Hybrid type-II ARQ/AMS and Scheduling using Channel Prediction for Downlink Packet Transmission on Fading Channels

Sorour Falahati and Arne Svensson
Communication Systems Group
Department of Signals and Systems
Chalmers University of Technology
SE-412 96 Göteborg, Sweden
{sorour, Arne.Svensson}@s2.chalmers.se

Nilo Casimiro Ericsson and Anders Ahlén
Signals and Systems
Department of Materials Science
Uppsala University, PO Box 528
SE-751 20 Uppsala, Sweden
{nce, Anders.Ahlen}@signal.uu.se

Abstract

In this paper, we propose a system suitable for Internet Protocols (IP) which consists of a hybrid type-II Automatic Repeat reQuest (ARQ) scheme combined with an Adaptive Modulation System (AMS) and a time-slot scheduler supplied by channel predictions, referred to as predictive HARQ-II/AMS. The performance of the system in a downlink for fast fading channels is investigated through simulations where in particular, BER, throughput and delay performance are studied. The proposed system is compared to two modified systems where one does not utilize Forward Error Correction/ARQ (FEC/ARQ) at link layer (referred to as predictive AMS) while the other one has no access to the channel prediction (referred to as blind HARQ-II/AMS). The results show that with perfect channel prediction both predictive AMS and HARQ-II/AMS satisfy the Quality of Service (QoS) requirements with some advantages to the former in terms of efficient usage of channel capacity. However, when imperfect channel prediction is utilized only our proposed system meets the users demands.

1 Introduction

Packet based traffic over wireless links, using IP is a major concern for future communications. In general, TCP/IP is designed for a highly reliable transmission medium in wired networks where packet losses are interpreted as congestion in the network which rarely occurs. However, this design assumption does not hold for wireless networks where the time-varying channel experiences severe fading. Therefore, TCP misinterprets the packet losses over the wireless radio link as congestion which leads to inefficient utilization of the available radio link capacity and therefore degradation of the system performance [1–3].

Exploiting the facilities provided at lower layers of wireless networks such as hybrid FEC/ARQ and adaptive modulation can improve the communication reliability. In this way, the major part of error recovery is performed at lower layers without being noticed at higher layers. Moreover, the performance can be further improved by utilizing channel predictions and a smart time-slot scheduler which uses the spectrum efficiently while satisfying the required QoS for different users [4–7].

In this work we propose a HARQ-II/AMS combined with a time-slot scheduler supplied by channel predictions. Basically, we assume the channel quality for each radio link can be predicted for a time interval of about *10 ms* into future [8, 9], and that they are accessible by the link layer. Based on the predicted values, the HARQ-II/AMS preliminary selects a Modulation and Coding Scheme (MCS) for each user which satisfies the QoS and provides high throughput. The scheduler uses the information about the individual data streams, along with the predicted values of the different radio links and proposed MCSs by the link layer to distribute the time-slots among users. The planning is performed so that the desired QoS and priority associated to different users are guaranteed while the channel spectrum is efficiently utilized.

In Section 2 the proposed system is described in detail. The simulation assumptions are explained in Section 3 where the results are presented and discussed. Finally, some conclusions are drawn in Section 4.

2 System description

To evaluate our approach we have chosen to mix several types of traffic. They all constitute sessions that exist for a predefined amount of time. Each session can have either deterministic, fixed values for packet sizes and packet inter-arrivals, or, they can be partly or completely drawn from some random distribution. For our purposes we have defined three traffic classes: *Voice*, *Data*, and *Media*. The traffic is generated using a Poisson distribution for the packet inter-arrival time and a Pareto distributed packet size, except for *Voice*, which is chosen to have a fixed packet size [10]. In the following, three main divisions of the system referred to as buffer, scheduler and Hybrid

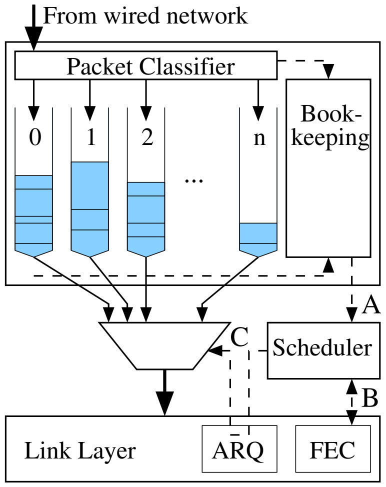

Figure 1: Schematic of the buffer and its queues, and how they interconnect to the scheduler and link layer.

type-II ARQ/AMS, are described in Sections 2.1, 2.2 and 2.3, respectively.

2.1 Buffer

At the incoming side from the wired network we maintain a buffer with separate queues for the different traffic flows, also distinguishing the packets with respect to the destination, so that each source-destination pair has a separate incoming queue. All incoming packets are scanned for size, priority, session identification number, and required service level from the link layer.

This information is stored in the buffer and can be accessed by the scheduler. The buffer controller is assumed to be able to submit a status report to the scheduling subsystem, described in Section 2.2, so that the scheduler can make an appropriate decision on which queues to choose for the next transmission frame.

The queues are emptied in a bit-by-bit manner, independently of the individual packet boundaries. The bitstream is passed to the link layer, along with information about the service requirements. The incoming buffer is described in Figure 1. At the receiving side of the wireless link, the packets have to be re-assembled, before passing them up to the network layer. This can be done since the scheduling decision is transmitted (broadcasted) to the receiving side, and it totally determines which byte belongs to which flow.

2.2 Scheduler

The system we propose makes use of a channel predictor and a multi-user time-slot scheduler that are mounted on top of a Hybrid type-II ARQ/AMS scheme. The scheduler creates a signaling pipe [3] between the network layer buffer and the link layer service, making them mutually aware of one-another. For instance, the network layer does not ask for a link service whenever there is data to transmit. Instead it notifies the scheduler of the incoming traffic by passing queueing information (A in Figure 1) about the amount of data and type of service that would be preferred by the packets. The scheduler then asks the link layer for a report (B in Figure 1) about how the channel conditions would meet the required service. This can be done since the link layer has access to channel prediction data of all the established connections.

The link layer builds up an $M \times N$ matrix of channel quality predictions, where M is the number of time-slots, and N is the number of traffic streams with ongoing sessions. These predictions cover the next following time-frame of $5ms$. The predictor has a prediction horizon of $10ms$. Figure 2 shows the predicted values for one link. Based on the predicted values of the Signal to Noise Interference Ratio (SNIR) for each of the N user's channels, and the target error rates , the initial code rate and signaling constellation are chosen in advance by the link layer, for a set of $M = 48$ future time-slots. It is then the task for the scheduler to, along with the queue sizes and priorities, distribute these time slots among the different queues in a fashion that maximizes some criterion, such as throughput or another measure of user satisfaction [10].

The scheduler performs the scheduling in two rounds. In the first round, each time slot is simply allocated to the user that can transmit at the highest rate in that time slot. If the buffered data in the queues had infinite size, this approach would actually maximize system throughput. However, since the buffers are not infinite, the scheduler runs a second round, where time-slots are redistributed from users that have been over-supplied (rich), to users that have been under-supplied (poor). We call this equalization to user satisfaction the *Robin Hood* principle: *"to take from the rich, and give to the poor"*.

2.3 Hybrid type-II ARQ/AMS

Even though the intention is to achieve reliable communication through suitable allocation of time-slots by the scheduler as well as appropriate selection of MCS by the physical layer, the possibility of erroneous reception can not be excluded. Therefore, a selective repeat ARQ protocol at link layer cooperated with an adaptive MCSs at the Physical layer, referred to as Hybrid type-II ARQ/AMS (HARQ-II/AMS), can perform partial data packets error recovery through a limited number of retransmissions. Further error recovery and complete end-to-end reliability is accomplished by the error and flow control mechanism defined at TCP which is not within the scope of this work.

HARQ-II/AMS is based on the Rate Compatible Convolutional (RCC) codes at parent rate 1/3 with constraint length 7 where the higher and lower rate codes are ob-

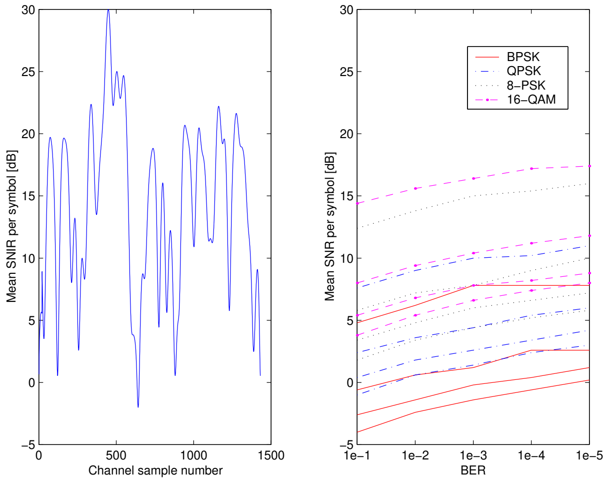

Figure 2: SNIR trend and the bit error probability for different combination of modulation and coding rates. For a fixed modulation, the BER corresponding to the different codes is decreasing by decreasing the coding rate.

tained by optimum puncturing or repetition of the parent rate code according to a puncturing or repetition pattern with period 2, respectively [11]. In this scheme, the transmission of a lower rate code in response to the retransmission request, is performed by sending only the *Incremental Redundancy* (IR) bits due to the rate compatibility property of the RCC codes. Moreover, due to the AMS, the transmitter chooses a modulation scheme among the alternatives 16-QAM, 8-PSK, QPSK and BPSK.

Selection of MCS is based on the user target BER and the predicted SNIR value accessible by the link layer. For each MCS, the minimum required SNIR values (referred to as the *SNIR limits*) for satisfying different BER requirements are specified, accordingly. For a given target BER, the corresponding *SNIR limits* are compared with the predicted SNIR, to select a MSC which provides the maximum throughput amongst the others.

However, due to the difficulties in analytical evaluation of the *SNIR limits* because of the large signalling constellation and convolutional coding, a numerical approach is preferred [11]. The simulated BER is therefore evaluated for all the combinations of 16-QAM, 8-PSK, QPSK and BPSK modulations and the code rates 1, 2/3, 1/2, 2/5, 1/3, 1/4, 2/9, 1/5, 2/11 and 1/6 for an AWGN channel which is the presumed channel model within a time-slot. Some of the simulated *SNIR limits* are illustrated in Figure 2 along with a realization of SNIR of a channel.

The selection of MCSs is performed in two phases as described in the following.

Phase one happens at point B in Figure 1, when the scheduler demands a report from the link layer. This report contains the channel prediction values for all the users with ongoing sessions with corresponding selected MCSs and the temporary priorities (i.e. the users with retransmission request).

Phase two occurs at point C in Figure 1, when the scheduler informs the link layer about the decision for the time-slot allocation, by reporting which users have been allocated time-slots and which time-slots.

At *Phase two* where the final decision for MCS is performed, two simplifying design assumptions are used. First, a *fixed modulation* is used within a time-slot while variable coding rate is allowed for different packets. The second one is the constraint on *rate compatibility* for retransmitting packets. For each user, the highest priority is assigned to retransmission requests (if they exist), following the policy of *"first in, first out"*. These packets are dynamically assigned to the time-slots, meaning that among the time-slots which can propose an appropriate MCS, the one offering the maximum throughput is selected. In this fashion, the retransmitting packets are occupying different time slots with corresponding MCSs. In case of failure, the transmission of the erroneous packet is postponed to the next frame.

After this stage, based on the partial or completely emptiness of the time-slots for each user and the appropriate MCS, the link layer drains *new data* from the corresponding buffers (the arrow at C in Figure 1). Hence, new packets are formed which fill the corresponding time-slots.

After completing this procedure for all the users, the frame is constructed by Cyclic Redundancy Check (CRC) and Convolutional encoding for error detection and correction, respectively, and modulation at the Physical layer. More details about the frame structure is given in Section 2.4. At each mobile host, the receiver performs optimum soft decoding by a Viterbi decoder at the parent rate which is aided by Channel State Information (CSI). The decoded bits are fed into the CRC decoder. In case of error detection, a NACK signal is fed back to the link layer transmitter if retransmission is permitted, and otherwise, an ACK is fed back.

Finally, the channel during frame transmission is assumed to be a time-varying fading channel. However, the fading is assumed slow enough for the SNR to be considered constant within each time-slot. We use one-tap channels in our simulations. This corresponds to assuming an AWGN channel in a time-slot.

2.4 Frame structure

Each time-slot comprises of user data, and a mid-amble of training data to aid the channel estimation process. Before and after the user data there are tails serving as guard intervals against bad timing in the reception or transmission. A descriptive image of the time-slot format is given in Figure 3. The figure also shows the location of the scheduling information, within the frame of 49 time-slots.

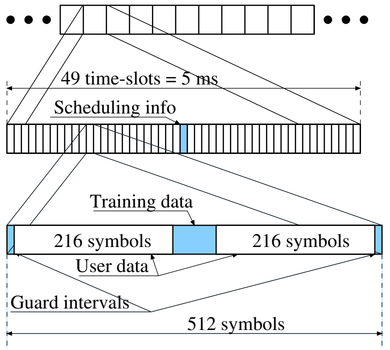

Figure 3: Frame structure

A frame consists of 49 time-slots, each of which can be dynamically assigned to one user, except for one of the time-slots in each frame, which is dedicated to broadcast scheduling information about the next frame. Furthermore, the assumption of a symbol rate at 5 Msymbol/sec which is used in this work, allows each time-slot to contain 512 symbols. Please note that a time-slot may consist of several packets. A compromise between the base-station and the mobile terminal leads to the suggestion that the scheduling information is transmitted in one of the time-slots in the middle of the frame. In this way the scheduler has time to perform the necessary calculations, and the mobile stations have time to adjust to the new schedule.

Since the scheduling information that is transmitted in the downlink is crucial for the efficiency of the multiple access scheme, it has to be well protected against errors. For this time-slot, only BPSK modulation with a low rate code should be used.

3 Simulation results

The following assumption are used during the simulation. The data packets at link layer contain 216 bits, including 12 CRC bits for error detection and 6 zero tail bits corresponding to the memory of the convolutional encoder. 16-QAM, 8-PSK, QPSK and BPSK modulations are employed where Gray coding is used for mapping the bits to the symbols. The maximum symbol energy and symbol rate are presumed to be constant in all the modulation schemes. The channel SNIR prediction is assumed to be lognormally distributed with a mean value equal to the true SNIR and a standard deviation of σ dB. The CSI used at the receiver is considered to be equal to the predicted values. Moreover, the channel is AWGN within a time-slot but fading during a frame transmission. For *Voice*, *Media* and *Data* traffic classes, target BERs of 10^{-3}, 10^{-4} and 10^{-5} are presumed, respectively. Additionally, the maximum allowed number of retransmissions at the link layer is chosen to be 3, 3, and 8 for *Voice*, *Media* and *Data* traffic classes, respectively. Finally, each simulation is carried out for 0.146 sec.

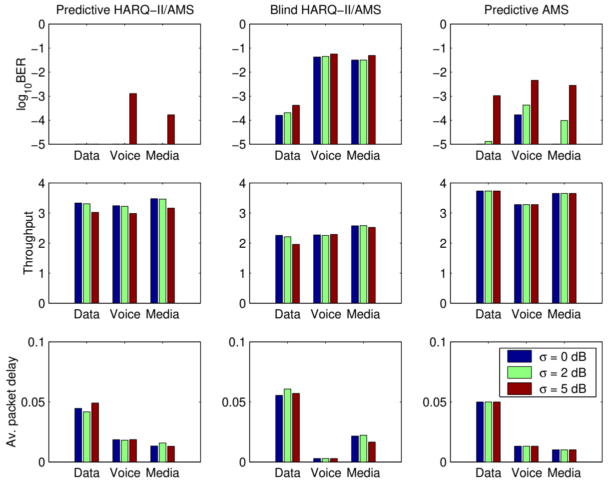

Figure 4: The BER, throughput (bits/symbol) and average packet delay for the *predictive HARQ-II/AMS*, *predictive AMS* and *blind HARQ-II/AMS* for σ, the standard deviation of the prediction error, equal to 0, 2 and 5 dB. The users belong to the classes *Voice*, *Data*, and *Media*.

For the performance assessment of the proposed system which is referred to as the *predictive HARQ-II/AMS* and in order to investigate the effect of coding, ARQ and channel prediction at the transmitter, we have considered two other systems for comparison.

One is the so-called the *predictive AMS* where no coding and retransmission protocols are allowed at lower layers but the link layer is provided with the channel predictions and an adaptive modulation system, similar to one presented in [10]. In this system 16-QAM, 8-PSK, QPSK and BPSK modulations are used where the corresponding *SNIR limits* are analytically evaluated as described in [12]. The *predictive AMS* reports the channel predictions and suggested modulation schemes to the scheduler for time-slot allocation. Furthermore, since no coding is involved here, the data packets at link layer generally contain 6 extra data bits due to the absence of tail bits, compared to the *predictive HARQ-II/AMS*.

The other one is the so-called *blind* or *non-predictive HARQ-II/AMS* where the link layer does not have access to any channel predictions and selection of the MCS is based on the predetermined values stored in a look-up table, similar to the one presented in [11]. In this system, the transmission starts using a modulation with large signalling constellation and a high rate code. During the retransmission, the size of constellation as well as the coding rate are reduced. The transmission is initialized with 16-QAM at rate 1. The next transmission attempts are performed by 16-QAM at rate $2/3$, followed by 8-PSK at rates $1/2$ and $2/5$, QPSK at rate $1/3$, $1/4$ and $2/9$ and finally BPSK at lower rates. Moreover, only the information according to temporary priorities due to the users with retransmission request is provided for the scheduler.

The simulations are carried out for 15 users with 5 users belonging to each of the traffic classes of *Voice*, *Data*, and *Media*. The BER, throughput and delay performance of three systems is evaluated both for perfect ($\sigma = 0$ dB) and imperfect channel predictions ($\sigma = 2$ and 5 dB). Here, the throughput is defined as the number of correctly received data bits per transmitted symbol[1]. The delay is given by the average packet delay which represents the average of all completely removed packets from the queue. The results are averaged for any of the *Voice*, *Data*, and *Media* classes and depicted in Figure 4.

The results show that with perfect channel predictions, both the *predictive HARQ-II/AMS* and *predictive AMS* meet the QoS requirements in contrast to the *blind HARQ-II/AMS*. This stems from the fact that having access to the information about the future characteristics of the channel results in a more appropriate choice of modulation and (or) coding rate. However, comparing the throughput and delay performance leads to the conclusion that on average, *predictive AMS* performs well enough and *predictive HARQ-II/AMS* is too conservative. The reason is that with ideal channel prediction, the required error rate can be obtained without ARQ and channel coding. The importance of the ARQ protocol and channel coding appears when erroneous channel predictions are introduced. In this case, the system robustness is increased by FEC/ARQ. It is shown that for $\sigma = 5$ dB only the proposed system guarantees the required QoS even if slightly higher throughput or lower delay are achieved by the others.

4 Conclusion

A combination of radio channel predictive time slot scheduling and HARQ-II/AMS, for IP (Internet Protocol) packet data, is presented and evaluated through simulations. The performance of the proposed system referred to as the *predictive HARQ-II/AMS* is evaluated and comparison is made with a system where no channel state information is available at the transmitter (the *blind HARQ-II/AMS*) or a retransmission protocol and channel coding are not provided for the system (the *predictive AMS*). Three traffic classes referred to as *Voice*, *Data*, and *Media* are chosen where each requires different BER.

The results show that our proposed system reduces the system delay due to the channel prediction which is available at the transmitter. Additionally, applying channel coding improves the error correction capability of the system and provides robustness against channel prediction errors, especially in situations where there are considerable changes in the channel conditions during the prediction time interval. Moreover, the error correction is done in an adaptive way, suitable for the time varying channel. Furthermore, employing the adaptive modulation system enhances the throughput by utilizing the channel capacity more efficiently. Therefore, the proposed system can provide a high throughput while keeping the BER and number of transmissions low and hence, guaranteeing the required QoS. The present system is designed to provide a constant QoS and throughput over time-horizons characterized by short time fading. The performance and robustness in the presence of slow fading and SNR variations on a longer time horizons remains to be studied. Added robustness will here be provided by suitably design transport protocols [13].

[1] Please note that in *predictive* or *blind HARQ-II/AMS*, only the packets without any detectable errors are regarded as correct ones. In *predictive AMS*, all the received packets are considered to be correct since no error detection codes are utilized.

References

[1] Hari Balakrishnan, Srinivasan Seshan, Elan Amir, and Randy H. Katz, "Improving TCP/IP Performance over Wireless Networks," in *Proc. MOBICOM*, Berkeley, CA, 1995, pp. 2–11.

[2] Andrew S. Tanenbaum, *Computer Networks*, Prentice Hall International, Upper Saddle River, New Jersey, 1996.

[3] Gang Wu, Yong Bai, Jie Lai, and Andy Ogielski, "Interactions between TCP and RLP in Wireless Internet," in *GLOBECOM*, Rio de Janeiro, Brazil, December 1999, pp. 661–666.

[4] C. Roobol, P. Beming, J. Lundsjö, and M. Johansson, "A Proposal for an RLC/MAC Protocol for Wideband CDMA Capable of Handling Real time and non Real Services," in *Proc. IEEE Vehicular Technology Conference*, May 1998, pp. 107–111.

[5] M. Kawagishi, S. Sampei, and N. Morinaga, "A Novel Reservation TDMA Based Multiple Access Scheme using Adaptive Modulation for Multimedia Wireless," in *Proc. IEEE Vehicular Technology Conference*, May 1998, pp. 112–116.

[6] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, "Packet Reservation Multiple Access for Local Wireless Communications," *IEEE Transactions on Communications*, vol. 37, no. 8, pp. 885–890, Aug. 1989.

[7] M. Naijoh, S. Sampei, N. Morinaga, and Y. Kamio, "ARQ Schemes with Adaptive Modulation/TDMA/TDD Systems for Wireless Multimedia Communication Services," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Helsinki, Finland, Sept. 1997, pp. 709–713.

[8] T. Ekman, G. Kubin, M. Sternad, and A. Ahlén, "Quadratic and Linear Filters for Mobile Radio Channel Prediction," in *Proc. IEEE Vehicular Technology Conference*, Amsterdam, The Netherland, Sept. 1999, pp. 146–150.

[9] T. Ekman, *Prediction of Mobile Radio Channels*, Licentiate thesis, Department of Signals and Systems, Uppsala University of Technology, Dec. 2000.

[10] N. C. Ericsson, "Adaptive Modulation and Scheduling of IP traffic over Fading Channels," in *Proc. IEEE Vehicular Technology Conference*, Amsterdam, the Netherlands, Sept. 1999, pp. 849–853.

[11] S. Falahati and A. Svensson, "Hybrid type-II ARQ Schemes with Adaptive Modulation System for Wireless Channels," in *Proc. IEEE Vehicular Technology Conference*, Amsterdam, The Netherlands, Sept. 1999, pp. 2691–2695.

[12] J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 3rd edition, 1995.

[13] A. Ewerlid, "Reliable Communication over Wireless Links," in *NRS'01*, Saltsjöbaden, Sweden, Apr. 2001.