SERVICE MULTIPLEXING OVER A TDMA AIR INTERFACE USING THE FLO CONCEPT

Glenn Platt & Jamil Y. Khan

School of Electrical Engineering & Computer Science

University of Newcastle

Callaghan, NSW, 2308, Australia

Email: glenn@ecemail.newcastle.edu.au, jkhan@ecemail.newcastle.edu.au

Abstract: This paper presents a scheduling scheme for a FLO based advanced TDMA system to enable service multiplexing from user terminals. The FLO specifications allows different traffic types to use appropriate transport formats to meet QoS requirements of each class of traffic. Service multiplexing can be achieved by selecting appropriate transport formats for each traffic class and then combining them in a single packet and transmitting over a single slot of a TDMA air interface. The paper introduces a scheduling scheme known as LQP which can offer higher throughput for using a FLO based air interface. The paper presents some initial simulation results obtained from an OPNET based simulation model.

I. INTRODUCTION

A major enhancement of the GSM/EDGE radio access network (GERAN) is the Flexible Layer One (FLO) concept, currently being considered for Release 6 of the 3GPP specifications. An entirely new physical layer protocol for GERAN, FLO greatly simplifies the introduction of new services to the GERAN system, allowing for "tailor-made" quality of service provision, significantly enhancing GERAN's flexibility, and ensuring it can continue to provide services on-par with other 3G systems.

The introduction of FLO provides a number of opportunities for unique new resource management techniques, particularly in the area of resource management and multiplexing within the traffic flows of one mobile station, these opportunities being the focus of this paper. Service multiplexing is a process where multiple logical connections can be mapped on a single physical link using appropriate source and channel coders. With the increase of transmission slot capacity of a GSM/EDGE air interface it is possible to accommodate more than one type of information bits on the UP and/or DOWN links to enable service multiplexing. To efficiently transmit different classes of information bits on the air interface it is necessary to select appropriate channel coding scheme for successful transmission. Using the FLO approach different traffic sources can select appropriate transport format and share a transmission slot with other traffic sources. However, the throughput of each traffic source depends on the transport format selection and the scheduling mechanism used in a transmitter. In this paper we propose a new scheduling mechanism known as Longest Queue Priority (LQP). Higher throughput can be achieved by combining the LQP technique

and the transport format selection technique. This mechanism is outlined in the sections below.

II. FLO'S IMPROVED MULTIPLEXING FUNCTIONALITY

As detailed elsewhere [1], FLO removes the fixed-configuration logical channels offered by the physical layer. Rather than logical channels with a fixed transport configuration, with FLO the GERAN physical layer now offers transport channels (TrCHs) to the MAC layer. The transport configuration of these transport channels is specified at call setup, and thus the transport configurations available for a particular connection can be tailored to the traffic to be carried. To limit complexity, a limited number of transport formats are allowed per transport channel, a valid combination being denoted a *transport format combination* (TFC). The set of valid transport format combinations across all transport channels is referred to as the *transport format combination set* (TFCS).

FLO can significantly improve the multiplexing functionality of the GERAN system. With FLO, multiple traffic streams can be carried by multiple transport channels that are multiplexed together on a time slot before being transmitted across the radio interface. As part of this multiplexing operation, FLO allows for prioritising between traffic sources on various transport channels, through the *rate matching* operation. The rate matching operation "adjusts" the size of an encoded data block on a transport channel so that when the block from this transport channel is multiplexed with the blocks from all the other concurrent transport channels, the resulting data will fit in one radio packet. The rate matching operation affects the QoS given to data on a particular transport channel- the size "adjustment" is done by bit puncturing (the removal of bits in the coded sequence) or repetition (the repetition of bits in the coded sequence), thus altering the size of the encoded block, and changing the level of protection given by the channel coding. The QoS provided to data on a particular transport channel is determined by a *rate matching attribute* (RMA), which controls how much data on a particular transport channel will be punctured or repeated, affecting the level of channel coding protection given to this data, and thus prioritising between data on different transport channels. Thus, when only one transport channel is active at a time, the coding rate for this transport channel only depends on the transport block size and on the available channel bandwidth. However, when more than one transport channel is active, the coding rate for

a particular transport channel also depends on the rate matching attributes associated to the other transport channels.

With its rate matching operation, the introduction of FLO to GERAN brings an entirely new transport format management and multiplexing role to the GERAN mobile station. With FLO, the mobile station MAC layer must control the access of concurrent upper-layer traffic streams to the limited transport formats available at any one time. Such an operation is non-trivial, as not only must the controlling entity consider issues such as what TFC to use and when to allow a particular upper layer stream to access a certain transport format in that TFC, it must also consider the inherent trade-offs unique to the rate matching operation that is the heart of the FLO multiplexing technique.

III. PRIOR WORK IN TRANSPORT FORMAT SCHEDULING

There has recently been several research works looking at the issue of scheduling transport formats for concurrent traffic streams within one MS. These publications have all been based on a fixed TFC allocation scheme suggested in the 3GPP specifications [2], where the TFC to be used is selected based on that TFC's ability to offer the best possible QoS to the highest priority data to be carried. If a number of TFCs offer the same best possible QoS to the highest priority data, the TFC is selected based on the ability of each TFC to offer the best possible QoS to the next highest priority data, and so on.

Of particular interest to our work is the work of S. Heier *et al* [3, 4, 5], that presents a number of interesting studies of MAC layer TFC management functionality in UTRAN. These publications all use a fixed TFC allocation technique, similar to that presented in the GERAN standards, for controlling access of the various active TrCHs to the available TFCs. The TFC allocation mechanism proposed in these references only moves between TFCs to match a significant shift in system load, for example when one TrCH stops transmitting for a short time. This early work has a number of limitations- the research does not appear to consider the effect of time-varying channels or transmission errors on the performance of the transport format management system, and the authors focus on the interaction between their proposed system and mechanisms such as TCP, rather than investigating metrics such as the QoS performance or multiplexing gains of their proposals.

Importantly, the research presented in references [3, 4, 5] identifies the impact transport format management schemes can have on the performance of a system such as FLO. However, the fixed TFC allocation scheme studied in the literature thus far is clearly relatively simple, and limited in its capability. The major limitation of this simple, priority-queuing style technique is inherent to its design- it is a fixed scheme, and as such cannot adapt its behaviour to changing conditions. Considering variable channel conditions, bursty traffic loads and the error-prone wireless environment, a dynamic TFC allocation mechanism that can adapt to such changing conditions should outperform the fixed scheme studied in the literature thus far. For the remainder of this paper such an allocation mechanism is referred to as a TFC

scheduler, where the interest is on scheduling access to the transport formats available in the TFCS.

One possible transport format scheduling scheme is outlined in the following section.

IV. A TRANSPORT FORMAT SCHEDULING SCHEME

At this point in time, it is worthwhile focusing the scope of these investigations. Reflecting the scheduling issue of interest, attention is focused on one individual mobile station, operating on a single GERAN dedicated basic physical subchannel (DBPSCH), carrying conversational, interactive and background classes of traffic, which represents a realistic UMTS traffic mix.

As per the GERAN FLO specifications, the maximum number of TFCs in the system is limited to 32 [2]. In scheduling between transport formats during an ongoing connection, the TFC scheduler is actually scheduling between transport formats with different rate matching attributes. For readability purposes, a particular transport format combination is denoted by the rate matching attributes assigned to each TrCH in the TFC. For example, with three TrCHs operating concurrently, TFC "532" has a rate matching attribute (RMA) of 5 for the first TrCH (denoted TrCH A), an RMA of 3 for the second TrCH (TrCH B) and an RMA of 2 for the third TrCH (TrCH C). In this TFC, with the highest rate matching attribute, TrCH A has the highest priority, followed by TrCH B and then TrCH C.

For manageability, and to allow an intuition as to the relative priority of a particular TrCH by inspection, the RMA of any TrCH is forced to lie in a constant range, between 0 and 5. With the GERAN specifications limiting the number of TFCs available, TrCH A, which is assigned to carry high priority conversational class traffic, is always assigned an RMA of 5, thereby ensuring it has maximum priority. By fixing the RMA of TrCH A, the TFC scheduling mechanism only ever operates on TrCHs B and C. Thus, it is the responsibility of the scheduling mechanism to select a particular TFC (be it 532, 505, 513, etc.) to be used for a particular transmission time interval (TTI). In doing this, the scheduling mechanism must prioritise amongst the three TrCHs, trying to maximise the QoS of the interactive and background traffic TrCHs, whilst also maintaining the QoS of the high priority conversational class TrCH.

To ensure the QoS of traffic on TrCH A, a set of TFCs denoted the *valid* set is selected, where all TFCs in the valid set can ensure the block error rate (BLER) of conversational class traffic is maintained less than 3% in the given channel conditions. Given the valid set of TFCs, and the current queue sizes of each TrCH, the scheduler must decide which particular TFC is best used to carry the data operating on the three TFCs, in order to maximise efficiency and utilisation, whilst maintaining QoS.

The proposed scheduling scheme comprises two parts- the dynamic scheduling algorithm itself, and a throttling mechanism. These are detailed in the sections below.

A. The LQP scheduler

The proposed scheduling scheme is based on the longest queue first (LQF) techniques, commonly used in packet scheduling mechanisms. However, unlike most LQF packet schedulers, which can block the access of the shorter queues to the transmission medium, the proposed scheme generally always transmits a packet from each queue, every TTI. QoS differentiation is performed by giving the queue with the longest length priority in TFC selection, and thus the proposed scheme is referred to as the *longest queue priority* (LQP) scheme.

In the LQP scheduling scheme, the scheduler selects between TFCs by first evaluating the weighted length of each queue, WL_i :

$$WL_i = \frac{new_i}{w_i} + ARQ_i, \quad w_i \in \mathbb{N}^+ \quad (1)$$

where new_i is the number of new (unsent) packets in the transmission queue of TrCH i , w_i is a user (operator) defined weighting factor (or *service priority*) of TrCH i , and ARQ_i is a count of the number of ARQ (retransmitted) packets awaiting retransmission for TrCH i .

The scheduler then compares the weighted length of all the TrCH queues, and decides which TFC to use based on the relative difference between the calculated weighted lengths. To reduce a particular TrCH's queue size, this TrCH should be offered the most favourable (least prone to error) transport format available, and thus, in the LQP scheme, the TrCH with the longest weighted queue length gains priority in the selection of which TFC to use for a particular TTI. In comparing TrCH queue sizes, the LQP scheduling algorithm works to keep the weighted TrCH queue lengths the same size, and which TFC is used at a particular time depends on the magnitude of the difference in weighted queue lengths. This difference, Δ_{WL} , is compared to predefined thresholds, denoted *switchpoints*, which determine which TFC to use. As an example, consider the set of valid TFCs contains RMAs 522, 530, 532, 503, and 523. If, Δ_{WL} is greater than switch point 1, the lowest threshold, then the LQP scheduler will use TFC 522, indicating the two TrCHs have relatively similar weighted queue lengths, and giving the same priority to both interactive and background traffic. If Δ_{WL} is less than switch point 1, but greater than switch point 2 (the middle threshold), indicating (for example), the interactive TrCH has a *slightly* larger weighted queue length than the background TrCH, then the scheduler would choose TFC 532, prioritising interactive traffic by offering it a more favourable (less prone to error) transport format. If Δ_{WL} is less than switch point 3, indicating the interactive TrCH queue's weighted length is *significantly* longer than that of the background TrCH, then the scheduler would use TFC 530, thereby not sending any background traffic at all, and greatly prioritising the interactive TrCH. Naturally, the reverse operation also applies.

Regarding erroneous packets, if traffic on one TrCH experiences a significant number of transmission errors, the weighted length of its scheduler queue will grow large due to the number of ARQ packets added to it. The LQP scheduler

Figure 1: LQP TCP scheduling system architecture.

will then strive to reduce the size of this larger queue, prioritising its packets by selecting an appropriate TFC that offers this queue a more favourable transport format.

B. Throttling mechanism

Whilst it is assumed appropriate call admission control mechanisms will not admit a new connection whose traffic load cannot be supported with current system load, a second mechanism, a throttling mechanism, is also introduced to ensure isolation between input streams in the proposed LQP scheduler.

The throttling mechanism works on the well-known token bucket basis. Such token bucket mechanisms have been studied widely, particularly for traffic shaping and QoS differentiation purposes [6, 7, 8].

QoS differentiation between the various TrCHs can also be performed at the throttling mechanism. By controlling the rate at which packets from the different TrCHs are passed on to the scheduler, it is possible to prioritise individual streams by varying their rate. Traffic with a high service rate will result in a longer queue size at the scheduler (compared to the queue size for the low rate traffic, and ignoring any ARQ packets), thus influencing the LQP scheduling algorithm's TFC decision (as the scheduler will choose a TFC that prioritises the traffic with the longer weighted queue length), and further prioritising the high priority traffic. This functionality can also apply to the short-term prioritisation of ARQ traffic- in times of significant errors, it is possible to throttle-back the number of new packets arriving at the scheduler, thus providing a higher weight to ARQ packets in the weighted queue evaluation. Figure-1 shows the operation of the LQP mechanisms.

V. PERFORMANCE ANALYSIS

To analyse the performance of the proposed TFC scheduling scheme, a detailed OPNET simulation model has been constructed. An event-based model, the simulation includes mobile stations, a base station, and realistic traffic

and channel models, further details of which can be found in [9]. Traffic is generated at sources, passes through the FLO mechanism, where it is channel coded, rate matched and multiplexed, and then transmitted across the radio interface. Received traffic is then evaluated for transmission errors, end-to-end delays, and so on. For comparison purposes, the proposed LQP TFC scheduling scheme is studied together with the fixed TFC selection mechanism suggested in the 3GPP specifications.

First evaluated is the performance of the two TFC management schemes in supporting high priority conversational class traffic. In this experiment, the simulation is run for a fixed length of time, and the effective throughput, percentage of blocks in error (BLER) and delays for transmitted traffic is evaluated. It was found that both schemes perform identically, and maintain very low conversational class traffic end-to-end delays of 0.005s- the arrival rate of conversational class traffic is significantly less than the service rate of the TFC scheduling mechanism. Similarly, both schemes maintained isolation between the conversational and other traffic classes- as the interactive class traffic load is increased, there was no noticeable effect on conversational class traffic delay. Next evaluated was the average BLER for conversational class traffic using the two TFC management schemes. It was found that both schemes maintain conversational class traffic loss well below 2% for all traffic loads, which is generally considered sufficient to maintain a conversational class connection [9].

The net effective throughput of the TFC management schemes is shown in figure 2. As can be seen from the figure, the fixed and LQP schemes offer the same average throughput. This is expected, as, on average, the two schemes will both remove a packet from all three TrCHs every TTI, and with the same channel conditions, the average number of transmission errors for the two schemes will also be the same.

Figures 3 and 4 show the interactive and background class traffic end-to-end delays for a variety of traffic loads. Figure 3 is zoomed in to the point where interactive class end-to-end delays grow very large- at very high loads (average interactive class packet arrival rates of greater than 60 packets/second), the average arrival rate of packets to the system is greater than the service rate of the scheduler, and thus the system becomes excessively congested.

Considering interactive class traffic delays, it can be seen that the LQP and fixed schemes perform similarly, with low end-to-end delay for interactive class traffic, provided the amount of arriving traffic does not exceed the capacity of the system.

Considering background class traffic delays, it can be seen that the LQP scheme maintains lower background class end-to-end delay, for all loads, than the fixed scheme. This is due to the flexibility of the LQP scheme- unlike the fixed scheme, the LQP scheme is able to adapt to changing channel and traffic conditions, and when conditions permit, select a TFC that provides better QoS to background class traffic, increasing the effective service rate of this traffic. Table I presents the average BLER for traffic carried by each

Figure 2: Net throughput for conversational class traffic.

Figure 3: Interactive class traffic average end-to-end delay

TABLE I

	Conversational class BLER	Interactive class BLER	Background class BLER
Fixed	0.025	0.046	0.145
LQP	0.025	0.065	0.093

of the schemes in this experiment. As can be seen, the LQP scheduler offers significantly better service to background class traffic than the fixed scheme, whilst offering comparable service to interactive class traffic. It is worthwhile noting that the slight difference in delay performance of the LQP and fixed schemes in carrying interactive class traffic can be attributed to slight differences in how the two schemes prioritise interactive class traffic against background class traffic. In the fixed scheme, the priority of interactive class traffic over background class traffic is predetermined by the design of the scheme, and cannot be changed. In the LQP scheme, the priority of interactive class traffic over background class traffic can be adjusted by altering the factor service priority factor w_i . This is shown in figure 5, where the factor $DelayRatio = \frac{BackgroundDelay}{InteractiveDelay}$ was calculated for different

background class traffic service priorities. As can be seen from figure 4, the background service priority parameter allows significant flexibility in the operation of an LQP scheduling system, allowing an operator to tightly control how various traffic classes are treated in the system. Figure 6 shows how different token generation rates in the throttle mechanism can also be used to adjust the QoS provisioning

Figure 4. Background class traffic average end-to-end delay

Figure 5. DelayRatio for different background class traffic service priorities

Figure 6: Background class traffic end-to-end delay for different token interarrival times

of the system- slow background class token generation rates will increase the delays experienced by background class traffic, improving the service available to interactive class traffic.

The QoS provisioning features discussed above are a significant advantage of the proposed LQP scheme- it offers the same QoS to conversational and interactive class traffic to the fixed scheme suggested in the GERAN standards, yet offers significantly better service to background class traffic, whilst providing significantly improved QoS provisioning flexibility.

VI. CONCLUSIONS

This paper has presented a multi-service traffic scheduling mechanism on an advanced TDMA air interface. Simulation results show that the proposed scheme can offer higher net throughput for multi-service traffic transmission, and improved flexibility when compared to the fixed TFC management scheme suggested in the GERAN specifications. Advanced QoS provisioning is facilitated by the service priority and token generation process in the LQP scheduler, whilst service isolation is inherent to the scheme's design. The proposed LQP algorithm is scalable and simple to implement. The proposed solution can offer significant benefit when used in a high data rate channel.

REFERENCES

- [1] G. Platt, K. Pedersen and B. Sebire, "Introducing the GERAN Flexible Layer One concept," *IEEE Wireless Communications Magazine*, vol. 10, issue 3, pp. 18-25, June 2003.
- [2] 3GPP TR45.902, Technical specification group GSM/EDGE radio access network; Flexible Layer One (Release 6), v6.5.0, April 2004.
- [3] S. Heier and M. Malkowski, "Interaction between UMTS MAC scheduling and TCP flow control mechanisms," *Proc. ICC'03*, vol. 2, pp. 1373-1376, April 2003.
- [4] S. Heier and M. Malkowski, "UMTS radio resource management by transport format assignment and selection," *Proc. of the 5th IEEE International Symposium on Wireless Personal Multimedia Communications*, vol. 3, pp. 1187-1191, October 2002.
- [5] S. Heier, C. Ellerbrock, and M. Malkowski, "UMTS medium access control quality of service scheduling," *Proc. IEEE PIMRC '02*, vol. 4, pp. 1893-1897, September 2002.
- [6] W. K. Wong, H. Zhu, and V. C. M. Leung, "Soft QoS provisioning using the token bank fair queuing algorithm," *IEEE Wireless Communications*, vol. 10, no. 3, pp. 8-16, June 2003.
- [7] A. W. Berger, "Performance analysis of a rate-control throttle where tokens and jobs queue," *IEEE JSAC*, vol. 9, no. 2, pp. 165-170, February 1991.
- [8] N. Yin and M. Hluchyj, "Analysis of the leaky bucket algorithm for on-off data sources," *Proc. IEEE GLOBECOM '91*, vol. 1, pp. 254-260, December 1991.
- [9] G. Platt and J. Y. Khan, "QoS implications of the new GSM/EDGE Flexible Layer One," *Proc. DSPCS '03*, December 2003.
- [10] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE performance*. John Wiley & Sons, 2002.