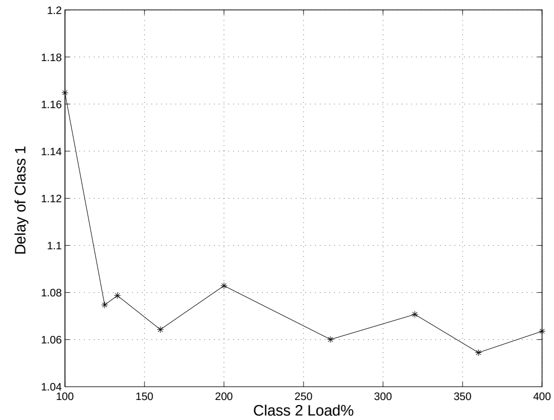

Fig. 10. Delay of traffic G1.

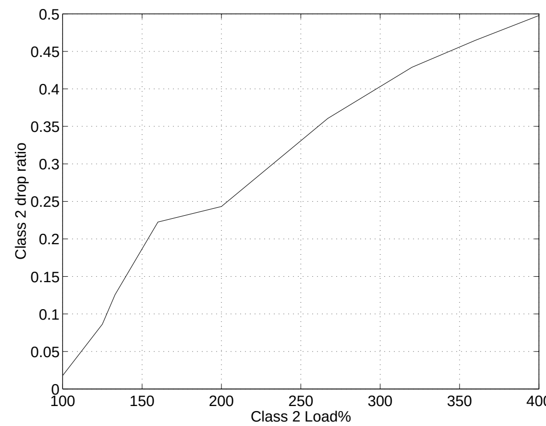

Fig. 11. Drop ratio of G2 vs. its load%.

ing the delay and allocated bandwidth of each of them when the rate of G2 traffic varies from 100% to 400% of its preset rate. In a DiffServ router the overload of misbehaving real-time traffic is dropped and packets are only accepted when the rate falls below a percentage of the assigned rate, as it does for the EF PHB. However, for other kinds of traffic, for example streaming aggregates, it might be useful to accept packets until the corresponding queue is full, while insuring this traffic does not degrade the service of the other streams. The size of this queue should be set to provide an acceptable maximum delay.

Figure 8 shows the delay observed by each application and Figure 9 illustrates their actual bandwidth assignments when the aggregate rate of G2 traffic varies from 100% to 400% of its base rate. The figures show that the PHB treatment of each aggregate is mostly independent of the behaviors of aggregate group G2, while they are offered a distinct range of QoS. Although the rate of G2 traffic is changing widely, the delay bound and bandwidth allocated to other traffic change within an acceptable range of their assigned QoS values. Figure 8 illustrates that the delay of each traffic groups does not degrade, other than the misbehaving streaming traffic. Figure 9 shows an initial decrease of bandwidth assignment for the streaming traffic, and correspondingly, a slight bandwidth increase for interactive and background aggregates. This is a direct result of the method of adjustment proposed for the DDB-FFQ scheduling scheme [17], which modifies the bandwidth according to the minimum FFQ-rate of the system.

Figure 10 depicts the delay observed by the conversational aggregate. It shows that initially the delay drops until the DDB-FFQ link-sharing mechanism comes into effect and then it stays constant.

Figure 11 shows the packet drop ratio of aggregate group G2 which increases when its rate increases. The conversational traffic observes no packet dropping and aggregate groups G3 and G4 observe very small packet drop rates due to their RED mechanism.

VI. Conclusion

In this paper, we have presented a basic methodology for deploying DiffServ in UMTS/GPRS backbone networks to support the core network's QoS requirements. The study addressed two implementation related issues: the structure of a DiffServ-aware UMTS backbone router and QoS mapping related issues. The router utilizes the DiffServ functional elements with novel algorithms to build five datapaths for five different PHBs, incorporating all the UMTS traffic classes, plus a management/signaling PHB. The scheduling block is based on a novel DDB-FFQ service discipline that is capable of decoupled delay and bandwidth assignments. This element is the key factor in achieving the desirable router behavior associated with each group. A simulation model of the backbone network has been used to evaluate the overall performance of the system when treating a packet according to its PHB group, mapped from its UMTS traffic class. We have presented simulation results on the edge-to-edge delay, bandwidth and loss ratio to show the effectiveness of the prototype in providing service differentiation in the backbone network.

References

[1] ETSI GSM 03.60 V6.0.0, "General packet radio service; service description; stage 2," March 1998.

[2] G. Brasche and B. Walke, "Concepts, services and protocols of the new GSM phase 2+ general packet radio service," *IEEE Commun. Mag.*, vol. 35, pp. 94–104, August 1997.

[3] H. Karranen, A. Ahtiainen, l. Laitinen, S. Naghian, and V. Niemi, *UMTS Networks, Architecture, Mobility and Services*, John Wiley & Sons, 2001.

[4] ETSI TS 123 107 V3.5.0, "UMTS QoS concept and architecture," January 2002.

[5] L.B. Milstein, "Wideband code division multiple access," *IEEE Trans. Selected Areas Comm.*, vol. 18, pp. 1344–1354, Aug. 2000.

[6] ETSI 101 146 V3.0.0, "UMTS terrestrial radio access (UTRA): concept evaluation," December 1997.

[7] E. Dahlman, P. Beming, J. Knutsson, F. Ovesjo, M. Persson, and C. Roobol, "WCDMA- the radio interface for future mobile multimedia communications," *IEEE Trans. Vehicular Technology*, vol. 47, pp. 1105–1118, November 1998.

[8] "IETF DiffServ working group," http://www.ietf.org/html.charters/diffserv-charter.html.

[9] "IETF IntServ working group," http://www.ietf.org/html.charters/intserv-charter.html.

[10] G. Priggouris, S. Hadjiefthymiades, and l. Merakos, "Supporting IP QoS in the general packet radio service," *IEEE Network*, vol. 14, pp. 8–17, September/October 2000.

[11] M. Puuskari, "Quality of service framework in GPRS and evolution towards UMTS," *in Proc. 3rd Euro. Pres. Mobile Commun. Conf.*, Paris, France, 1999.

[12] K. Venekn, D. D. Vleeschauwer, and J. D.Vriendt, "Designing a diffserv capable IP-backbone for the UTRAN," *in Proc. 3G Mobile Commun. Technologies Conf.*, pp. 47–52, London, UK, March 2001.

[13] S. Maniatis, C. Grecas, and I. Verieris, "End-to-end quality of service issues over next generation mobile internet," *Symp. Commun. & Vehicular Technology*, pp. 150–154, Leuven, Belgium, Oct. 2000.

[14] Y. Bernet, S. Blake, D. Grossman, and A. Smith, "An informal management model for DiffServ routers," *IETF Internet Draft*, February 2001.

[15] S. Blake, D. Black, M. Carlson, E. Davies, W. Weiss, and Z. Wang, "An architecture for differentiated services," *RFC 2475*, December 1998.

[16] F. Agharebparast and V.C.M. Leung, "Improving the performance of RED deployment on a class based queue with shared buffer," *Proc. IEEE GlobeCom'01*, vol. 4, pp. 2363 –2367, San Antonio, Tx, November 2001.

[17] F. Agharebparast and V.C.M. Leung, "Efficient fair queueing with decoupled delay-bandwidth guarantees," *Proc. IEEE GlobeCom'01*, vol. 4, pp. 2601 –2605, San Antonio, Tx, November 2001.

[18] F. Yu and V.C.M. Leung, "Connection admission control for pcs-to-internet protocol internetworking," *in Int'l Journal of Wireless Information Networks*, Dec. 2001.

[19] Y. Lina, H.C.H. Rao, and I. Clamtac, "General packet radio service: architecture, interface and development," *Wireless Communications and Mobile Computing*, vol. 1, pp. 77–92, John Wiley & Sons Journal, January-March 2001.

[20] S. Floyd and V. Jacobson, "Random early detection gateway for congestion avoidance," *IEEE/ACM Trans. Networking*, vol. 1, pp. 397–413, August 1993.

[21] H. Zhang, "Service disciplines for guaranteed performance service in packet switching networks," *Proc. IEEE*, vol. 83, pp. 1374–1396, October 1995.

[22] I. Stoica, H. Zhang, and T.S.E. NG, "A hierarchical fair service curve algorithm for link-sharing, real-time, and priority services," *IEEE/ACM Trans. Networking*, vol. 8, pp. 185 –199, April 2000.

[23] S. Floyd and V. Jacobson, "Link-sharing and resource management models for packet networks," *IEEE/ACM Trans. Networking*, vol. 3, pp. 365–386, August 1995.

[24] D. Stiliadis and A. Verma, "Efficient fair queueing algorithms for packet switched networks," *IEEE/ACM Trans. Networking*, vol. 6, pp. 175–185, April 1998.

[25] D. Stiliadis and A. Verma, "Rate-proportional servers: a design methodology for fair queueing algorithms," *IEEE/ACM Trans. Networking*, vol. 6, pp. 164–174, April 1998.

Farshid Agharebparast (farshid@ece.ubc.ca) received the M.A.Sc. degree from the University of British Columbia, Vancouver, Canada in 2001 and the B.Sc. degree from the Isfahan University of Technology, Iran in 1995, all in Electrical Engineering. He is now a Ph.D. student at the University of British Columbia. He is a recipient of numerous awards including the NSERC postgraduate scholarship. His main research interest is QoS in computer and wireless communications.

Victor C.M. Leung (vleung@ece.ubc.ca) received the B.A.Sc. (Hons.) degree in electrical engineering from the University of British Columbia (U.B.C.) in 1977, and was awarded the APEBC Gold Medal as the head of the graduating class in the Faculty of Applied Science. He attended graduate school at U.B.C. on a Natural Sciences and Engineering Research Council Postgraduate Scholarship and obtained the Ph.D. degree in electrical engineering in 1981. From 1981 to 1987, Dr. Leung was a Senior Member of Technical Staff at Microtel Pacific Research Ltd. (later renamed MPR Teltech Ltd.), specializing in the planning, design and analysis of satellite communication systems. He also held a part-time position as Visiting Assistant Professor at Simon Fraser University in 1986 and 1987. In 1988, he was a Lecturer in the Department of Electronics at the Chinese University of Hong Kong. He joined the Department of Electrical Engineering at U.B.C. in 1989, where he is a Professor, holder of the TELUS Mobility Industrial Research Chair in Advanced Telecommunications Engineering, and a member of the Centre for Integrated Computer Systems Research. He is also a project leader and a member of the Board of Directors in the Canadian Institute for Telecommunications Research, a Network of Centres of Excellence funded by the Canadian Government. His research interests are in the areas of architectural and protocol design and performance analysis for computer and telecommunication networks, with applications in satellite, mobile, personal communications and high speed networks. Dr. Leung is a senior member of IEEE and a voting member of ACM. He is an editor of the IEEE Journal on Selected Areas in Communications - Wireless Communications Series, and an associate editor of the IEEE Transactions on Vehicular Technology.