Priority Queuing for IP-Based Service Differentiation in the UMTS Radio Access Network

Abedellatif Samhat, Tijani Chahed, and Gerard Hébuterne

GET/Institut National des Télécommunications
9 rue Charles Fourier - 91011 Evry CEDEX - France
{abedellatif.samhat, tijani.chahed, gerard.hebuterne}@int-evry.fr

Abstract. In this work, we investigate service differentiation under IP for the transport of both real-time and non real-time user traffic in the UMTS Terrestrial Radio Access Network (UTRAN). Therein, stringent delay bounds are to be met for both types of traffic, albeit tighter for the voice traffic. For the sake of simplicity, we suggest, model and analyze the use of priority queuing, as an efficient way to implement service differentiation. Our results are validated empirically on a test-bed emulating the UTRAN transport functionalities. This is carried out following a mean value as well as a percentile analysis. Based on these results, we draw the proper dimensioning of the UTRAN so as to meet the target QoS requirements.

1 Introduction

The Universal Mobile Telecommunications System (UMTS) promises to enable a wide range of multimedia applications and seamless service delivery in multiple mobile environments while granting them Quality of Service (QoS). An intensive research activity is currently investigating the basic design of UMTS, and a significant amount of this activity focuses on the UMTS Terrestrial Radio Access Network (UTRAN) [1] and the transport technology that shall be used therein. Specific to UMTS are the stringent delay bounds for the transport of various types of user traffic real-time as well as non real-time, with various QoS needs, over the UTRAN. This is imposed by the WCDMA advanced radio control functions. The transport in the UTRAN should meet these requirements in a cost effective way in terms of efficiency and maximal utilization of the bandwidth. This latter is typically formed of E1 links. Early works on the UTRAN focused solely on AAL2/ATM, especially in the release 99 of IMT-2000 standards [1]. AAL2 offers an elegant way both to multiplex voice traffic and to differentiate between voice and data traffic in the UTRAN, as reported in some analytical works [2] [3] or others carried out by simulations [4] [5] [6].

With the advent of IP as a de facto networking technology and its presence in 3G core network, IP is making its way to the UTRAN. Its standardization is still under way and should be finalized in Release 6 of the 3GPP standards [7]. IP-based UTRAN is further supported by the Mobile Wireless Internet Forum (MWIF) [8]. Contrary to most works in this area, that use simulations, we developed in [9] an analytical model for the transport of real-time voice traffic over the UTRAN using IP adopting a similar approach than the one in [2] for the AAL2/ATM case. Our results show the feasibility of IP as a transport technology in the UTRAN as well as its efficiency.

N. Mitrou et al. (Eds.): NETWORKING 2004, LNCS 3042, pp. 490–501, 2004.
© IFIP International Federation for Information Processing 2004

In this paper, we consider IP service differentiation between voice and data traffic in the UTRAN. The main difficulty lies in the fact that, in this case, both traffic types have stringent delay bounds to be met, albeit tighter for the real-time voice. We hence suggest priority queuing as a simple yet efficient way to meet these objectives. Data traffic shall however not be severely penalized owing to its elastic nature and the low bit rate, well controlled nature of voice traffic. We then validate our proposed model through an empirical work on a test-bed emulating the UTRAN transport functionalities. This is done in two fashions : mean value as well as percentile analysis. We eventually use these results to draw proper dimensioning of the UTRAN in order to meet the required performance.

The remainder of this paper is organized as follows. In Section II, we present the main features of the UTRAN and recall the issues related to QoS, mainly in terms of delay. In Section III, we present our model for the UTRAN which we analyze in Section IV. We consider both voice and data traffic and evaluate their delay performance. In Section V, we show empirically as well as numerically the benefits of priority queuing over a non-priority scheme for the transport of both traffic types. We then present in section VI the required dimensioning of the UTRAN in this case. Section VII eventually concludes the paper.

2 Radio Access Network in UMTS

2.1 Architecture

The UTRAN, as described in [1], interconnects the Uu interface (air interface) and the Iu interface. It contains the Node B, the Radio Network Controller (RNC) and their interconnection. The RNC is responsible for the control of radio resources of UTRAN. It plays a very important role in power control, handover control, admission control and load control. RNC interfaces the core network via Iu interface and uses Iub interface to control one Node B. The Iur interface between RNCs allows soft handover between RNCs. Node B is equivalent to the GSM base station (BS/BTS), and it is the physical unit for radio transmission and reception with cells. Node B performs the air interface processing, which includes channel coding, interleaving, rate adaptation and spreading. The connection with the user equipment is made via Uu interface, which is actually the WCDMA radio interface.

The user plane protocol stack in the UTRAN is shown in Figure 1. RLC (Radio Link Control) establishes the RLC connection between UE and RNC. The MAC layer deals with logical channels. It handles the mapping between the logical channels and the transport channels. Transport channels are categorized depending on the transmission format. The output of the MAC layer consists of sets of Transport Block (TB) periodically generated every Transmission Time Interval (TTI) of the transport channel. For each transport channel, FP layer assembles the bursts transmitted in one TTI into one FP frame which is transmitted to the transport network layer, the IP transport technology in our case.

Why IP? IP is already present in the 3G core network; it would make it easier both for users and operators to have an all-IP setting. Besides, IP as the common layer 3

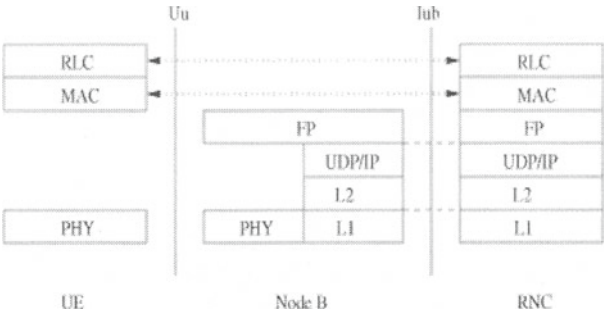

Fig. 1. Protocol stack

protocol in the UTRAN brings flexibility to an operator in choosing a Layer 1/2 backhaul technologies, including options of IP over synchronous optical network (SONET) or IP over wavelength-division multiplexing (WDM). Eventually, with IP, several possibilities exist to provide sufficient QoS, DiffServ for instance.

2.2 QoS Requirements and Service Differentiation

The real-time nature of voice-oriented applications and the WCDMA radio control functions impose rather stringent delay requirement on the UTRAN transport network for both types of traffic, real-time as well as non-real-time: 3GPP specifies 5 ms delay bound for real-time traffic in addition to minimal jitter. The delay requirement for data traffic can be as small as 10ms, possibly 50ms, owing to the radio functions of the outer-loop power control and soft-handoff control. This implies that challenging demands are imposed for the fulfillment of QoS requirements. Service differentiation should thus give voice traffic a higher priority while guaranteeing both delay targets for both traffic types, as explicited in the next section.

3 Model for UTRAN System

We propose to model the Iub interface of the UTRAN as shown in Figure 2. FP frames belonging to voice and data traffic are separately packetized and processed through a transfer unit. They however share a common output link that includes a scheduler that should enforce their respective QoS needs while optimizing the bandwidth utilization.

In this work, we suggest the use of Priority Queuing (PQ) as an efficient yet simple means of implementing service differentiation. With strict, non-preemptive PQ, IP voice packets are always served first since they have a tighter latency constraint; IP data packets are only served when the voice queue is empty. We note that the main factor influencing the jitter for voice traffic is the inability to pre-empt a data packet that has just begun receiving service when the voice packet arrives. To minimize this effect, data packet sizes should be kept low, for example on a 2Mbps link, a 1500 byte data packet introduces a 6ms link blocking, too high a value with regards to our voice delay budget. Service differentiation should also make sure that data traffic is not starved. In this case, proper dimensioning shall be implemented.

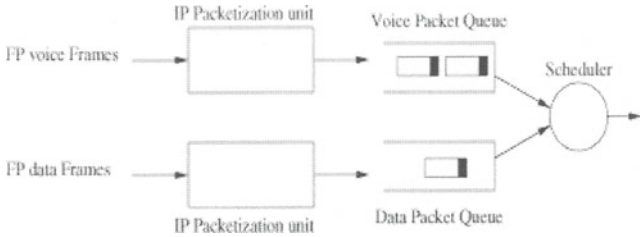

Fig. 2. IP model in the UTRAN

4 Analysis

4.1 High-Priority Voice Traffic

Voice traffic is generated by Adaptive Multi-Rate (AMR) codecs at 12.2 kbps. The transport channels carrying speech traffic shall be assigned a 20 ms TTI value. One DCH (Dedicated CHannel) is allocated to each user. Voice traffic consists of a succession of ON and OFF periods. When a user is in ON period, the MAC layer transfers a 31 bytes speech frame each 20ms; this 31 bytes speech frame is encapsulated into one FP voice frame by adding a 5-bytes header. The individual channel's FP frames arrive periodically, every TTI, at the IP packetization unit. As voice traffic is symmetric, the model holds for both uplink and downlink.

For voice case, the IP packetization unit (Figure 2) is a multiplexing or assembly unit where IP voice packets are filled with FP frames, each FP frame has an additional 3-byte as Multiplexed Header (MH) including one byte termed User IDentifier (UID) unique to each user. The maximum size for IP voice packets is s_v and the payload for this completely filled packet corresponds to n FP frames. A Timer Common Usage (TCU) is associated with the assembly unit so as to avoid unacceptably large packetization delay. Then some IP packets can be partially filled and the packetization delay is bounded by the TCU value.

For large number of simultaneously ON DCH channels N_v, i. e. large number of active users, the arrival process of FP voice frames is modeled by a Poisson process with mean rate r_v equal to $\frac{N_v}{20}$ FP frames/ms. The arrival process of the IP packets to the voice queue is affected by the use of the TCU in the packetization unit. In [9], we derive analytically the IP packet arrival rate taking into account the TCU effect; including all IP packet configurations : partially as well as completely filled ones. We now reproduce briefly our work (see [9]for details).

Let the time axis be discrete, scaled according to Time Units (TUs). Let λ be the mean arrival rate of frames to the packetization unit, in units of frames per TU. Since arrivals are Poisson with mean rate λ, the probability that k FP frames arrive in one TU in the assembly unit is given by $P_k = \frac{\lambda^k}{k!}e^{-\lambda}$.

Let Y be the process that describes the number of FP frames in the assembly unit and let the following matrices denote the dynamics of Y during one TU :

- Let A_0 be a $1 \times (n-1)$ transition matrix from $Y = 0$ (empty unit) to $Y > 0$ (non empty unit) during one TU without IP Packet generation. Then, $A_0 = \begin{pmatrix} P_1 & P_2 & \dots & P_{n-1} \end{pmatrix}$.
- Let A_{00} be an $(n-1) \times (n-1)$ transition matrix from $Y > 0$ to $Y > 0$ during one TU without IP Packet generation.

$$A_{00} = \begin{pmatrix} P_0 & P_1 & P_2 & \dots & P_{n-2} \\ 0 & P_0 & P_1 & \dots & P_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & P_0 \end{pmatrix} \tag{1}$$

- Let A_{0i} be a $1 \times n$ transition matrix from $Y = 0$ to Y during one TU with the generation of i IP packets. Then, $A_{0i} = \begin{pmatrix} P_{in} & P_{in+1} & \dots & P_{in+n-1} \end{pmatrix}$.
- Let A_i be an $(n-1) \times (n)$ transition matrix from $Y > 0$ to Y during one TU with the generation of i IP packets.

$$A_i = \begin{pmatrix} P_{in-1} & P_{in} & P_{in+1} & \dots & P_{in+n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_{in-n+1} & P_{in-n+2} & \dots & \dots & P_{in} \end{pmatrix} \tag{2}$$

- Let A_{+0} be an $(n-1) \times n$ transition matrix from $Y > 0$ to Y during one TU when TCU expires.

$$A_{+0} = \begin{pmatrix} \sum_{k=0}^{n-2} P_k & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{k=0}^{0} P_k & 0 & 0 & \dots & 0 \end{pmatrix} \tag{3}$$

Just after the beginning of a TU, FP frames arrive and IP voice packet(s) is (are) immediately generated if there are enough frames to constitute them. Let Y_m be Y just after the mth generation of a group of IP packets and let t_m be the corresponding TU. Let T_m be the time between the mth and $(m+1)$th generation of a group of IP packets, i.e., $T_m = t_{m+1} - t_m$. Let S_m be the number of FP frames that arrive during T_m and let C_{m+1} be the number of IP packets generated in the $(m+1)$th group packet generation. We have,

$$Y_{m+1} = \begin{cases} 0 & \text{if TCU expires and } C_{m+1} = 1 \\ Y_m + S_m - nC_{m+1} & \text{otherwise} \end{cases} \tag{4}$$

where $C_{m+1} = \lfloor \frac{Y_m + S_m}{n} \rfloor$. Let $r = (r(0), r(1), r(2), \dots\dots, r(n-1))$ be the corresponding stationary probability vector of Y, with $r_0 = r(0)$ and $r_+ = (r(1), r(2), \dots, r(n-1))$. For e a $n \times 1$ all-1 column vector, the mean number g of IP packets in a group and the mean inter-group generation time τ are given by

$$g = r_+ (\sum_{t=1}^{TCU} A_{00}^{t-1} \sum_{i=1}^{\infty} i A_i + A_{00}^{TCU-1} A_{+0}) e$$

$$+ r_0 \sum_{s=1}^{\infty} P_0^{s-1} (\sum_{t=1}^{TCU} A_0 A_{00}^{t-1} \sum_{i=1}^{\infty} i A_i + A_0 A_{00}^{TCU-1} A_{+0} + \sum_{i=1}^{\infty} i A_{0i}) e \quad (5)$$

$$\tau = r_+ (\sum_{t=1}^{TCU} t A_{00}^{t-1} \sum_{i=1}^{\infty} A_i + TCU A_{00}^{TCU-1} A_{+0}) e + r_0 \sum_{s=1}^{\infty} P_0^{s-1}$$
$$(\sum_{t=1}^{TCU} (s+t) A_0 A_{00}^{t-1} \sum_{i=1}^{\infty} A_i + (s + TCU) A_0 A_{00}^{TCU-1} A_{+0} + s \sum_{i=1}^{\infty} A_{0i}) e \quad (6)$$

The mean rate of IP packet generation at the assembly unit λ_v packets per TU, which is equal to the mean IP packet arrival rate to the voice queue, is given by $\lambda_v = \frac{g}{\tau}$.

At this point, we make a change in the time scale as follows. We assume that the service time is constant and equal to the time needed to serve an IP packet with s_v size at output link capacity C, denoted by Service Time Unit (STU). Normalizing by STU, the mean packet arrival rate in one STU is then equal to the server load ρ_v. $n\rho_v$ is the mean arrival rate of FP frames. This is an over-estimation as some IP packets can leave the assembly unit with less than n FP frames.

Using the independence approximation (see [2]) which states that FP frames arrive according to a Poisson process with mean rate $n\rho_v$, we calculate the probability θ_i that i IP packets arrive to the transmission queue in one STU. That is approximately equivalent to $i \times n$ FP frames arriving to the system in one STU. We have,

$$\theta_i = \begin{cases} \sum_{j=0}^{n-1} \frac{n-j}{n} \frac{(n\rho_v)^j}{j!} e^{-(n\rho_v)} & i = 0 \\ \sum_{j=1}^{n-1} \frac{j}{n} \frac{(n\rho_v)^{(i-1)n+j}}{((i-1)n+j)!} e^{-(n\rho_v)} + \sum_{j=0}^{n-1} \frac{n-j}{n} \frac{(n\rho_v)^{in+j}}{(in+j)!} e^{-(n\rho_v)} & i = 1, 2, \ldots \end{cases} \quad (7)$$

We now define the embedded Markov chain X to be the number of IP packets present in the transmission queue at the end of the service of one packet. The matrix P of transition probabilities $[p_{ij}]$ $(i, j = 0, 1, 2, \ldots)$ is given by (see [11] page 178)

$$P = \begin{pmatrix} \theta_0 & \theta_1 & \theta_2 & \theta_3 & \ldots \\ \theta_0 & \theta_1 & \theta_2 & \theta_3 & \ldots \\ 0 & \theta_0 & \theta_1 & \theta_2 & \ldots \\ 0 & 0 & \theta_0 & \theta_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (8)$$

The stationary probability that there are i IP packets in the voice queue is given by $\pi = \pi P$ where $\pi = (\pi(0), \pi(1), \pi(2), \ldots)$. Taking into account that the mean arrival rate is ρ_v and the average number of IP packets in the queue is q, we apply Little's formula and obtain d_q, the average queuing time in the transmission unit

$$d_q = \frac{q}{\rho_v} = \frac{\sum_{i=1}^{\infty} i\pi(i)}{\rho_v} \quad (9)$$

Let M be the time it takes for n frames to arrive. As the frames arrival process is Poisson with rate r_v, M is distributed according to an Erlang$(n-1, r_v)$ distribution with mean $E(M) = \frac{n-1}{r_v}$. The mean delay d_v of individual FP frames at the multiplexer including packetization at the assembly unit and queuing at the transmission queue is given by

$$d_v = \begin{cases} d_q + \text{E(M)} & \text{if E(M)} \leq \text{TCU} \\ d_q + \text{TCU} & \text{if E(M)} > \text{TCU} \end{cases} \tag{10}$$

Owing to the presence of data traffic with packet size s_d, the total mean delay D_v on voice traffic is given by

$$D_v = d_v + \frac{s_d}{2C} \tag{11}$$

4.2 Low-Priority Data Traffic

Data traffic refers to various applications. It ranges from background traffic, such as mail, to interactive traffic, such as Web surfing. In this work, we focus on downlink Web browsing traffic, transported over a Downlink Shared CHannel (DSCH), shared by several users. By reference to the preceding voice traffic model over dedicated channels, the TTI value is 40 ms in this case. The DSCH channel bit rate can be 64 kbps, 144kbps or 384kbps. At the FP layer, and as is the case of voice, Web traffic over DSCH can be modeled as a succession of ON and OFF periods. In one ON period, for a 64 kbps channel, a frame of 320 bytes is generated every 40 ms and is encapsulated into one FP data frame by adding a 5-byte header.

We focus on the case where one FP data frame is encapsulated into one IP packet by adding UDP/IP header. The packet size s_d amounts to 353 bytes. Assuming that the time for forwarding IP packets from the packetization unit to the transmission queue is negligible, the system is then equivalent to the transmission queue only. For N_d simultaneously ON data channels, the arrival process of IP data packets to the data transfer queue can be modeled by a Poisson process and the queue itself is viewed as an $M/D/1$ queue with vacations. The vacations correspond to epochs when the server is busy serving higher-priority voice traffic. Let V be a random variable indicating the duration of vacations. The values of V correspond thus to the length of the busy period for the voice queue. Normalizing the time axis, both for arrival and service processes, to the fixed time needed to serve a data packet of size s_d at link capacity C, and using the Pollaczek-Khinchin formula for an $M/D/1$ queue, we calculate the mean queuing delay d_d for a data packet as

$$d_d = \frac{2 - \rho_d}{2(1 - \rho_d)} \tag{12}$$

where ρ_d is the data workload. The mean queuing delay for an $M/D/1$ queue with vacations is given by [12]

$$D_d = d_d + \frac{E[V^2]}{2E[V]} \tag{13}$$

where $E[V]$ and $E[V^2]$ are respectively the first and second moments of the vacation interval V and correspond to the first and second moments of the busy periods for the voice queue and are thus given by

$$E[V] = \sum_{i=1}^{\infty} (i * STU)\pi(i) \tag{14}$$

$$E[V^2] = \sum_{i=1}^{\infty} (i * STU)^2\pi(i). \tag{15}$$

5 Numerical Results and Empirical Validation

In this section, we first describe the UTRAN test-bed that we use for the empirical validation of our results. We then validate our model through a comparison between the analytical and the empirical results. Our analytical work was based on a mean value analysis. A percentile analysis can give further insights into the problem. It is carried out empirically in this work and is presented in the third subsection. Note that throughout this section and for the case of voice traffic, we set n, the number of FP voice frames per IP packet equal to 8 and the TCU timer value equal to 2 ms. Those values yield an optimal link utilization, as shown in our work in Reference [9].

5.1 Emulating UTRAN Transport on a Local Test-Bed

Our test-bed emulating the UTRAN transport functionalities is based essentially on PCs operating with FreeBSD Operating System (OS). All elements of the test-bed are synchronized using GPS equipment, essential for accurately measuring delays at very small scales. The traffic at the FP layer is generated using the server-client UDP/IP model with socket programming. The assembly and the transmission units are implemented as follows: at the assembly unit, threads are implemented to execute the multiplexing algorithm including the assembly process and the timer process. The PQ algorithm is implemented through the use of the ALTQ tool [13] on FreeBSD system. Tcpdump and other locally-coded software are used to record desired parameters.

5.2 Mean Value Results

In Figure 3, we compare the analytical mean delay D_v for high-priority voice traffic and the empirical one obtained from experiments on the test-bed for a number of simultaneously active voice channels N_v=100 and in the presence of data traffic. PQ ensures service differentiation. The x-axis shows different values of the link capacity C. As expected, the mean delay decreases when the link capacity increases. One can see that the empirical mean delay is less than the analytical one, this is explained by the fact that analytically there is an over-estimation due to the assumption that all packets are completely filled with FP frames. This assumption leads to an analytical packetization delay higher than the empirical one.

Figure 4 compares the analytical and the empirical mean delay D_d for data traffic as a function of data load ρ_d. The three sets of curves correspond to three different values of ρ_v, the load of voice traffic : zero, medium and high. The output link capacity is fixed at 2 Mbps. Curves are truncated at values where the total load is strictly less than 1, i.e.,

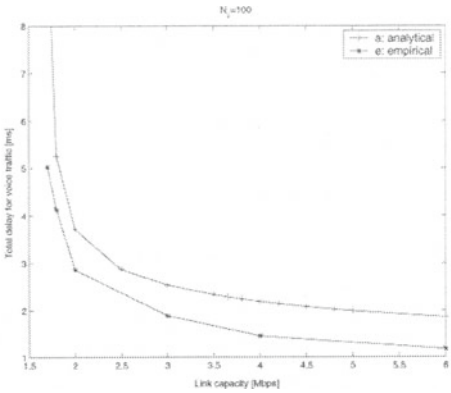

Fig. 3. Mean delay for voice traffic: analytical versus empirical

$\rho = \rho_v + \rho_d < 1$. Note that the empirical curves are generated through a Poisson arrival process. We note that the mean delay D_d increases as the data workload increases. When no voice traffic is present, the results obtained analytically match very closely with the empirical ones when the arrival process of data packets is Poisson. When the load of voice traffic is medium or high, the analytical and empirical curves differ slightly. This is mainly due to the terms related to vacations in the analytical case.

In Figure 5, we reproduce the same performance measures as in the previous one. The curves are however both empirical and correspond to two different traffic generation schemes: Poisson versus periodic. The latter represents best the real system as it reproduces the periodic nature of arrivals per TTI. The results, in terms of mean delay, show that the Poisson approximation is a good one. We note that for a low load of data traffic, i.e. a low number of data channels, the periodic traffic is less bursty than Poisson. When the number of data channels increases, the periodic traffic can be more bursty.

5.3 Percentile Results

Figures 6 and 7 confirm further the benefits of the use of service differentiation through a percentile analysis. In effect, those Figures draw the 95 percentile of three quantities :

1. Total delay of voice traffic including packetization, queuing and possible jitter in the presence of data traffic (denoted by total delay in the figures).
2. Queuing (or transmission) delay of voice traffic and possible jitter in the presence of data traffic (denoted by trans delay in the figures).
3. Delay of data traffic.

Figure 6 shows the results of the above mentioned delays with and without the use of PQ for ρ_v=0.36 and output link capacity equal to 2 Mbps. The x-axis shows the total load $\rho = \rho_v + \rho_d$. We observe that when no differentiation is implemented, the voice transmission delay and the data delay are almost equal. The total voice delay is higher due to the packetization delay component and violates the voice delay budget for almost

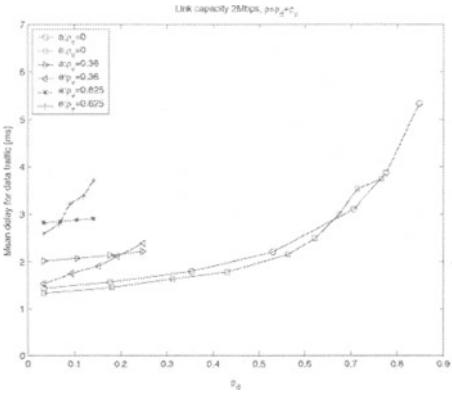

Fig. 4. Mean delay for data traffic: analytical versus empirical

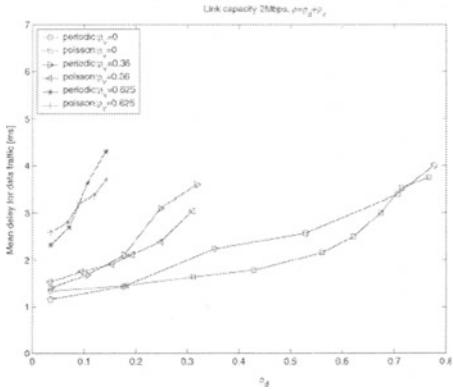

Fig. 5. Mean delay for data traffic: Poisson versus periodic

all values of ρ_d. Using PQ, voice delays both total and transmission get less than the 5ms delay bound. At the expense of a large delay for data traffic, yet below the data delay budget. The same observations can be made for Figure 7 where ρ_v=0.5. In this case however, the total load is less than the one in the previous case. This is due to the fact that voice traffic has an even tighter delay than the data one and gets more critical as its percentage increases with respect to the one of data in the total load.

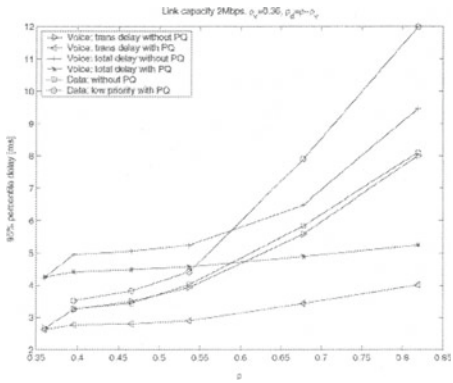

Fig. 6. 95 percentile delay of voice and data traffic with and without PQ

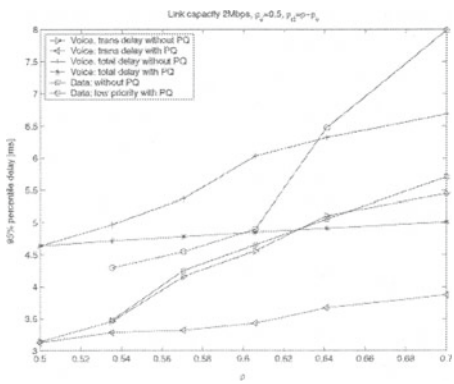

Fig. 7. 95 percentile delay of voice and data traffic with and without PQ

Figure 8 shows the Cumulative Distribution Function (CDF) of the total delay of voice traffic when ρ_v=0.36 for different values of ρ_d. The difference between curves is the jitter due to the presence of data traffic. When no data traffic we note that 98.15 % of voice traffic has a delay less than 5 ms. When ρ_d=0.2 and 0.36, this figure is about

96.75% and 95.53% respectively. Figure 9 shows the CDF of the delay of data traffic when ρ_d=0.2 for different values of ρ_v. We note that when the load of voice traffic is low or medium the data delay bound is respected. For high loaded systems, when ρ_v=0.625 or larger, one can observe that about 88% of the data traffic experiences a delay less than 10 ms .

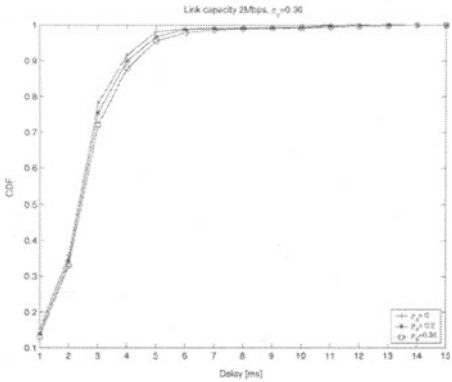

Fig. 8. Cumulative Distribution Function of the total delay of the voice traffic

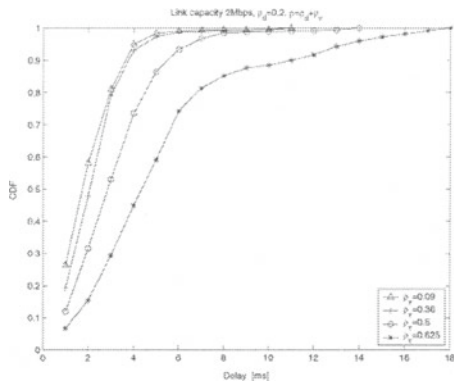

Fig. 9. Cumulative Distribution Function of the delay of the data traffic

6 Dimensioning of the UTRAN

The volume of incoming traffic, be it voice or data or both, present in the UTRAN is typically dictated by the air interface following the availability of the air resources. The latter are scarce and expensive and so the UTRAN wired capacity should not be restrictive. Instead, proper ressource provisioning following accurate dimensioning should be implemented so as to accommodate the demand.

Consider traffic i, $i \in \{v, d\}$, standing for voice and data respectively and for a QoS criterion defined as the probability of delay violation by any type of traffic, i.e., $P(D_i > b_i) \leq \epsilon_i$, where b_i is the maximum delay that can be tolerated by traffic i in the UTRAN. We can, at this point, distinguish between three cases :

1. presence of voice traffic only in the UTRAN. In this case, b_v =5ms. Let ρ_{0v} denote the maximum load of voice traffic that the UTRAN can accomodate in this case.
2. presence of data traffic only. In this case, b_d =10ms. Let ρ_{0d} denote the maximum load of data traffic that the UTRAN can accomodate in this case.
3. presence of both voice and data traffic in the UTRAN. PQ is here to prioritize voice over data. In this case, in the worst case, the voice delay bound is $(b_v - \frac{s_d}{C})$ ms due to the non pre-emptive nature of PQ. Let ρ'_{0v} be the maximum load of voice traffic corresponding to this delay budget. As of data traffic, the delay bound is $(b_d - b_v)$; let ρ'_{0d} be the load of data traffic in this case.

For an offered load $\rho = \rho_v + \rho_d$, proper dimensioning should satisfy the following constraints:

- for $\rho_d = 0$, $\rho \leq \rho_{0v}$.
- for $\rho_v = 0$, $\rho \leq \rho_{0d}$.
- for $\rho_d > 0$ and $\rho_v > 0$, $\rho_v \leq \rho'_{0v}$ and $\rho_d \leq \rho'_{0d}$.

7 Conclusion

In this work, we developed an analytical model for IP service differentiation in the UTRAN where both real-time voice and non-real-time data traffic are to meet severe QoS constraints in terms of delay. We validated our model on a local test-bed emulating the UTRAN transport functionalities. Our results show that PQ is essential to make the voice traffic meet its delay requirements. PQ ensures not only service differentiation between voice and data traffic but also helps minimize jitter for voice traffic. We also drew proper dimensioning for the UTRAN so as to make voice and data traffic meet their delay constraints.

References

1. 3GPP TS 25.430 V3.6.0 (2001-06) 3rd GPP; Technical Specification Group Radio Access Network; UTRAN Iub Interface: General Aspects and Principles (Release 1999).
2. H. Saito, Performance Evaluation and Dimensioning for AAL2 CLAD, IEEE INFOCOM'99, New-York, 1999.
3. A-F. Canton, S.Tohmé, D. Zeghlache, T. Chahed, Performance analysis of ATM/AAL2 in UMTS Radio Access Network, in press, IEEE PIMRC 2002, Lisbon, 2002.
4. R. Makké, S. Tohmé, J-Y. Cochennec, S. Pautonnier , Performance of the AAL2 Protocol within the UTRAN, IEEE ECUMN 2002.
5. O. Isnard, et all, Handling Traffic Classes at AAL2/ATM layer over the logical Interfaces of the UMTS Terrestrial Radio Access Network, IEEE PIMRC, 2000.
6. S. Nananukul, S. Kekki, Simulation Studies of bandwidth Management for the ATM/AAL2 Transport in the UTRAN. Vehicular Technology Conference, 2002.
7. 3GPP TS 25.933 V1.7.1 (2002-01), IP Transport in UTRAN Work Task Technical Report.
8. Mobile Wireless Internet Forum IP in the RAN as a Transport Option in 3rd Generation Mobile Systems, Technical Report MTR-006, Release v2.0.0, Ratified June 18, 2001.
9. A. Samhat, T. Chahed, Performance Evaluation of IP in the UMTS Terrestrial Radio Access Network, In Proc. 18th International Teletraffic Congress (ITC), Berlin, September 2003.
10. A. Samhat et all, Transport in UMTS Radio Access Network: IP versus AAL2/ATM, 8th International Conference on Cellular and Intelligent Communications (CIC), Seoul, 2003.
11. L. Kleinrock, Queueing Systems, Volume I: Theory, John Wiley and Sons, 1975.
12. D. Bertsekas, R. Gallager, Data Networks, second edition, Prentice-Hall, 1992.
13. Home page of ALTQ project http://www.csl.sony.co.jp/ kjc/ projects.html.