UMTS access network architecture for multimedia services

Roberto Cusani[1], Filomena Del Sorbo[2], Francesco Delli Priscoli[3], Giuseppe Lombardi[4], Dario Pompili[5]

[1]robby@infocom.uniroma1.it, [2]iofilu@yahoo.it, [3]dellipriscoli@dis.uniroma1.it, [4]lombardi@dis.uniroma1.it, [5]dpompili@tiscalinet.it

Computer Science Department, University of Rome, "La Sapienza", ITALY
Via Buonarroti, 00184 Rome

The authors are listed in alphabetical order

Abstract

In this paper a scheduling mechanism is provided within a proposed UMTS access network architecture, employed to perform a QoS experiment. W-CDMA is the strongest candidate for the air interface technology of UMTS which will comprise both a terrestrial part (T-UMTS) and a satellite part (S-UMTS). Dynamic Resource Scheduling (DRS) is proposed as a framework that will provide QoS provisioning for multimedia traffic in W-CDMA systems.
The proposed resource allocation algorithm, applied with a congestion control mechanism, allows the access network to guarantee the appropriate QoS contract agreed at connection set-up between the users and the UMTS access network operator, obtaining a fair distribution of the available resources between the users and a high link utilization.
OPNET 7.0 has been used to carry out and analyze the experiment results.

Introduction

The convergence of mobile and IP-based technologies has now become the major driving force behind the standardization of third generation (3G) communication systems. The next generation mobile system in Europe will be known as the Universal Mobile Telecommunication System (UMTS). UMTS will build on the success of the second generation mobile network GSM/GPRS to provide circuit and packet switched services to the mobile user; it will provide multimedia services at data rates up to 2 Mbit/s, and will comprise both a terrestrial part (T-UMTS) and a satellite part (S-UMTS). W-CDMA is the strongest candidate for the air interface technology of UMTS. Dynamic Resource Scheduling (DRS) is proposed as a framework that will provide QoS provisioning for multimedia traffic in W-CDMA systems. This scheduling mechanism is provided within a proposed UMTS access network architecture, employed to perform a QoS experiment; OPNET 7.0 has been used to carry out and analyze the experiment results.

Overview of the proposed architecture

In the following, we will focus our attention on a packet-switched traffic control experiment, which foresees Congestion Control and Scheduling mechanisms, in order to guarantee appropriate QoS contracts agreed at connection set-up between the users and the UMTS access network operator. The end-to-end QoS contract can be split in various QoS sub-contracts each holding for one of the cascade sub-networks supporting the connections (of course, the various QoS subcontracts must have QoS contractual conditions such that the end-to-end target QoS is met). So, for instance, the end-to-end QoS contract can be split in two subcontracts: one holding for the Core Network and the other one for the UMTS Access Network, i.e. the UMTS Radio Access Network (URAN).
We will describe a possible functional description of the QoS experiment that could be performed, focusing on the following QoS parameters:

- the maximum transfer delay within the wireless network (i.e. from the time the bit is entered in the wireless network to the time in which it has gone out), which is guaranteed for the bits relevant to a particular connection;
- the number of packets lost.

In our experiment only the URAN QoS constraints will be analyzed, considering the simultaneous presence of two users, each one handling several connections; a dummy traffic is added to the traffic generated by those connections, representing the remaining users' connections in the system: they are treated as a whole and considered as interfering users. It is important to highlight that the dummy traffic affects the capacity of

the physical channels of the two considered users, and produces different interference levels that cause variable Bit Error Rate (BER).
The presented approach gives no possibilities to the implementation of a re-negotiation procedure of the physical resources during a trial. To give realistic numerical values for the bit rate of the physical channel we can state that the total supported bit rate is 128 Kb/s in the uplink direction and 360 kb/s in the downlink direction and it has to be shared between the two considered users and the interfering users.
A fundamental element is that during the execution of a trial of the experiment it is not possible to change the statistical distribution of the traffic: according to this point of view the only modifications of the previous characteristics could be performed only at the beginning of the experiment. To obtain significant results from the experiment we are introducing, it is necessary to repeat the simulation varying the above-mentioned parameters.
Three types of traffic are considered in the experiment: voice over IP (VoIP), web browsing (WB) and file transfer (FTP) traffic.
Two levels of dynamic scheduling are provided: one scheduler within each user equipment (UE) dealing exclusively with the traffic sources of the considered user and one scheduler, on the network (NW) side, dealing with the two users' and interfering users' traffic, whose way of working depends on a dynamic code assignment scheme.
An appropriate congestion control mechanism must be implemented in the RRC (Radio Resource Control) subsystems, both on the UE (User Equipment) and on the RNS (Radio Network Subsystem) sides, to provide that, on the one hand, the QoS contracts established with the connections are respected, and that, on the other hand, the capacity the real user can avail of is efficiently exploited. In the proposed simulation architecture, the above-mentioned congestion control mechanism are implemented by means of Dual Leaky Buckets (DLBs), whose input parameters could be either fixed or dynamically changing.
The outgoing packets of a DLB are stored in a queue (named QoS queue to differentiate it from the internal DLB queue), where they stay until they are allowed to be transmitted. It is important to establish that all the queues that are present in the following functional designs have a First In First Out (FIFO) way of working, every one utilizes one DLB, which acts as Congestion Control Executor under the control of the Congestion Control Handler, so the congestion control has an IntServ approach.
The decision about the DLB parameters' values for every connection is taken by the Congestion Control Handler, taking into account a static information (the QoS contract parameters) and a dynamic one (the measurement results) if the DLB input parameters change dynamically, only the static information if the DLB input parameters are fixed. In both solutions the choice of the input values for all the DLBs is executed from the transmitting entity without the action of the receiving entity. This approach is suitable for a satellite system that has a high transfer delay.
The dynamic information is provided by the Measurements Executor, which performs, on a frame-by-frame basis, measurements on the QoS queues and stores them in the Measurements Register (MR); that information are provided to the QoS Sorter and to the Scheduler.
If the source (e.g. a particular active application) breaks the established contract and it produces an excessive amount of traffic in comparison with the agreed values, the surplus traffic is considered as "Non-Compliant Traffic". In case of congestion the Non-Compliant Traffic is immediately discarded by the DLBs and it is definitively lost.
A QoS Sorter is foreseen for every user whose basic function is to arbitrate between packets that are ready for transmission and are stored in the QoS queues. The QoS Sorter works on a frame-by-frame basis, choosing, among the users' queues, the one from which a certain number of packets are allowed to be transmitted; in the following we refer to that queue as the *leading queue*.
An entity on the network side, the Scheduler, decides the amount of data to be taken from the selected queue, taking into account the requirements of the two users, and of the interfering users (as a whole). The Scheduler is the entity that provides a flexible resource allocation scheme, basing on the available physical resources.
The concatenation and the segmentation functions are provided after the QoS queues: because of the different size of the data packets that are manipulated (IP packets), it is necessary an adaptation method to fit the data length to the used packet.
A logical representation of the architecture previously described is given in Figure 1 for the uplink direction and in Figure 2 for the downlink direction (for both directions, only the transmitting side is depicted for sake of simplicity), highlighting the logical entities within the protocol stack.

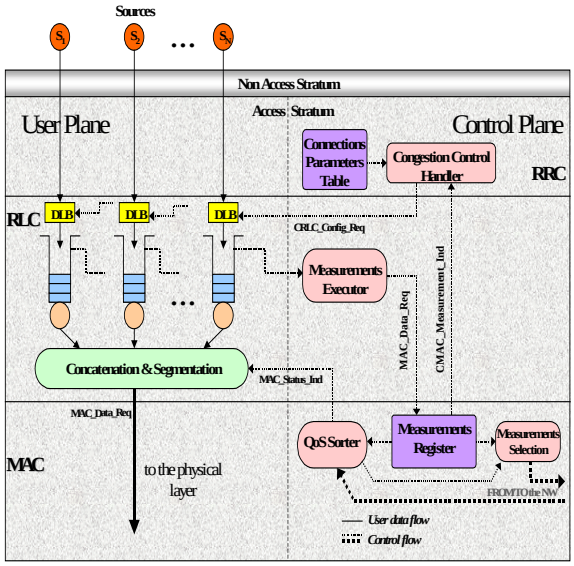

Figure 1. UE Logical Architecture, Uplink

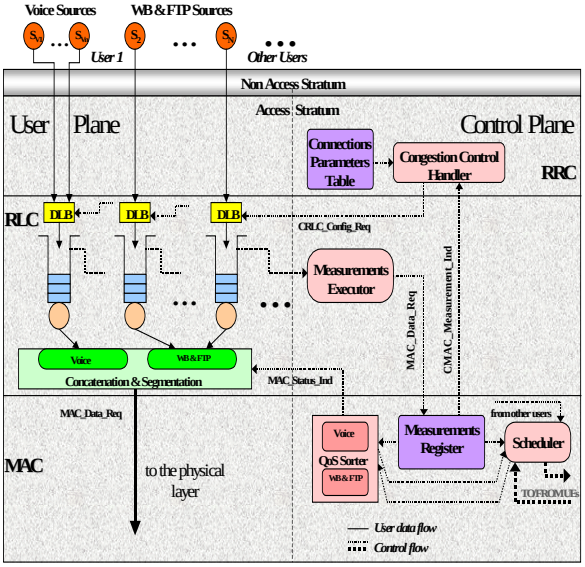

Figure 2. RNS Logical Architecture, Downlink

As already mentioned, two possible Congestion Control strategies can be implemented:

Closed Loop strategy
The DLB parameters change dynamically, on the basis of measurements performed on the QoS queues.

Open Loop strategy
The dynamic information is not foreseen and the congestion control is provided only by means of static parameters established at connection set-up.

The only architectural difference between the two described strategies is that in the Open Loop strategy the CMAC_Measurement_Ind primitive shown in the previous figures is not present.
In the following we will analyze the Open Loop Strategy, which is the strategy adopted in the simulation.

Traffic sources

In the proposed architecture, we consider three different traffic classes: voice (VoIP), web browsing (WB) and file transfer (FTP).
Voice is transmitted using a constant packet size of 240 bits, in transparent mode. It is not foreseen an RLC header. Segmentation is not provided, thus the Concatenation & Segmentation functional block provides only packets concatenation.
There are some differences between downlink and uplink architectures. In the uplink, a queue is provided for every voice source in the transmitting side; in the downlink network side, instead, all voice sources are gathered in the same queue, so in the packet header a field is necessary to identify the different streams. In both directions, WB and FTP traffic uses IP packets in unacknowledged mode and segmentation can occur. One queue is activated for every source.

QoS Queues

In the proposed architecture, for each user we consider variable number of QoS queues, from 0 up to 15 (according to the C/T field described in [1]):

- for voice connections
  - ✓ Uplink: a variable number of VoIP queues (one queue for each VoIP source);
  - ✓ Downlink: one queue gathering all the voice connections, if there are any (from 1 up to 12 simultaneous total connections for the two users);
- a variable number of WB queues (one queue for each WB source);
- a variable number of FTP queues (one queue for each FTP source).

Primitives

Interactions between different layers are described in terms of primitives, where the primitives represent the logical exchange of information and control between the layers. The (adjacent) layers are connected to each other through Service Access Points. Primitives, therefore, are the conveyers of the information exchange and control through SAPs (not shown in Figure 1 and Figure 2 for simplicity).
In the following, we describe only primitives and parameters which are of interest for the simulation architecture proposed. To this aim, we give a lightly modified version of the considered primitives as defined in 3GPP specifications, ignoring all those parameters and primitives out of the scope of the experiment.
Two types of primitives are used for the present document, as follows:

- ***Request primitive***
  This type is used when a higher layer is requesting a service from a lower layer.

- ***Indication primitive***
  This type is used by a lower layer providing a service to notify its higher layer of activities concerning that higher layer [2].

Radio Resource Control layer

The exploited Radio Resource Control (RRC) functionality is to inform the Radio Link Control (RLC) about DLBs parameters.
The Congestion Control Handler (CCH) gets static information from the Connections Parameters Table and, if the Closed Loop strategy is applied, dynamic information about the status of the QoS queues (buffer occupancy) from the Measurements Register (MR), the

latter by means of the CMAC_Measurement_Ind primitive. It then sets the parameters (average rate, peak rate and token bucket size) of the Dual Leaky Buckets, through the CRLC_Config_Req primitive, thus allowing them to keep limited the number of packets in the QoS queues (the DLB implements the congestion control functionality).

Radio Link Control layer

RLC provides transmission towards MAC (Medium Access Control) layer, implementing segmentation functionality for WB and FTP packets. The Measurements Executor (MEx) performs measurements on the QoS queues and uses the MAC_Data_Req primitive to store in the MR, located in the MAC layer, the status (buffer occupancy) of all the QoS queues and, for each queue, the timestamp (i.e. the simulation time stored in the packet when the source sends it to the RLC layer) of the head packet (in addition, information about the number of voice connections is needed in the downlink). The Concatenation & Segmentation block receives from the QoS Sorter, by means of the MAC_Status_Ind, the following information:

- *uplink*: the leading queue index and the number of bits to be transmitted from the leading queue. These values are used for concatenation and segmentation;
- *downlink*: the number of bits to be transmitted from the voice queue, the leading queue index and the number of bits to be transmitted from the leading queue (WB or FTP). These values are used for concatenation and segmentation.

The Concatenation & Segmentation block calculates, on the basis of the information received from the QoS Sorter, the number of packets to be removed without segmentation from the queue and provides segmentation, if needed, for the last packet to be transmitted. It is worth highlighting that, as previously stated, segmentation is not provided for voice packets.
User data after the Concatenation & Segmentation block are forwarded to the MAC layer by means of the MAC_Data_Req primitive.

Medium Access Control layer

It exchanges information with both the RRC and the RLC layer. If the Congestion Control Closed Loop strategy is applied, the MR provides to the CCH, using the CMAC_Measurement_Ind primitive, the status of all the QoS queues. MR stores (MAC_Data_Req primitive) information about the status of all the QoS queues and, for each queue, the timestamp of the head packet (and the number of voice connections, in the downlink), thus allowing the QoS Sorter to calculate the deadline of each head packet; those information are provided by the MEx. MAC layer then supplies to the Scheduler (placed in the Network side) the length of the leading queue (and the number of voice connections, in the downlink).
In the uplink, the QoS Sorter calculates the leading queue index, comparing the deadlines of the head packets, stored in the MR, and provides it to the Measurements Selection block (MSel); the MSel then selects the leading queue length and supplies it to the Scheduler. In the downlink, the MSel functionality is included in the Scheduler itself. The Scheduler, using information about the traffic of all users connections handled, can inform the QoS Sorter about the number of bits to be transmitted. The QoS Sorter forwards this information to the Concatenation & Segmentation block (through MAC_Status_Ind primitive).
User data are forwarded from the MAC layer to the Physical layer by means of the PHY_Data_Req primitive.

Selection of the Leading Queue

In order to provide transmission of packets from the QoS queues respecting the QoS constraints established at connections set-up, an appropriate scheduling algorithm within each User Equipment must be performed. This functionality is provided by the QoS Sorter. The algorithm here proposed is named Head Packet Earliest Deadline First (HP-EDF) and is based on the Earliest Deadline First (EDF) algorithm ([3], [4], [5]). This scheduler is a form of a dynamic priority scheduler where the priorities for each packet are assigned as it arrives. Specifically, each packet is assigned a deadline, which is given by the sum of its arrival time and the delay guarantee associated with the flow the packet belongs to. This scheduler determines the order in which queued packets are forwarded over outgoing links at switches and routers. This order determines the packets' waiting time in the link's queue, and ultimately the delay that the link scheduler can guarantee. The EDF selects the packet with the smallest deadline for transmission on the link and hence the name.

Project choices: Head Packet EDF

Our architecture provides multiplexing of different logical channels on different MAC-PDUs; a MAC-PDU is transmitted each frame period, i.e. each 10 ms. Thus, packets contained in a MAC-PDU belong to the same logical channel (i.e. to the same queue). We applied Earliest Deadline First (EDF) algorithm to select the queue by which the packets to be transmitted are extracted. At each frame, EDF is applied to the head packets of a user's QoS queues; the queue whose head packet has the smallest deadline is selected to transmit its packets in the current frame. We name this algorithm Head Packet EDF (HP-EDF). Application of

this policy implies that packets transmitted in a frame are not the packets which EDF algorithm, if applied, would have chosen (EDF algorithm should be applied every time a packet is selected for transmission). In fact, the deadline of the packets belonging to the selected queue, except the head packet, is not considered; thus, it would be possible that their deadlines are not the smallest ones. After a head packet (and so the queue with the right of priority) is selected by EDF at the beginning of the frame, a packet not belonging to the selected queue could have the smallest deadline.

Uplink: UE side
A two-level HP-EDF based algorithm is applied by the QoS Sorter for the leading queue choice:

- *level 1*: HP-EDF is applied to voice queues only, if there are any, and a voice queue is chosen as the queue having the right of priority for transmission; if no voice connection is active, the leading queue is chosen applying level 2;
- *level 2*: HP-EDF is applied to WB and FTP queues and a leading queue is chosen.

This choice has been made because of the stricter time delay requirements of voice traffic than the web browsing and file transfer one.

Downlink: NW side
HP-EDF is applied only to WB and FTP queues and a queue is selected for transmission in the current frame. HP-EDF is not applied to voice queues.

Simulation results

We performed the simulation using the parameters shown in Table 1 and considering the two users have an active connection for each type of traffic.

Table 1. Traffic sources and QoS parameters

Traffic type	Average rate [Kb/s]	Maximum rate [Kb/s]	Maximum delay	Packet loss probability
VoIP	8	16	30 ms	10^{-4}
Web	UL: 29 DL: 53	UL: 56 DL: 100	UL: 2 s DL: 1 s	10^{-7}
FTP	UL: 16.5 DL: 36.5	UL: 20 DL: 40	UL: 3.5 s DL: 3 s	10^{-7}

In the following we will show the graphics representing the end-to-end delay. The packet loss probability constraint is respected because packets are discarded neither from the QoS queues nor from the DLB buffers.

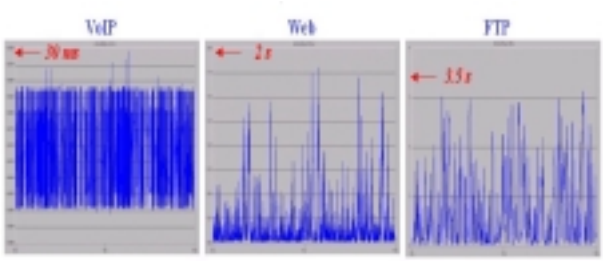

Figure 3. Uplink end-to-end delays

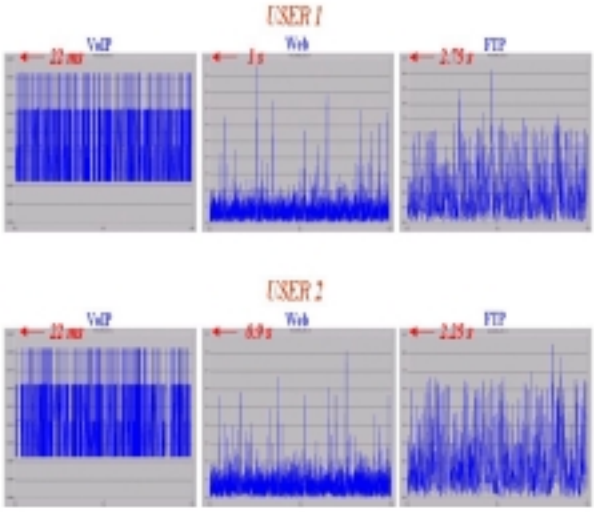

Figure 4. Downlink end-to-end delays

Conclusions

The proposed resource allocation algorithm, applied with a congestion control mechanism, allows the access network to guarantee the appropriate QoS contract agreed at connection set-up between the users and the UMTS access network operator, obtaining a fair distribution of the available resources between the users and a high link utilization.
The simulation results show that the maximum end-to-end delay and packet loss probability parameters are respected for each kind of traffic.

References

[1] 3G TS 25.321 V3.3.0: MAC protocol specification

[2] 3G TS 25.302 V3.4.0: Services provided by the Physical Layer

[3] R. Guerin and V. Peris, “Quality of Service in Packet Networks Basic Mechanism and Directions”

[4] V. Firoiu, J. Kurose and D. Towsley, “Efficient Admission Control of Piecewise Linear Traffic Envelopers at EDF Schedulers”, January 1998

[5] G. Mamais, M. Mrkaki, G. Politis, I. Venieris, “Efficient Buffer Management and Scheduling in a Combined IntServ and DiffServ Architecture: A Performance Study”, 1999