Downlink Packet Scheduling for a Two-Layered Streaming Video Service in UMTS

Jordi Pérez-Romero, Oriol Sallent, Ramón Agustí
Universitat Politècnica de Catalunya (UPC),
Dept. of Signal Theory and Communications, Campus Nord
Barcelona, Spain
Tel: +34 93 401 71 95, email: [jorperez, sallent, ramon]@tsc.upc.es

ABSTRACT
This paper presents a packet scheduling algorithm to manage the quality of service of a two layered video service in the downlink FDD mode of UMTS. The algorithm takes into account the expected interference and the OVSF code usage to schedule appropriately the different transmissions.

I. INTRODUCTION
W-CDMA access networks, such as the considered in UTRA-FDD proposal [1], provide an inherent flexibility to handle the provision of future 3G mobile multimedia services. In this scenario, Radio Resource Management (RRM) strategies will play a key role when optimizing air interface utilization. They are particularly relevant when dealing with packet multimedia flows that face different QoS requirements. Within this context, in this paper we focus on the QoS provision by means of UMTS of a downlink streaming video service that has two different quality layers, a basic flow that provides the minimum quality requirements and an enhancement layer that supplies additional information to improve reception depending on the available bandwidth [2]. Consequently, appropriate scheduling algorithms should be devised to manage the transmissions of the different flows while maximizing the use of the scarce downlink radio resources. It is worth noting that few studies aligned with 3GPP specifications are available in the open literature dealing with this problem. So the paper is organized as follows: in section II an overview of the downlink RRM is provided emphasizing how the power levels and codes are shared. Decisions about who should transmit and its transmission parameters (i.e., transport format TF and power level) are the responsibility of the packet scheduler. Thus, any downlink RRM strategy should take into account the following two aspects to work properly:

II.A. Interference Management
Within a W-CDMA cell, all users share the common bandwidth and each new connection increases the interference level of other connections, affecting their quality expressed in terms of a certain $(E_b/N_0)$. For $n$ users transmitting simultaneously at a given cell, the following inequality for the i-th user must be satisfied:

$$P_t^i + \chi_i + \rho \times \left( \frac{P_t^i - P_{th}}{L_x(d_i)} \right) \geq \left( \frac{E_b}{N_0} \right)_i$$

(1)

$P_t^i$ being the base station transmitted power, $P_{th}$ being the power devoted to the i-th user, $\chi$ representing the intercell interference observed by the i-th user, $L_x(d_i)$ being its path loss, $r$ the channel coding rate and $P_N$ the background noise. $SF$ relates the bit duration to the chip period. $\rho$ is the orthogonality factor due to the fact that orthogonal codes are used in the downlink direction. Differently from the uplink case, in downlink the intercell interference is user-specific since it depends on the user location, the base station transmitted power is shared by all users and the power allocations depend on the user location as well. Then, it is obtained that:

$$P_{th} \geq \frac{P_w + \chi_i + \rho \times \frac{P_t^i}{L_x(d_i)}}{SF}$$

(3)
Adding all $n$ inequalities it holds that the total transmitted power to satisfy all the users demands should be:

$$
\sum_{i=1}^{n} \left( \frac{P_n + \chi_i}{SF_i} \right) L_p(d_i) + \rho \left( \frac{E_r}{N_o} \right)_i r 
$$

$$
P_{T, max} \geq P_T = \frac{1 - \sum_{i=1}^{n} \frac{\rho}{SF_i} + \rho \left( \frac{E_r}{N_o} \right)_i r}{1 - \sum_{i=1}^{n} \rho} \left( \frac{E_r}{N_o} \right)_i r
$$

Claiming in (4) for the inherent positivity of $P_N$ (i.e. $P_N > 0$) leads to:

$$
\eta_{DL} = \sum_{i=1}^{n} \left( \frac{\rho \times L_p(d_i)}{SF_i} \right) P_T < 1
$$

The later expression is commonly known as the downlink load factor [4]. Additionally, physical limitations into the power levels are given by the maximum base station transmitted power, $P_{T, max}$. The total transmitted power by the base station can be expressed in terms of the load factor as:

$$
P_T = \frac{P_n \sum_{i=1}^{n} L_p(d_i)}{(1-\eta_{DL}) \sum_{i=1}^{n} \frac{\rho}{SF_i} + \rho \left( \frac{E_r}{N_o} \right)_i r}
$$

where it can be observed that as the load factor increases the power demands also increase. Notice that, depending on how users are distributed in the cell, the downlink load factor is modified and also the required transmitted power varies.

II.B. Code Management

Apart from managing appropriately the power levels, another important scarce resource in the downlink are the OVSF codes. According to the properties of these codes, their availability is guaranteed whenever the Kraft’s inequality is fulfilled, given by [5]:

$$
\sum_{i=1}^{n} \frac{R_{b,i}}{R_b} \leq SF_{max}
$$

where $n$ is the number of users, $R_{b,i}$ their transmission bit rates and $R_b$ the minimum bit rate (corresponding to spreading factor $SF_{max}=512$). In any case, the above inequality only guarantees the code availability, but in certain cases, depending on how codes are assigned some reallocations may be required.

III. SERVICES AND TRANSPORT CHANNELS

In UTRA FDD there are three types of channels to carry out downlink transmissions, namely [1]:

a) DCH (Dedicated CHannel): devoted to services with stringent transfer delay requirements, such as conversational services.

b) DSCH (Downlink Shared CHannel): devoted to services with tolerant transfer delay requirements, such as interactive services. It is associated to a DCH channel through which physical layer control information is transmitted. Transmission through these channels is subject to a packet scheduling policy.

c) FACH (Forward Access CHannel): devoted to services without QoS requirements.

Depending on the type of service to be provided, the previous channels should be managed and allocated appropriately. In this paper we are focusing on streaming video services, whose quality requirements deal with the achieved bit rate, the percentage of lost packets and the jitter of the delay rather than the end-to-end delay. It is considered that streaming service allows an initial set-up delay that gives room to some packet transmissions before the video is reproduced. These packets can be stored in the mobile terminal buffer and the reproduction rate can be adjusted to the source rate. Then, the user can be unaware of the possible packet retransmissions because the stored buffer allows for a continuous packet flow. Of course, the retransmission capability would be limited by the initial buffering. Thus, this property gives some more room for scheduling the streaming service as packet retransmissions may play a role.

In order to differentiate quality levels, we assume for this service a two layered video application that is characterized by two different flows: a basic layer, with the minimum requirements for a proper visualization, and an enhancement layer, that contains additional information to improve the quality of the received images. We will assume that the basic layer will be transmitted through DCH channel while the enhancement layer will be transmitted only if there is capacity in the DSCH channels. For a certain user, the DSCH will be associated to the DCH channel carrying the basic layer. For this service a constant bit rate generation model is assumed. In order to save resources, the DCH channel operates at a fixed bit rate equal to the source bit rate, which means that a fixed number of transport blocks should be transmitted in each Transmission Time Interval (TTI). Consequently, there is no margin for carrying out retransmissions in the DCH channel. However, retransmissions may be useful to avoid having a very stringent BLER target that would limit the number of users in the system. Furthermore, the end-to-end delay of the service allows the use of retransmissions. As a result, it is assumed that the possible retransmissions of the basic layer can be carried out in the DSCH channel together with the enhancement layer, and having a higher precedence than the latter.

IV. PROPOSED PACKET SCHEDULING ALGORITHM

The proposed strategy allocates resources to the different flows that make use of the DSCH channel. It operates on a frame by frame basis (i.e., a frame is 10
worth noting that the quotient transmitted Transport Blocks in the previous TTI. It is the Service Credit in the previous TTI, which depends on the BLER target to be achieved – a measurement of each user’s path loss $L_p(d_i)$ – a measurement of the other-to-own cell interference factor for each user:

$$f_{DL,\lambda} = \frac{\chi \cdot L_p(d_i)}{P_T}$$ (8)

- the number of current transmissions in DCH channels, together with their corresponding transport format (TF) and Eb/No target

Taking into account all these parameters, and according to a generic scheduling behavior presented in [6] the algorithm performs the following steps in each frame:

**IV.A. Prioritization**

The first step consists in ordering the different users’ requests in the DSCH depending on some priority criterion that takes into account the required QoS of each user. In particular, the priority table is derived from highest to lower priority according to:

a. The higher the number of basic layer TBs to be retransmitted the higher the priority will be.

b. For the same number of basic layer TBs, the priority is established according to the service credit concept, explained below. The higher the service credit of the enhancement layer the higher the priority.

The service credit concept consists in monitoring the QoS that each flow has received in terms of bit rate and measuring the difference between the expected bit rate and the offered bit rate. The higher the difference, the higher the resources to be allocated (or equivalently the number of transport blocks that should be transmitted).

So the “credit” that the system owes to the flow should be computed. This leads to the definition of the “service credit” (SCr) [7], that accounts for this difference and can be computed as follows in each TTI:

$$SCr(k) = SCr(k-1) + \frac{R_g}{TB} - NumTx(k-1)$$ (9)

where $SCr(k)$ is the Service Credit for TTI=$k$, $SCr(k-1)$ is the Service Credit in the previous TTI, $R_g$ is the guaranteed bit rate measured in bits/TTI, $TB$ is the number of bits of a Transport Block for the considered RAB and $NumTx(k-1)$ is the number of successfully transmitted Transport Blocks in the previous TTI. It is worth noting that the quotient $R_g/TB$ reflects the mean number of transport blocks that should be transmitted per TTI in order to keep the guaranteed bit rate. As a result, $SCr(k)$ is a measure of the number of Transport Blocks that the connection should transmit in the current TTI to keep the guaranteed bit rate. For example, if $TB = 320$ bits, $R_g = 32$ kb/s, and TTI=40 ms, 4 service credits are added in each TTI.

After computing the service credit, and assuming a total of $x$ transport blocks in the buffer, the number of transport blocks to be transmitted in the current TTI would be:

$$numTB = \min(x, SCr(k), TB_{max})$$ (10)

$TB_{max}$ being the maximum number of Transport Blocks that can be transmitted per TTI depending on how the RAB is defined. Finally, the selected TF would be the one that allows to send $numTB$ blocks.

The output of this phase is an ordered list of requests for the different users, each containing a TF value.

**IV.B Resource allocation**

Once requests are ordered, the next step consists in deciding whether or not they are accepted for transmission in the DSCH channel and which is the accepted TF. The limitations explained in section II dealing with interference and code availability are taken into account in this phase. To this end, it is required to estimate the expected load factor and transmitted power level once all the requests are accepted. Then, the expected load factor whenever there are $n$ transmissions in the system in frame $t$ (including both DCH and DSCH transmissions) is:

$$\eta(n,t) = \sum_{i=1}^{n} \left( \frac{\rho + f_{DL,i}(t-1)}{SF_i} \right) + \rho \left( \frac{E_p}{N_0} \right)_{\phi}$$ (10)

Similarly, the expected power is given by:

$$\bar{P}_T(n,t) = \frac{P_S}{(1-\eta(n,t))} \sum_{i=1}^{n} \frac{L_p(d_i)}{SF_i} + \rho \left( \frac{E_p}{N_0} \right)_{\phi}$$ (11)

It should be pointed out that the differences between the expected load factor and the real value can be due to the inaccuracies in the measurement of the other-to-cell interference factor $f_{DL,i}$ and the path loss.

With this restrictions in mind, the algorithm executes for each request the rules in figure 1, assuming a total of $n$ already granted transmissions. At the beginning, for the initially selected TF, the Kraft’s inequality (7) is evaluated, afterwards, the expected load factor is compared with a threshold $\phi$ and finally the expected transmission power level should be below a fraction $\delta$ of the maximum transmitted power. If all three conditions hold, transmission is granted for this request during one TTI, otherwise, the transport format is reduced by one, or equivalently, the transmission bit rate is reduced. If this is not possible, the request should wait for the next frame.
It should be mentioned that control parameters $\phi$ and $\delta$ (both <1) should be appropriately set in order to take into account the possible fluctuations between the expected values and the real measurements.

$Kraft's$ $inequality$ $with$ $n+1$ $users$

$\eta(n+1,t) < \phi$

$P_t(n+1,t) < P_{t,\max}$ $\delta$

\begin{tabular}{|c|c|}
\hline
TrCH type & DSCH \\
\hline
TB sizes, bit & 320 bits (payload) + 16 bits (MAC/RLC header) \\
\hline
TFS & \begin{tabular}{l}
TF0, bits \ 0x320 \ \\
TF1, bits \ 1x320 (8 Kbps) \ \\
TF2, bits \ 2x320 (16 Kbps) \ \\
TF3, bits \ 4x320 (32 Kbps) \ \\
TF4, bits \ 8x320 (64 Kbps) \ \\
TF5, bits \ 16x320 (128 Kbps) \\
\end{tabular} \\
\hline
TTI, ms & 40 \\
\hline
\end{tabular}

The characterization of the physical layer, including the rate 1/3 turbo code effect is taken from [9]. The BLER target is set to 1%. The total number of users in the scenario has been varied between 100 and 160 (for less than 100 users both the streaming and enhancement layers can be provided at their expected rate of 32 kbps, so the interest in the analysis arises when the system has more than 100 users).

One of the most relevant parameters in the design of the packet scheduling algorithm relays on the threshold $\phi$ of the estimated load factor $\eta(n+1,t)$ when deciding the granted transmissions. Particularly, if $\phi$ is too high, the difference between the estimated and the real values can lead the system to a situation where no available power exists that satisfies at the same time all the users requirements, thus obtaining BLER values higher than the target one for both basic and enhancement layers. On the other hand, if $\phi$ is too low less problems will exist for basic transmissions at the expense that a lot of enhancement requests will be postponed. This trade-off can be observed in Figs. 2 and 3. The first one presents the average bit rate obtained during a streaming session for the enhancement layer depending on threshold $\phi$. \(\phi=1\) has been assumed. The basic layer is not presented since its achieved bit rate is always 32 kbps, with a slight reduction for the case 160 users and $\phi=1$. This reduction can be observed in Fig. 3 where the percentage of lost packets (due to expiring the maximum delay) is presented for both basic and enhancement layers as a function of the number of users. The reason for this degradation is the increase in the BLER due to errors in the estimation of the offered load. Something similar occurs for the enhancement layer, where most of the packets are lost. From both figures it can be concluded that the selection $\phi=0.95$ provides the best behavior since it achieves the maximum bit rate for the enhancement without degrading the quality of the basic layer. Notice also that thanks to the retransmissions, the packet loss ratio is zero for the basic flow whenever $\phi<=0.95$ even for high loads.
Furthermore, Fig. 4 shows the jitter of the packet delay, that is one of the main QoS requirements for a streaming service. For the basic flow, it can be observed that, a part from the case $\phi=1$, the jitter is below 1 TTI (40 ms). For the enhancement case, the jitter is somewhat higher than the basic due to the packet scheduling operation. The value of the jitter impacts the buffer dimensioning at the receiver side to guarantee that each flow is delivered to the user in a continuous way. Consequently, the maximum allowable jitter will depend on the specific buffer capabilities.

VI. CONCLUSIONS
This paper has presented a packet scheduling strategy that deals with the provision of QoS to a two layered video streaming application by jointly considering the use of DCH and DSCH channels. It has been shown that managing retransmissions appropriately can be suitable to reduce packet losses for both layers. In order to decide the granted transmissions, the algorithm takes into account an estimation of the expected load factor, that must be below a threshold. The influence of this threshold has been analyzed to maximize the rate of the enhancement flow while at the same time meeting the requirements of the basic layer.

ACKNOWLEDGEMENTS
This work is part of the ARROWS project, partially funded by the European Commission under the IST framework and by the Spanish Research Council under grant TIC2001-2222.

REFERENCES
[1] 3GPP TS 25.211, “Physical channels and mapping of transport channels onto physical channels (FDD)”