PROVISION OF MBMS OVER THE GERAN: TECHNICAL SOLUTIONS AND PERFORMANCE

Leonardo Provvedi*, Chris Rattray*, Jürgen Hofmann†, Sergio Parolari‡

* Roke Manor Research Ltd., Old Salisbury Lane, Romsey, SO51 0ZN, UK - E-mail: leonardo.provvedi@roke.co.uk, chris.rattray@roke.co.uk
† Siemens AG, St. Martinstrasse 76, D-81541, Munich, Germany - E-mail: juergen-hofmann@siemens.com
‡ Siemens Mobile Communications S.p.A., Via Monfalcone 1, 20092 Milano, Italy - E-mail: sergio.parolari@siemens.com

Keywords: MBMS, GERAN, multicast, broadcast.

Abstract — **3GPP is currently in the process of standardising the Multimedia Broadcast / Multicast Service (MBMS). In this paper the technical solutions adopted for the provision of these services over the GERAN *A/Gb mode* are presented. Simulation results are provided showing the data rates that can be achieved over the radio interface using the data transfer mechanisms already available in Release 5. Some of the concepts (to be included as part of Release 6) that have been proposed in 3GPP to increase the data rates, such as outer coding in the RLC layer or point-to-multipoint (p-t-m) with feedback, are also discussed, and their performance gain demonstrated through simulations.**

1 Introduction

The 3rd Generation Partnership Project (3GPP) is currently standardising the Multimedia Broadcast / Multicast Service (MBMS) as part of the new features to be included in Release 6 of its specifications. The objective of MBMS is the efficient use of the radio resources by allowing the simultaneous distribution of identical multimedia data to multiple receivers using the same radio channel(s). Therefore, in addition to the existing procedures for point-to-point (p-t-p) transmission to a single user, new procedures are defined to support point-to-multipoint (p-t-m) transmission to multiple users.
It is expected that MBMS will generate a significant new revenue stream for operators by allowing the efficient broadcast or multicast of popular multimedia services such as news, traffic information and sports clips.
MBMS is a service provided over the PS domain, and therefore will be provided over the GPRS system. The architecture to support MBMS is shown in Figure 1. New network elements need to be introduced, such as the BM-SC (Broadcast/Multicast Service Centre), which is a source for MBMS data. The service has been designed so that it can be provided independently of the Radio Access Technology used (UTRAN or GERAN). In this paper we will focus on the provision of MBMS services over the *A/Gb mode* of the GERAN (GSM/EDGE Radio Access Network).
In [4], requirements and recommendations on possible MBMS-specific extensions to the Radio Access Network (RAN) are given. Among other requirements, MBMS solutions to be adopted should minimize the impact on the RAN physical layer and maximize the reuse of existing protocols and functionalities.
MBMS is considered limited to the Streaming and Background QoS classes, as defined in TS 23.107 [1].
The possibility to provide p-t-m transmission had already been envisaged in Release 97 of GPRS, but the technical solution was seen as too complex at the time, and all the efforts were concentrated on the standardisation of p-t-p services. However, now that GPRS is a mature technology, the time is right for the introduction of p-t-m services in the specifications.

2 Transport of MBMS over the GERAN

For MBMS over the GERAN, the Service Data Units (SDUs) are the LLC frames received over the Gb interface (the interface between the GPRS Core Network and the GERAN). Figure 2 shows how LLC frames are derived from application layer packets (in this example RTP/UDP/IP packets are considered) and their processing at the various layers of the (E)GPRS protocol stack. It is expected that a single IP packet will be contained in one LLC frame.
After optional header compression at the SNDCP and framing at the LLC layer, the resulting LLC frame is split by the RLC layer into equal-sized segments of length *M* bytes. The segment size depends on the coding scheme used at the physical layer (CS-1 to CS-4 for GPRS and MCS-1 to MCS-9 for EGPRS). In order to avoid stuffing in the final segment of an LLC frame, the remaining byte positions are often filled up with data from the next frame. Each segment is then mapped onto the data part of an RLC/MAC block. Finally, a CRC is appended before passing it to the channel encoder. The encoded block of constant size (456 bit for GMSK modulation or 1368 bits for 8-PSK modulation) is then interleaved over four bursts, which are mapped on the

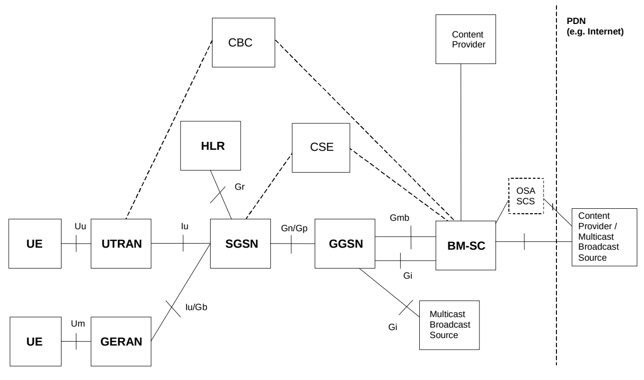

Figure 1: MBMS Architecture [2]

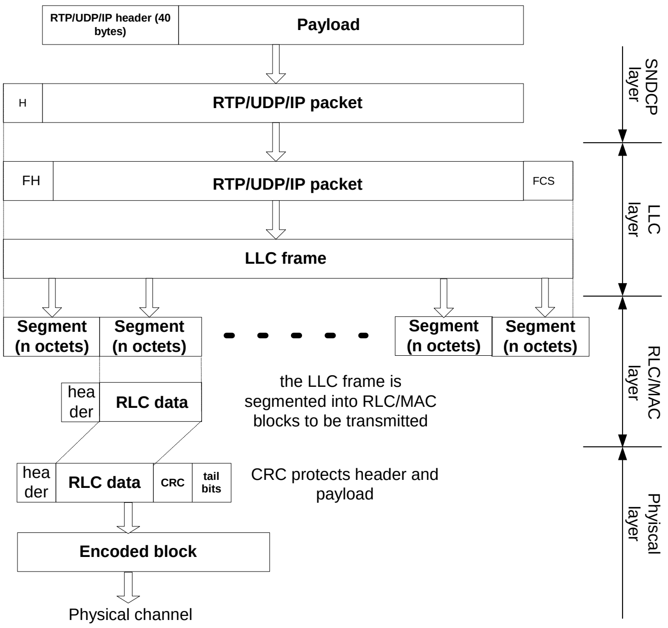

Figure 2 – Segmentation and channel coding for streaming or "download & play"

assigned timeslot in four successive TDMA frames. In case of transmission over multiple timeslots, the RLC segments are distributed over all the allocated timeslots.
At the receiver, channel decoding is performed on each of the received and deinterleaved blocks. If any residual bit errors are detected via the CRC, the RLC/MAC block is discarded. The loss of a block will result in the loss of the SDU that the block is part of, thus resulting in the loss of an IP packet. For typical MBMS services, the SDU error rate should not exceed the target of 10^{-2} or 10^{-3} (depending on the application). The error rate at the application layer is expected not to be higher than the SDU error rate.

3 RLC/MAC procedures for data transfer

For the p-t-m transmission of MBMS data over the GERAN, the following options have been chosen:

1) Without ARQ at the RLC/MAC layer.
2) With ARQ at the RLC/MAC layer.

In the first option, retransmissions of individual RLC/MAC blocks (as in the acknowledged mode of (E)GPRS) is not supported. In the second option, the existing procedures for acknowledged mode, which are valid for data transfer to a single user, need to be modified to cover the case of data transfer to multiple users. For both options, the LLC layer always operates in unacknowledged mode.
In order to use ARQ, feedback is required from the mobile stations listening to MBMS to inform the network of whether RLC/MAC blocks have been received correctly or not. The use of feedback is specific to the GERAN, as in the UTRAN it has been decided that the retransmission of individual RLC blocks will not be supported in Release 6 (see TS 25.346 [3]).
In the following sections, the two options are described in detail.

4 Option 1: transmission without MS feedback

Without feedback, the retransmission of individual RLC/MAC blocks as in the acknowledged mode of (E)GPRS cannot be supported in p-t-m transmission. In this case, transmission is in the downlink only, with no transmission from MSs in the uplink. An initial analysis shows that, even using the most robust coding schemes available (CS-1 or MCS-1), the error rate for the RLC/MAC blocks is such that it would not be possible to achieve the required QoS (e.g. 1% SDU FER) over the whole cell area using the existing procedures for unacknowledged mode. Hence, additional techniques have been considered to improve efficiency and service quality.

4.1 Repetition Redundancy

One possibility to reduce the Block Error Rate (BLER) at the RLC/MAC layer is the method of "repetition redundancy", in which each RLC/MAC block is transmitted a predefined number of times q. A mobile station accumulates correctly received blocks from each transmission to assemble an LLC frame.
Assuming that the errors in the RLC/MAC blocks are independent, and that no soft combining of different received replicas of each block is performed (i.e. each block is decoded independently of the others), the probability of an SDU error P_{SER} is given by the formula:

$$P_{SER} = 1-(1-P_{BLER}^{q})^{L} \qquad (1)$$

where L is the number of RLC/MAC blocks that make up an SDU and P_{BLER} is the Block Error Rate. However, in the receiver there is the possibility to recombine different replicas of the same block, for example using Chase combining. In this case the formula becomes:

$$P_{SER} = 1-\left(1-P_{BLER}^{(q)}\right)^{L} \qquad (2)$$

where $P_{BLER}^{(q)}$ is the probability that a block is received in error after all the q replicas have been combined. This leads to a better performance because $P_{BLER}^{(q)} < \left[P_{BLER}^{(1)}\right]^{q}$.
For EGPRS, it is possible to use *incremental redundancy* (IR), whereby a different puncturing pattern is applied in different retransmissions. As the different retransmissions of the same block are combined in the receiver, the effective coding rate decreases thus reducing the BLER. In the case of IR, Equation (2) still applies.

4.2 Outer Coding

In order to reduce the SDU FER as seen by the higher layers, one alternative that has been investigated is the use of outer coding at the RLC layer. Reed-Solomon (RS) codes have been selected because of their correcting properties and because of their flexibility to adapt to different application requirements as well as different channel conditions.

Reed-Solomon codes are non-binary cyclic codes and are particularly useful in correcting burst errors. The codes are defined by the parameters $(N, K) = (2^m - 1, 2^m - d_{\min})$ where m is the number of bits per symbols; the block size is N symbols, with K systematic information symbols and $N–K$ parity symbols. RS codes are used in many applications because they are *maximum distance separable* (MDS), i.e. they have the largest minimum distance $d_{\min}$ than any other code with the same N and K. The symbol *error* correction probability of the code can be determined as

$$t = \left\lceil \frac{d_{\min} - 1}{2} \right\rceil = \left\lceil \frac{N - K}{2} \right\rceil$$

If symbol *erasures* (i.e. when the location of errors is known) are considered, then a Reed-Solomon code with N-K parity symbols is capable of correcting s erroneous symbols and r erased symbols provided that the following condition is met:

$$2s + r \leq d_{\min} - 1 = N - K$$

When only symbol *errors* are considered (i.e. the location of the errors is not known to the receiver), the performance of m-bit symbol Reed-Solomon code (N,K) for a channel with statistically independent symbol errors is approximately given by:

$$P_E \approx \frac{1}{N} \sum_{j=t+1}^{N} j \binom{N}{j} p^j (1-p)^{N-j}$$

where P_E is the symbol error probability after decoding and p is the symbol error probability before decoding [7]. On the other hand, if only symbol *erasures* are considered (i.e. the receiver knows the location of all errors) up to N-K symbols can be recovered and P_E is given by:

$$P_E \approx \frac{1}{N} \sum_{j=d_{\min}}^{N} j \binom{N}{j} p^j (1-p)^{N-j} \ .$$

In addition to improved performance, erasure decoding has the benefit of lower complexity compared to error-erasure decoding.

Reed-Solomon codes with different rates can be obtained from a single mother code by *shortening* (i.e. by inserting dummy symbols before the encoding process) or *puncturing* (i.e. by deleting symbols after the encoding process). Typically, the nomenclature for the mother code parameters is (N,K), whilst for the shortened or punctured code is (n,k). The mother code can be implemented in hardware, thus speeding up the encoding and decoding processes.

One possibility for introducing outer coding in the GERAN is to apply RS coding to a sequence of RLC/MAC blocks, column-wise as shown in Figure 3. Each symbol of the code consists of 8 bits, i.e. m = 8. Thus, for MCS-1, 22 RS encoding operations are applied to the sequence of k RLC/MAC blocks. The outer coding generates $n–k$ parity blocks from k systematic blocks, which are transmitted separately and reduce the throughput by a factor of k/n, where

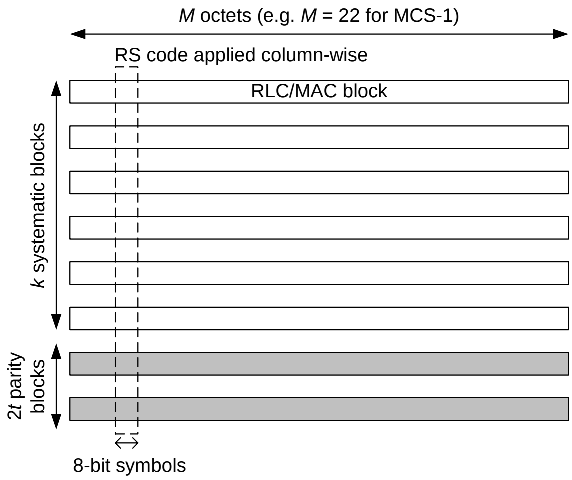

Figure 3: Outer coding applied at the RLC layer

n is the length of the code block. At the receiver, the RLC checks whether each block is in error using the CRC information provided by Layer 1. If a block is found in error, it is discarded, and in each of the 22 RS codewords the symbol corresponding to that RLC/MAC block is marked as an erasure. Reed-Solomon erasure decoding is then performed.

A similar scheme for introducing outer coding in the GERAN is presented in [6]. It is worth noting that TSG GERAN has agreed not to support outer coding for MBMS in Release 6, but to consider it as a possible enhancement for future releases.

4.3 Simulation Results

Figure 4 and Figure 5 show the performance of repetition redundancy and outer coding for GPRS and EGPRS (GMSK), respectively, for a target SDU FER of 1%. For repetition redundancy the performance with a different number of repetitions is presented. For GPRS, results are presented with and without Chase combining; for EGPRS, incremental redundancy without feedback has been used. In general, the use of incremental redundancy leads to better performance than Chase combining. For outer coding the code parameters (n,k) are adjusted through puncturing or shortening to trade throughput against C/I (carrier to interference ratio), while maintaining a fixed SDU FER of 1%. Each point on the curves represents a particular code RS(n,k). The results presented here are for SDUs of 510 octets (500 octets for the IP packets, including headers, plus 10 octets for the SNDCP and LLC headers [9]). The simulations have been performed using the TU3 radio channel profile with ideal Frequency Hopping.

For an SDU FER of 1%, Chase combining results in a gain of approximately 4 dB with respect to the case of no combining. Through the use of outer coding, C/I gains in the region of 6 dB are achievable for GPRS compared to simple repetition redundancy; for EGPRS (with GMSK modulation), gains between 4.5 and 7.5 dB are observed. Further results for repetition redundancy and outer coding with GPRS, EGPRS (GMSK) and EGPRS (8PSK) can be found in [8]. With 8PSK

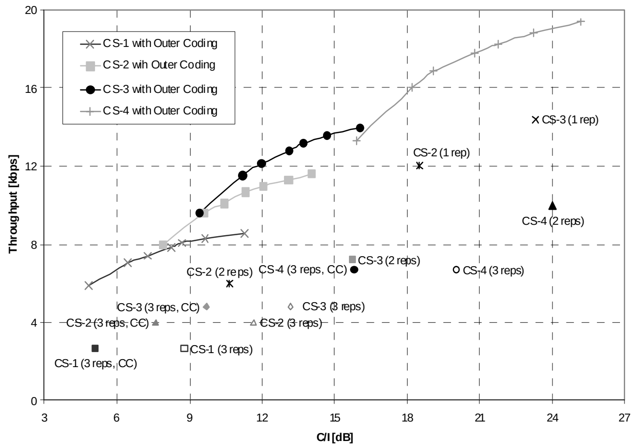

Figure 4: Performance of repetition redundancy and outer coding using GPRS coding schemes, CS-1 to CS-4 for 1% SDU FER (rep = repetition, CC = Chase combining)

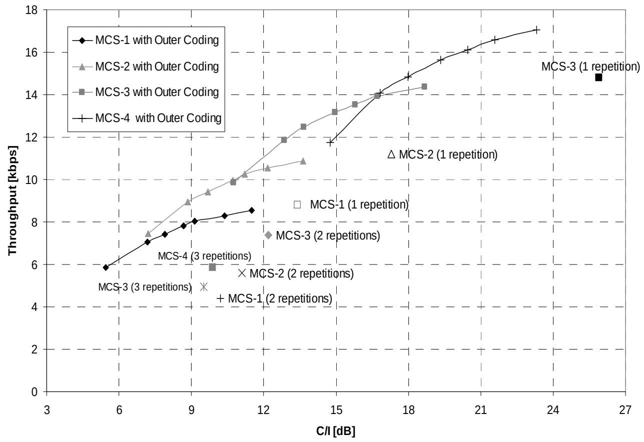

Figure 5: Performance of repetition redundancy (with IR) and outer coding using EGPRS (GMSK) coding schemes, MCS-1 to MCS-4 for 1% SDU FER

modulation the gains are not as great due to the larger payload sizes, leading to improved performance with incremental redundancy. Typically with EGPRS (8PSK) the performance of outer coding varies between no gain at all to gains of 1-2 dB depending on the coding scheme used.

5 Option 2: transmission with MS feedback

In this case, a selective retransmission technique is used (similar to the one used for the acknowledged mode of GPRS). Mobile stations provide feedback to the network (by means of the PACKET DOWNLINK ACK/NACK message) to inform of whether each RLC/MAC block has been received correctly or not. Blocks that have been indicated by at least one mobile station as not having been received correctly may be retransmitted. Feedback is meant to enhance the p-t-m delivery; however, the goal is not to realize a fully acknowledged protocol: even if one (or more) MS indicates that some RLC/MAC blocks have not been received, their retransmission can be skipped by the network. In particular, no persistent transmission is assumed, and each block can be transmitted only a limited number of times. This prevents a single MS experiencing bad radio conditions from adversely affecting the throughput of the overall p-t-m transmission. Further details on the procedure can be found in [5].

It is expected that the number of retransmissions (and therefore the throughput) of a feedback-based solution depends on both the number of terminals receiving the service and their radio channel quality. Simulations have been performed considering a different number of MSs and different C/I conditions (for simplicity, in the simulations the C/I ratio is assumed to be the same for all mobile stations). In all the simulations the assumed timeslot configuration is 4 DL + 1 UL (to allocate the feedback channel). Results are given only for MCS-6 (the coding scheme is kept fixed during the data transfer), and the maximum number of retransmissions per RLC/MAC block is set to 3. The acknowledgment window size is set to 512.

Simulation results in terms of throughput per timeslot are given in Figure 6, while the corresponding SDU error rates are presented in Figure 7. Again, SDUs of 510 bytes are considered, and the channel model is TU3iFH.

The achieved throughput is not constant (as in the case of transmission without feedback) but depends on the number of users and their radio conditions, converging (in the considered scenario) to a minimum value of 9.5 kbps in the

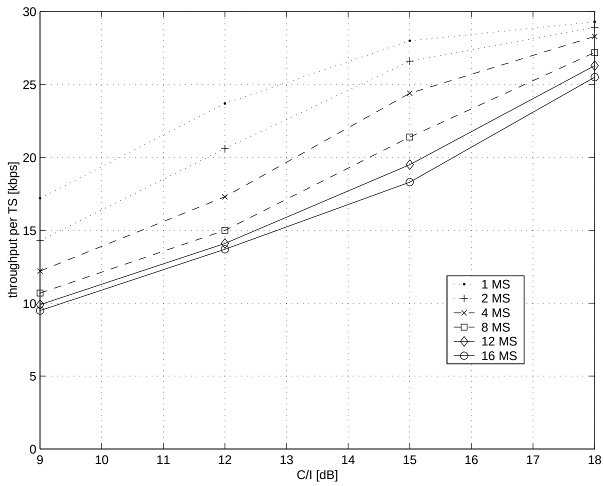

Figure 6: Throughput per TS (MCS-6, max 3 retransmissions)

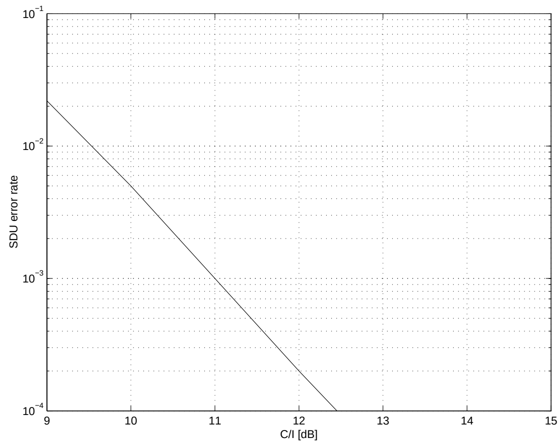

Figure 7: SDU error rate (MCS-6, max 3 retransmissions)

case of 16 users, all experiencing a C/I of 9 dB; but as the radio link quality improves and/or the number of users decreases, the throughput increases.

A scenario with several MSs in fairly good radio conditions (15 MSs with a C/I of 15 dB) and a single MS in moderate radio conditions (1 MS with a C/I of 9 dB) has also been simulated. The throughput is 17.2 kbps per timeslot, compared to 18.3 kbps of 16 MSs with a C/I of 15 dB. Although this confirms that the performance is adversely affected by terminals experiencing bad radio conditions, the throughput is still higher than the one obtained for C/I = 9 dB with a downlink-only solution (i.e. with no feedback).

The procedure being specified in TSG GERAN currently allows the possibility to support only up to 16 users (users need to be addressed in order to be polled for feedback, and only 16 identifiers are available). As the number of users increases beyond this number, it is expected that a better option is to switch to a downlink-only solution.

6 Data rates

The values for the throughput provided in the previous sections refer to the case of a single timeslot. The data rates for an MBMS session can be increased by transmitting over multiple timeslots. If feedback is not used, the maximum number of timeslot that a single MBMS session can be transmitted upon is 6; if feedback is used, the maximum number of timeslot in the downlink is 4, due to the need to transmit in the uplink on one timeslot [5]. Table 1 shows typical data rates achievable with the techniques described in this paper for a C/I of 12dB with multiple timeslots.

	4 TS (kbit/s)	6 TS (kbit/s)
Repetition (MCS-3, 2 reps)	30.0	45.0
Outer coding (MCS-3)	44.8	67.2
Feedback (MCS-6, 4 MSs)	69.2	–
Feedback (MCS-6, 16 MS)	54.8	–

Table 1: Typical data rates for C/I=12dB with 4 and 6 timeslots.

The investigations have revealed that when the number of users receiving MBMS is low, p-t-m with feedback provides the best performance. However, as the number of users increases the best option is to switch to data transfer with no feedback, with outer coding providing better performance than repetition redundancy.

7 Further considerations

The decision as to which of the options for data transfer is selected for a particular service is made by the network. The network starts the establishment of an MBMS session by sending a notification message, which is a broadcast message sent to all MBMS users in the cell informing them that a session is about to start. The notification message may optionally initiate a counting procedure to determine how many users interested in the service are present in each cell; once the number is known, the network can decide whether to establish a p-t-m connection and, in that case, whether to use data transfer with or without feedback [5]. However, if a high number of users is expected to be present in a cell, the counting procedure can be skipped, and the network may directly use downlink-only transmission.

In the case of p-t-m transmission, some mobile stations (e.g. those at the cell boundary) may experience worse radio quality than others, resulting in more corrupted radio frames, even when feedback is enabled, and hence in more corrupted packets that cannot be used by the application. In order to make the p-t-m transmission more robust, two solutions have been devised. The first is an optional FEC scheme at the application layer. Such a scheme could also be useful to recover from longer interruptions due, for example, to cell change of the MS or to a paging procedure for a circuit switched call. The FEC schemes currently under consideration in 3GPP TSG SA WG4 include schemes based on RS codes, LDPC (Low Density Parity Check) codes and Raptor codes [9].

If the FEC at the application layer cannot recover the data, or in case it is not activated, after the end of the MBMS p-t-m session the application layer in the MS may initiate a “file repair” session: a GPRS connection is established between the MS and the BM-SC, so that the MS can request the missing packets, which are then sent over a normal GPRS TBF. This further reduces or eliminates the residual error rate. It has to be noted that this option is only suitable for “download & play” services, but not for real-time streaming applications.

Acknowledgements

This work has been performed as part of the Siemens GERAN standardization project. The authors wish to thank their colleagues within Siemens and in TSG GERAN for many fruitful discussions on the topic.

References

[1] 3GPP TS 23.107, “Quality of Service (QoS) concept and architecture”

[2] 3GPP TS 23.246, “Multimedia Broadcast / Multicast Service (MBMS); Architecture and functional description”

[3] 3GPP TS 25.346, “Introduction of the Multimedia Broadcast Multicast Service (MBMS) in the Radio Access Network (RAN); Stage 2”

[4] 3GPP TR 25.992, “Multimedia Broadcast / Multicast Service (MBMS); UTRAN/GERAN requirements”.

[5] 3GPP TS 43.246, “Multimedia Broadcast / Multicast Service (MBMS) in the GERAN; Stage 2”

[6] H. Jenkac, G. Liebl, T. Stockhammer, W. Xu, “Flexible Outer Reed-Solomon Coding on RLC Layer for MBMS over GERAN”, *IEEE Vehicular Technology Conference (VTC Spring)*, Milano, Italy, May 2004

[7] B. Sklar, *Digital Communications - Fundamentals and Applications*, Prentice Hall, 2001

[8] Tdoc GP-040164, “Forward Error Correction in the GERAN for MBMS”, 3GPP GERAN#18, Reykjavik (Iceland), 2-6 February 2004

[9] Tdoc S4-040348, “Permanent Document on: Simulation guidelines for the evaluation of FEC methods for MBMS download and streaming services, v1.2”, 3GPP SA4#31, Montreal (Canada), 17-21 May 2004