

Convergence of UMTS and Internet Services for End-to-end Quality of Service Support

S. I. Maniatis, E. G. Nikolouzou, I. S. Venieris

National Technical University of Athens, 9 Heron Polytechniou, 15773 Athens, Greece

Ph.: +30 1 7722424, Fax: +30 1 7722534, email: sotos@telecom.ntua.gr ,
enik@telecom.ntua.gr , ivenieri@cc.ece.ntua.gr

ABSTRACT

Internet evolution delineated through the last years has urged the wireless network community to adopt the support of the IP protocol. The deployment of IP multimedia services with guaranteed quality of service in both the access and core network is one of the primary requirements for 3G wireless networks. This paper focuses on issues regarding interoperability with external, "wired" QoS-enabled IP networks, in order to support end-to-end services over both the wireless and the wired environments. The paper first briefly describes the RCL network architecture that supports QoS based on the concept of Differentiated Services, focusing on the supported network services. Subsequently, the UMTS QoS architecture and traffic classes are presented. Last, the interoperability aspects between the two networks are discussed, mainly at the QoS management level, supported by simulations for two traffic classes. The main outcome of the paper is that interoperability can be achieved only if both networks offer services with common characteristics and QoS requirements.

1. INTRODUCTION

Internet has had an overwhelming effect on the way people interact, communicate, and work over the past few years. The Internet is based on the simplicity of the Internet Protocol (IP), which only offers best-effort services. The lack of any Quality of Service (QoS) mechanisms has urged the IETF community to seek for QoS support through network layer mechanisms. The most apparent mechanisms are the Integrated Services (IntServ) [1] and the Differentiated Services (DiffServ) [1]. The next generation Internet will most probably exploit these mechanisms to provide services with QoS characteristics to the users.

The third generation (3G) wireless networks support IP, but they also promise guaranteed quality for IP multimedia services, not only over the air interface, but also in the access and core network. The Universal Mobile Telecommunications System (UMTS) Release 99 onwards specifies mechanisms for QoS support within all the aforementioned parts. Another significant contribution of this release is the specification of the UMTS QoS classes, or traffic classes [2]. The traffic classes are intended for specific applications that produce traffic that exhibits a well-known behavior (e.g. Voice over IP).

Having in mind the picture of the all-IP future network (wired and wireless), the need of the

appropriate interworking between UMTS and external, wired IP networks is considered essential. Although the IP protocol assures interoperability at the network level, specific mechanisms must be provided to offer interoperability at the QoS level, too. The need for interoperability with external IP networks has also been foreseen within the 3G standardization forums, and standardization about this issue has been progressing in the Third Generation Partnership Project (3GPP) [3].

This paper addresses the interoperability issue between 3G wireless networks and external wired ones, mainly at the QoS management level. We first briefly describe the RCL architecture, a novel architecture for the support of QoS for the next generation wired Internet. The architecture defines five network services that aim to support applications with different traffic behavior. Moreover, it introduces the Resource Control Layer (RCL), which offers a scalable approach for the management of the resources in the network, per flow policy-based admission control, configuration of edge routers, monitoring of the network and interaction with host applications and end-users.

The definition of specific network services and traffic classes in both networks is the first step towards the proper interworking between them. The main focus of the paper is the mapping between the network services and traffic classes in the two worlds. We have to stress here that it is not the aim of this paper to propose the RCL architecture as the counterpart of UMTS in the wired world, but to use it as reference architecture for our purposes. Moreover, interoperability has also to be provided at the signaling level, to allow the setup of QoS in both networks, leading to end-to-end support. Both networks provide control-plane modules that can be used to integrate end-to-end signaling. These modules and the interworking issues are briefly investigated in this paper. However, it is still open whether dynamic signaling or a more static mechanism, based on Service Level Agreements (SLA), or a combination of both is going to be used for QoS negotiation.

In the rest of the paper, section II presents the RCL architecture and the entities that compose the RCL. In section III, we focus on the concept of Network Services and Traffic Classes in the RCL architecture. In section IV, the UMTS QoS architecture and the QoS classes are described. The interworking aspects are then addressed in section V, and simulation results are presented, proving the gain that is acquired in the provision of end-to-end QoS for voice and video traffic. Finally, the conclusions are given.

2. THE RCL ARCHITECTURE

The aim of this section is to give a short introduction of the wired network architecture, namely RCL architecture. The reader is referred to [11] for a thorough presentation. The RCL architecture consists of two functional areas: the data plane that is responsible for transmitting IP packets, and an overlay control plane, namely the Resource Control Layer (RCL) that is based on the Bandwidth Broker (BB) concept [4]. Although the classical BB architecture proposes a concentrated approach where one BB is responsible for an administrative domain, RCL is designed as a distributed BB, to overcome scalability problems.

As depicted in Fig. 1, the three key components of the RCL are: the Resource Control Agent (RCA), the Admission Control Agent (ACA) and the End-User Application Toolkit (EAT).

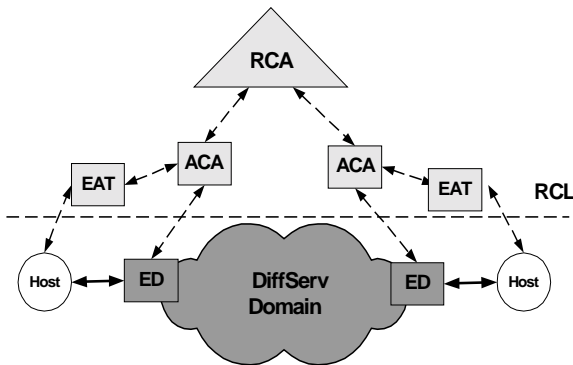


Fig.1: RCL Structure and main interactions

The *Resource Control Agent* represents the ultimate principle of the domain concerning the management of network resources. In order to simplify the task of the RCA to handle the resources efficiently, the network is divided into sub-areas that form a tree structure, where each sub-area is assigned initially its own resources, according to traffic load forecasts and results retrieved by measurements. The initial assignment of resources to sub-areas may not reflect the actual traffic load, so an intelligent load-balancing algorithm is utilized for the re-distribution of resources among them, as described in detail in [5]. The RCA is therefore divided in logical entities (Resource Pools, RP). Each RP manages the resources of the respective sub-area. The root of the tree is in charge of the available resources in the whole network, while each leaf of the tree structure (Resource Pool Leaf, RPL) is associated to one Admission Control Agent, which is in turn associated to one edge device of the network.

The *Admission Control Agent* mainly performs user authentication and authorization, reservation handling, and admission control. User authentication is initially performed when contacting the ACA, and then a reservation request, specifying the traffic requirements of the new flow, can be placed. Subsequently, user authorization for that type of request is checked and a reservation state is created. Policing and admission control are made only at the edges of the network, therefore the corresponding ingress and egress points (ingress-egress ACAs) of the flow are identified and the local resources (in the RPL) are checked to ensure that the new flow can be accommodated. The core network is provisioned in order to ensure that once the admission control at the edges succeeds, no bottleneck will be

created in the core network. Upon a successful reservation request, the corresponding ACAs consequently configure the edge routers appropriately to accommodate the new flow.

Reservation requests are forwarded to the ACAs from the *End-User Application Toolkit*, which mediates between end-users or applications and the network. The EAT interacts with the ACA to be aware of the available network services. A reservation request specifies the flow identifiers, the selected network service and the traffic profile for it. Special support is foreseen for legacy applications as well as for end users that are not aware of traffic description details, through the concept of Application Profiles.

3. NETWORK SERVICES AND TRAFFIC CLASSES

The RCL architecture provides quality of service guarantees to the users by offering a number of transport options for user IP traffic. These transport options are called Network Services, and are constructed by applying traffic conditioning to create aggregates that experience a known Per-Hop-Behavior (PHB) at each node within the DS domain.

In the RCL architecture five Network Services (NS) have been defined in order to provide service guarantees to different applications: Premium Constant Bit Rate (PCBR), Premium Variable Bit Rate (PVBR), Premium MultiMedia (PMM), Premium Mission Critical (PMC) and Best Effort (BE) [6]. Applications can be grouped into this relatively small number of services, with applications in each service having similar requirements on the network in order to perform effectively. In other words, flows belonging to each NS have similar traffic characteristics.

The implementation of those NSs is realized with the use of some network's mechanisms, which are the Traffic Classes (TCLs). A TCL is defined as a composition of a set of admission control rules, a set of traffic conditioning rules and a PHB. In the RCL architecture five TCLs are introduced: TCL1, TCL2, TCL3, TCL4 and TCL5, which correspond to PCBR, PVBR, PMM, PMC and BE, respectively. Each TCL maintains a separate queue at the router output ports. The scheduling mechanism actually implemented is a combination of the Priority Queuing (PQ) and Weighted-Fair Queuing (WFQ) [7], which is called PQWFQ (Fig. 2). A queue is dedicated for TCL1, which has strict priority over the others, while the other TCLs are scheduled with the WFQ. A WFQ weight is assigned to each TCL and each queue is managed by different queuing strategy (Drop-Tail, Random Early Detection (RED), Weighted-RED (WRED) [8, 9]).

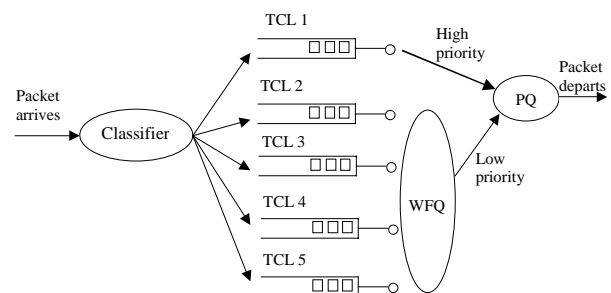


Fig. 2: Design of router output port

The WFQ weights are configured based on the sharing of each link's resources among the different traffic classes. In this way, the maximum amount of traffic for each TCL allowed to transit onto a link is calculated. The routers are configured with those weights during the start-up procedure. The weights are somehow static, since they are not updated dynamically. According to the WFQ weights, the Admission Control (AC) rate limits for TCLs at ingress EDs are also initially set. The AC procedure [11] is intended to restrict traffic, in order to avoid that a bottleneck arises in the edge-link (i.e. the link between the network and the ingress or egress ED), as well as in any of the internal-links. The AC rate limits may be changed dynamically as a result of resource pool operations.

3.1 Traffic Classes Implementation

The PCBR is for constant and variable bit rate applications with low bandwidth flows. These applications have strict delay and delay variation requirements. Therefore, the delay should be less than 150msec for the 99% of in-profile packets, while the packet loss should be less than 10-8. In addition, flows should have small packets, so as not to provoke long transmission delays. IP telephony is the basic application supported by this NS. Since peak rate allocation is suitable for these low bandwidth flows, the traffic conditioning mechanism of TCL1 is realized with the use of a token bucket (TB) as meter and dropper, as depicted in TABLE I. The value of x_1 lies in the range of {1,5}; a possible value could be $x_1 = 1$, while a larger value would allow a small amount of burstiness. Packets that do not find enough tokens in the bucket are dropped. The traffic conditioning mechanism is realized in the routers with the use of the Committed Access Rate (CAR) mechanism. Packets of TCL1 are enqueued in a single FIFO drop-tail queue.

TABLE I

CAR profiles	
TCL1 (Single Token Bucket)	$r = \text{Peak Rate (PR)}$ of the flow $b = x_1 * M_1 = \text{Bucket Size for PR, BSP}$ M_1 : maximum allowed packet size (<256 B)
TCL2 (Dual Token Bucket)	$r1 = \text{Sustainable Rate (SR)}$ of the flow $b1 = \text{Bucket Size for SR in bytes (BSS)}$ $r2 = \text{Peak Rate (PR)}$ of the flow $b2 = x_2 * M_2 = \text{Bucket Size for PR, BSP}$ M_2 : maximum allowed packet size (<1000 B)
TCL3 (Single Token Bucket)	$r = \text{Sustainable Rate (SR)}$ of the flow $b = \text{Bucket Size for SR in bytes (BSS)}$ $M3$: maximum allowed packet size (<1500 B)
TCL4 (Dual Token Bucket)	$r1 = \text{Sustainable Rate (SR)}$ of the flow $b1 = \text{Bucket Size for SR in bytes (BSS)}$ $r2 = \text{Peak Rate (PR)}$ of the flow $b2 = \text{Bucket Size for PR, BSP}$ $M4$: maximum allowed packet size (<1500 B)

The PVBR is appropriate for unresponsive VBR sources with medium to high bandwidth requirements. They require low delay, delay variation and packet loss, even though greater than PCBR. An end-to-end delay less than 250msec and a packet loss less than 10-6 are guaranteed. Video and teleconferencing are possible applications supported by this NS. Since peak rate allocation is not appropriate for those high bandwidth

flows, a dual TB as meter and dropper is proposed, which controls both the peak and sustainable rate (TABLE I). The value of x_2 is in the range of {1,5}. The depth of the first bucket defines the burstiness allowed for the sender's flow (BSS). If there are enough tokens in the first and second TB to accommodate a packet, it is marked as in-profile, otherwise it is dropped. The intention is to limit the sender's traffic in order to be conformant to the profile of the first TB (SR, BSS), while the second TB (PR, BSP) allows an amount of burstiness. Packets of TCL2 are also enqueued in a single FIFO drop-tail queue.

The PMM is expected to carry a mixture of TCP and non-TCP traffic. These flows require a minimum bandwidth, which must be delivered at a high probability. Independently of the transport protocol, flows are expected to implement some kind of congestion control mechanism and their aggressiveness should be similar to the one of TCP, assuming that they are roughly TCP-friendly. This NS delivers applications such as video/audio streaming and ftp. The drop probability for in-profile packets should be low (less than 10-3), while out-of-profile packets do not experience any QoS guarantees. A single TB as a meter and marker is proposed, which polices the sustainable rate (TABLE I). Flows conforming to this profile will be marked as in-profile otherwise as out-of-profile. The bucket size (BSS) should be very high to satisfy the bursty nature of TCP traffic. Packets of TCL3 are enqueued in a single FIFO queue, which is managed by WRED with two sets of parameters (minth, maxth, maxp). One set is for the in-profile and the other for out-of-profile packets. Out-of profile packets are not dropped, but marked with a different DSCP.

The PMC service supports mainly transactions and database queries. Thus, flows of the PMC are non-greedy, have short lifetimes, low bandwidth requirements and roughly homogeneous congestion control (TCP). Those applications are guaranteed with minimal loss probabilities for in-profile packets (less than 10-4) and small queuing delays. For TCL4, WRED with two sets of parameters is used in order to discriminate out-of-profile packets against in-profile packets. Regarding traffic conditioning, a dual TB is proposed (TABLE I), where the first bucket works as a sustainable rate policer and the second bucket works as a peak rate policer. The value of x_4 for TCL4 is fixed in the range of {1,5}. A packet that requires fewer tokens than available in the first and second TB is marked as in-profile, otherwise is marked out-of-profile and forwarded into the net. The SR should be small in order to disable greedy sources to transmit in-packets with a high rate into the network while the BSS should be large enough to allow several back-to-back packets to enter the network without being marked as out-of-profile.

Finally, the BE requires no quality of service guarantees, and best effort packets are enqueued in a single FIFO queue managed by the RED algorithm.

4. UMTS QOS ARCHITECTURE AND TRAFFIC CLASSES

3GPP standards [3] propose a layered architecture (depicted in Fig. 3) for the support of end-to-end QoS for the packet domain, through interaction of Bearer Services (BS) established between UMTS modules at

different layers. Each bearer service mainly specifies the control signaling, user plane transport and QoS management functionality, among others.

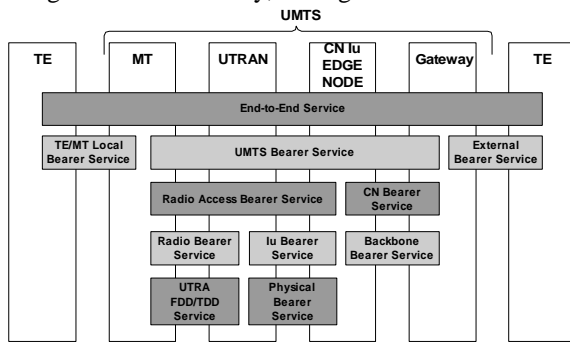


Fig.3: UMTS QoS Architecture

The end-to-end service may be conveyed over several networks (not only UMTS). The External Bearer Service deals with the interoperability aspects with external IP Bearers. For this purpose, the Gateway (GGSN) is supplied with the IP BS Manager module that is responsible for controlling the external IP bearer, and with the Translation function to convert between external and internal primitives (like service attributes). Moreover, the GGSN offers an IP Policy Control point that coordinates the events in the application layer with resource management in the IP bearer.

Respectively to GGSN, the User Equipment (UE), that is TE and MT in Fig. 3, may also be supplied with an IP BS Manager and a translation function, performing respective operations. However, this is optional, as it depends on the participation of UE in the IP QoS procedures (e.g. signaling). In case the UE does not provide these functions, QoS in the UMTS network is provided with mere UMTS procedures, i.e. through PDP context activation, while the end-to-end IP QoS bearer service towards the remote terminal is controlled by the GGSN only. There is also the option that the UE performs IP QoS operations, like DS marking or exchanging RSVP signaling messages. The UE relative scenarios are explained in [10] and are not covered here. We rather focus on the interworking operations in the GGSN only.

4.1 UMTS QoS Classes

The UMTS specifications define four QoS classes: Conversational, Streaming, Interactive and Background. The distinguishing factor among these classes is mainly the delay sensitivity. The Conversational class is the most sensitive, while Background is the least sensitive. Conversational and Streaming classes are intended for real time traffic. They both preserve time relation (variation) between information elements of the stream, but Conversational has more strict and low delay requirements. Example applications are IP telephony for the former and streaming video for the latter.

For the Interactive and Background classes, transfer delay is not the major factor. Instead, they both preserve the payload content. Interactive follows a request-response pattern and defines three priorities to differentiate between bearer qualities, while it does not provide explicit quality guarantees. Background's main characteristic is that the destination does not expect the data within a certain time. Example applications are

Web traffic for Interactive and download of emails for Background.

Ongoing work within 3GPP defines the attributes that characterize the classes in the UMTS BS and the other bearers. Only UMTS BS attributes are relevant to the context of this paper, as only these are mapped to the RCL architecture traffic attributes. The UMTS BS attributes are the maximum bit rate, guaranteed bit rate, delivery order, maximum SDU size, SDU format information, SDU error ratio, residual bit error ratio, delivery of erroneous SDUs, transfer delay, traffic handling priority, and allocation/ retention priority.

The specifications [2] explain in detail each attribute, as well as their role for each QoS class. However, some important information is also cited hereafter. Regarding the maximum bit rate, traffic is conformant to that, as long as it follows a token bucket algorithm where token rate equals the maximum bit rate and bucket size equals the maximum SDU size. Guaranteed bit rate differs from the maximum one as it sets the bucket size to $k \cdot [\text{maximum SDU size}]$, to cater for burstiness of sources. The SDU error ratio refers only to conforming traffic and is defined as the fraction of erroneous or lost SDUs. Finally, the transfer delay is the maximum delay for 95th percentile of the distribution of delay for all delivered SDUs and is meaningful for non-bursty sources.

5. END-TO-END QOS SERVICE

The end-to-end QoS service requires interworking between UMTS and the external IP network most importantly at the QoS management level, but also at the user and control plane. The UMTS Bearer Services can be mapped to the network services implemented in the RCL architecture, providing end-to-end quality of service guarantees. Table II summarizes the proposed mapping.

From Table II, we see that the mapping of Network Services can be easily accomplished. Conversational and Streaming classes are associated one-to-one to PCBR and PVBR, respectively. Interactive can be mapped to either PMM or PMC, according to the traffic handling priority. Although the Interactive class does not take into account the bit rate for admission control and policing, it requires a specific packet loss. This QoS requirement can only be satisfied by using either PMC or PMM, and definitely not by the BE. Background can be mapped to either PMC or BE, depending on the packet error loss ratio requirements.

Coming to user plane interworking, it can be provided in the IP layer. The IP Policy control in UMTS and the appropriate setup of CAR in edge routers by ACA in the RCL network provide the commitment of the networks to the requested traffic attributes, after successful admission control in both domains.

Admission control is accomplished through control plane signaling. Control plane interworking can be offered by interfacing between the IP BS Manager and the Translation function in the UMTS domain, and the ACA and RCA components of RCL. The way of interworking mainly depends on the involvement of the UE at the QoS procedures, which is not examined in this paper.

In case that the UE does not provide any QoS functionality, interworking is accomplished through

GGSN only. QoS negotiation between UMTS and the RCL network can be provided dynamically through the exchange of signaling messages between the IP BS Manager and the ACA, as they both provide such functionality. On the other hand, scalability reasons, due to the possible large number of IP flows, could direct to a static solution, based on Service Level Agreements between the operators of both the UMTS and the RCL network. A compromising solution suggests that signaling is exchanged only for real-time traffic (Conversational and Streaming classes), in order to guarantee the low delay requirements.

Table II

	UMTS Classes	Wired Architecture NS	
	Conversational	PCBR	
Characteristics			
Maximum bitrate (kbps)	< 2048	Maximum per flow 200kbps	
Maximum packet size (bytes)	<=1500 or 1502	256	
Packet error ratio	$10^{-2}, 7 \cdot 10^{-3}, 10^{-3}, 10^{-4}, 10^{-5}$	$< 10^{-8}$	
Transfer Delay(msec)	100 maximum	150 maximum	
	Streaming	PVBR	
Characteristics			
Maximum bitrate (kbps)	< 2048	Maximum per flow 1000kbps	
Maximum packet size (bytes)	<=1500 or 1502	1000	
Packet error ratio	$10^{-1}, 10^{-2}, 7 \cdot 10^{-3}, 10^{-3}, 10^{-4}, 10^{-5}$	$< 10^{-6}$	
Transfer Delay(msec)	250 maximum	250 maximum	
	Interactive	PMM	PMC
Characteristics			
Maximum bitrate (kbps)	< 2048 - overhead	Maximum per flow 250kbps	Maximum per flow 50kbps
Maximum packet size (bytes)	<=1500 or 1502	1500	1500
Packet error ratio	$10^{-3}, 10^{-4}, 10^{-6}$	$< 10^{-3}$	$< 10^{-4}$
Traffic Handling Priority	1,2,3	1	2,3
	Background	PMC	BE
Characteristics			
Maximum bitrate (kbps)	< 2048 - overhead	Maximum per flow 50kbps	
Maximum packet size (bytes)	<=1500 or 1502	1500	
Packet error ratio	$10^{-3}, 10^{-4}, 10^{-6}$	$< 10^{-4}$	

5.1 Simulation Scenario and Results

We conducted some simulation scenarios to verify the acquired gain when UMTS traffic is properly associated with services of the RCL network, as opposed to the case when UMTS traffic is passing over the current “best-effort” Internet infrastructure, as delineated in Fig.4. The Opnet 7.0 simulation tool is used for the simulations.

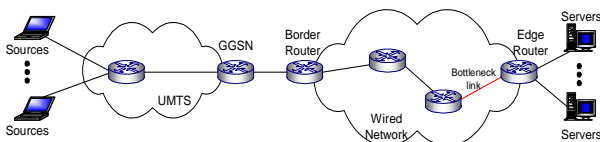


Fig. 4. The simulation network

We have to stress here that the UMTS infrastructure is not fully simulated, especially the radio interface and the access network. This is not inconsistent with the concept of the UMTS architecture, which specifies that the access network and the core network are independent. Therefore, we only simulate the packet-

based core network, as an IP-based network that implements specific QoS mechanisms to cope with the provision of the UMTS QoS classes. To be more specific, we simulate the Core Network Bearer Service between the SGSN and GGSN. SGSN and GGSN are regular IP routers that implement traffic conditioning, packet scheduling, and the appropriate queue management techniques, as it will be described later in this section. Moreover, we assume that the UMTS network is not congested, meaning that traffic experiences no queuing delay at the UMTS routers output interfaces.

The RCL network topology is more complex and is examined in both congested and uncongested cases. In order to simulate congestion in the wired network, the access link between the Edge Router and the core network is considered the bottleneck link, with capacity of 2Mbps, while all the other links are 44Mbps.

The paper focuses on the real-time traffic classes only. Specific applications in workstations are selected appropriately for these services. Constant bit rate voice flows of 64kbps with constant length packets of 218 bytes using the G.711 encoder scheme are used for the Conversational-PCBR pair. The Streaming-PVBR pair uses video flows with constant length packets of 512 bytes and exponential packet inter-arrival time with mean 0.039 sec, having an average bandwidth of 105Kbps. For Background-BE, flows are used with constant length packets of 1500 bytes, exponential packet inter-arrival time with mean 0.07msec and average bandwidth of 171Kbps.

In the RCL network, the PQ serves only the TCL1 traffic, while WFQ is configured with two queues, one for the TCL2 traffic and one for the TCL5 traffic. The WFQ weights are configured based on the sharing of bandwidth between TCLs. A small percentage of the link capacity is actually dedicated for the TCL1, and here actually restricted to 13% of the link capacity. A 20% percentage of the link is dedicated to TCL2, while TCL5 is configured to occupy the rest (67%) of the link resources. TCL1 and TCL2 do not utilize their whole reserved bandwidth. Their utilization is 0.75 and 0.8 for TCL1 and TCL2 accordingly, which is based on the admission control algorithms and the target performance. Best effort traffic utilizes its reserved resources from 50% to 140%, transmitting 4 to 11 flows, which occupy 684Kbps to 1881Kbps. Obviously, utilizing more than 100% of the reserved resources leads to a congested situation.

Under the RCL architecture, voice and video are subject to admission control with the use of the CAR, which is appropriately configured for each NS. For the total offered load, the CAR is configured for PCBR as: $r = 192\text{kbps}$, $b = 218\text{Bytes}$ and for PVBR (dual token bucket), the first token bucket is configured with $r = 315\text{kbps}$, $b = 5120$ bytes, and the second with $r = 340\text{kbps}$, $b = 1024$ Bytes. The first TB is configured with the average bandwidth of the flows, while the second one with the peak rate of the flows.

The involved traffic mechanisms (CAR, scheduling) are also configured in routers in the UMTS network. The use of the WFQ is proposed as the scheduling algorithm with three queues; one for voice, one for video and one for BE traffic with corresponding weights

13%, 20% and 67%. The CAR is set up for voice and video traffic as a single token bucket (proposed in the appendix of [2]) and configured for voice as $r = 4.29\text{Mbps}$, $b = 218\text{Bytes}$ and for video as $r = 7.04\text{Mbps}$, $b = 512\text{Bytes}$. The decision to have different mechanisms in the UMTS and the RCL architecture was taken on purpose, not only because this is closer to the real situation, but also to show that the end-to-end results are not technically manipulated.

A number of simulations have been conducted, in order to measure the end-to-end delay for voice and video traffic and the packet loss for all kinds of traffic, using different loads of best-effort traffic. Moreover, two simulation scenarios are carried out, where the wired network of Fig. 4 is either the RCL architecture or the best-effort Internet. In all simulations, after a transient period, the value of delay for video and voice is stabilized.

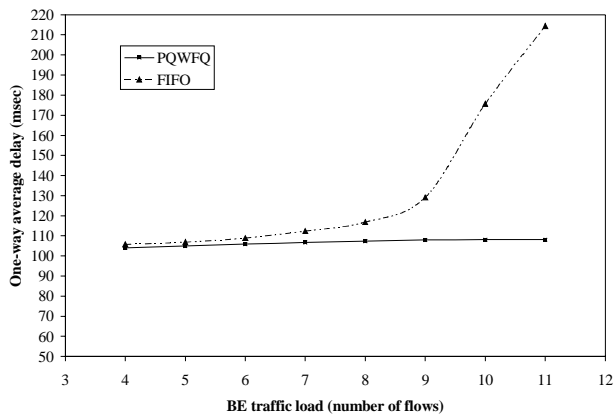


Fig. 5a. Voice end-to-end delay (msec)

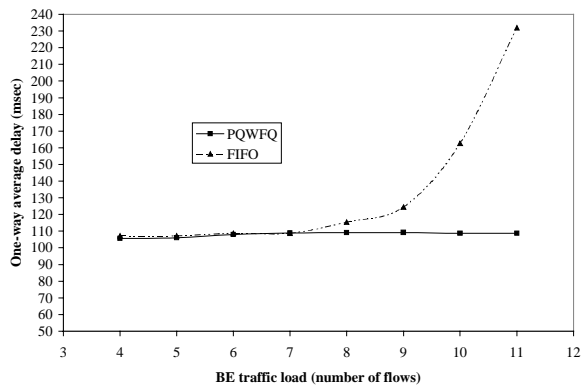


Fig. 5b. Video end-to-end delay (msec)

In Fig. 5a and 5b, the end-to-end delay for voice and video is depicted. It is obvious that the RCL architecture (PQWFQ) outperforms the Internet infrastructure (FIFO), providing an acceptable end-to-end delay for both voice and video. Using the RCL network, the end-to-end delay for voice is 108msec under a heavy load of BE traffic, which is lower than the maximum 150msec that it can tolerate. On the contrary, using the Internet, the end-to-end delay increases continually and reaches an unacceptable value of 214.4 msec.

The same behavior is also observed for the video traffic, as depicted in Fig. 5b. Under the RCL architecture, the end-to-end delay rises up to an acceptable value of 109.19msec, while the Internet architecture provides a delay of 231.99msec. We have

to stress here that also FIFO achieves a lower than the maximum value of end-to-end delay for TCL2 (250 msec), but this is accomplished with an unacceptable packet loss rate, as more than 20% of TCL2 packets are dropped under FIFO (Fig. 6b).

As resulting from the simulations, the packet loss for voice and video traffic (Fig. 6a and 6b) is less than 10^{-5} with the use of the RCL architecture, while it is really high under the Internet architecture, where no traffic discrimination is considered. This is justified from the fact that Internet does not guarantee any bandwidth to neither application, so when the BE traffic increases, it occupies a greater portion of the bandwidth in the bottleneck link causing a higher packet loss to the voice and video traffic. On the other hand, the packet loss for BE traffic is less under Internet than the RCL architecture, since under Internet the excess BE traffic storms the bottleneck link.

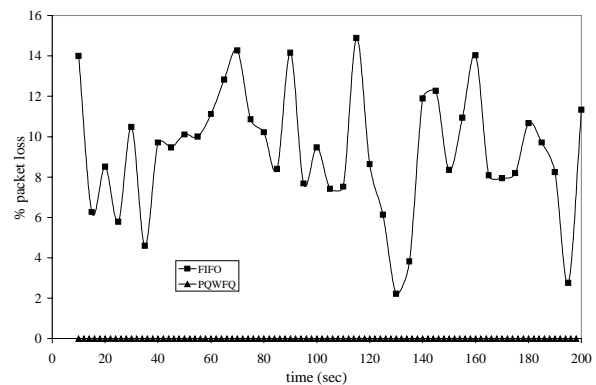


Fig. 6a. Packet loss for Voice

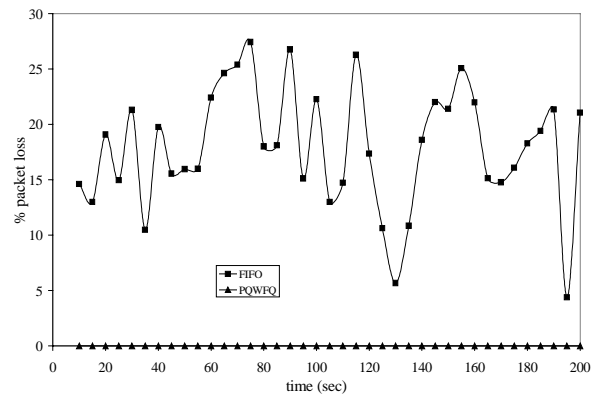


Fig. 6b. Packet loss for Video

6. CONCLUSIONS AND FUTURE WORK

In this paper we discussed the implementation of end-to-end QoS services over the forthcoming UMTS infrastructure and an external IP network, the RCL architecture. We identified the need to have Network Services and Traffic Classes at both networks that constitute the end-to-end path, as well as interworking at the QoS management level. A possible mapping between the traffic classes of the two worlds has then been presented, and simulations for the real-time classes have shown that the specific end-to-end quality requirements of NSs are satisfied using the RCL architecture.

As a potential future point of focus, we are mainly interested in the elaboration of the traffic conditioning and scheduling mechanisms in the UMTS core network, not only for the already implemented QoS classes (the real time classes), but also for Interactive and Background.

Moreover, simulations need to be conducted to evaluate the mapping of the Interactive class to the PMM or PMC network service. It would be challenging to study the behavior of the flows and the performance achieved, when the traffic of the priority-based Interactive class is carried over the rate-based PMM or PMC service.

ACKNOWLEDGEMENTS

The work regarding the RCL architecture was performed in the framework of IST Project AQUILA [11] (Adaptive Resource Control of QoS Using an IP-based Layered Architecture - IST-1999-10077) funded in part by the EU. The authors wish to express their gratitude to the other members of the consortium for valuable discussions.

REFERENCES

- [1] IETF web site: <http://www.ietf.org>, DiffServ and IntServ Working Groups.
- [2] 3GPP, "QoS Concept and Architecture," TS 23.107, Release 5, April 2001.
- [3] 3GPP web site: <http://www.3gpp.org>.
- [4] R. Neilson, J. Wheeler, F. Reichmeyer, S. Hares, "A discussion of bandwidth broker requirements for Internet2 Qbone deployment," Internet2 Qbone BB advisory Council, August 1999.
- [5] E. Nikolouzou, G. Politis, P. Sampatakos, I. Venieris, "An adaptive algorithm for resource management in a differentiated services network," IEEE ICC2001, Helsinki, Finland, June 2001.
- [6] Internet Draft, draft-aquila-sls-00.txt, "Definition and usage of SLSs in the AQUILA consortium", May 2001
- [7] J. Bennett, H.Zhang, "Hierarchical packet fair queueing algorithms," SIGCOMM, August 1996, pg.143—156.
- [8] V. Firoiu, M. Borden, "A study of active queue management congestion control," Infocom 2000, March 2000.
- [9] S. Floyd, V. Jacobson, "Random early detection gateways for congestion avoidance," IEEE/ACM Transaction on Networking, August 1993.
- [10] 3GPP, "End-to-End QoS Concept and Architecture," TS 23.207, v.2.0.0, June 2001
- [11] Aquila web site: <http://www.ist-aquila.org>