Adaptive Modulation and Scheduling of IP traffic over Fading Channels

Nilo Casimiro Ericsson
Signals and Systems, Uppsala University,
PO Box 528, SE-751 20 Uppsala, Sweden
nce@signal.uu.se

ABSTRACT

In future packet based wireless communication systems, transmission in the downlink will often dominate the traffic load. An obstacle in this context is the time-variability of the channel.

To achieve a high throughput also over fading channels, adaptive methods for adjustment of, for example, the modulation alphabet, and the coding complexity, can be used.

In this paper we investigate the effect of adaptive modulation, along with time-slot scheduling of IP-like traffic in a scenario involving several types of mobile hosts and one base station. We extend a study of the impact of adaptive modulation and scheduling on the bit-error rate, to include models for packet length and packet inter-arrival, to find the queueing delay imposed by our proposed scheduling algorithm.

Our scheduler keeps the bit error rate at attractively low, prespecified levels, well suited for Forward Error Correction (FEC) codes. Moreover, the scheduler splits the bandwidth between different types of traffic in a desirable way, according to the traffic situation.

1 INTRODUCTION

Fading channels confront us with the problem of lost packets and the need for frequent retransmissions. One strategy to combat time-variability is to use averaging: Spread-spectrum signalling can average out variations of the noise and interference level, while coding and interleaving can compensate for the temporary loss of signal strength due to fading dips. Such strategies can combat bad signalling conditions, but are inefficient when the channel conditions are good. In this paper we explore a strategy, where the time-variations of the channel, due to short-term fading, are estimated and the signalling scheme is adapted accordingly. This does not exclude that adaptation to slower variations can take place simultaneously, such as slow power control to compensate for long-term fading and shadow fading, though, in this paper, we focus at the fast variations. We can exploit temporarily good transmission conditions to obtain higher throughput, while reducing the demands on the channel when its condition is bad. Assuming a system making use of either Frequency Division Duplex (FDD) or Time Division Duplex (TDD) with separate (ideal) control channels, the current channel parameters can be estimated and predictions about their future evolutions can be stored for subsequent transmission in the control channel. The bit-rate can be tailored to the current channel conditions by, for example, adjusting the modulation complexity, while keeping the transmitted symbol energy at a constant level. The further into the future the terminal can perform accurate predictions of the channel parameters, the more flexible and efficient the selection of the symbol alphabet will be. Moreover, the traffic on the control channel can be efficiently planned to minimize the signalling overhead.

For a predicted value of the Signal to Noise and Interference Ratio (SNIR) of each channel, the modulation level is maximized under the constraint of a required probability of symbol error, for example, $P_M \leq 10^{-5}$. If no modulation level attains the required probability of symbol error, then transmission is deferred until later when the SNIR is higher. The reason for using this strategy is that it will stabilize the error probability, thus keeping the error rate at a low and constant level, avoiding retransmissions. The averaging strategies mentioned above do not have this feature. On the contrary, they would yield a higher traffic load when conditions are bad, since the increasing error rate would increase the requests for retransmissions. In the case discussed here, when many mobile terminals are connected to the same base station, sharing the same frequency, the strategy will be to allocate the channel to the mobile that can make the best use of it.

We assume that accurate long-term predictions of the rapidly varying channel SNIR are available, allowing us to *schedule* the transmissions for one or more users. A non-linear method for achieving channel predictions is described by Ekman and Kubin in [3]. Channel predictions are demonstrated to be accurate for horizons farther than $10ms$ ahead in time. It will be possible to allocate resources also beyond the next fading dip.

Similar approaches to adaptive modulation, which compensate for fast fading, have been proposed by Ue, Sampei and Morinaga [1], and, Chua and Goldsmith [2]. However, they use an *instant estimate* of the channel based on the received power instead of *predicting* it, and, they *estimate* the received symbol alphabet instead of *scheduling* the transmission.

Our proposed scheme will result in some overhead due to the transmission of scheduling decisions over separate control channels. It is crucial to the performance of this system that the control information is correctly transmitted. This was also indicated by Torrance and Hanzo in

[4].

A nice feature of our approach is that the adaptive modulation strategy is embedded in the schedulation operation. The decision of the modulation format is merely a first natural step in the scheduling process, where we decide which throughput different users can achieve in each time-slot. The following steps use this information to make a fair resource distribution.

2 SYSTEM DESCRIPTION

The era of wireless Internet has just started, and it will be supported by the Wireless Application Protocol (WAP). This protocol ensures that a minimal amount of data is transmitted over the wireless channel to save bandwidth.

In the future, however, in fourth generation wireless (4GW) systems, applications will not distinguish between wired and wireless terminals. If IP is chosen to be the protocol of the future, then it will also have to reach wireless terminals. This doesn't mean that we should just consider the wireless link as any other (wired) connection. Radio bandwidth is still a precious resource, and we need to be careful when using it.

In our approach, we assume that we have a proxy-server that will handle retransmissions due to errors in the backbone network (Internet), before passing the packet to the mobile terminal over the wireless channel. In this way, we will avoid retransmissions over the wireless channel due to errors caused by the wired network.

The key innovations introduced in our system are:

- A **channel estimator/predictor** that gives accurate predictions of the channel quality for a considerable extent of time.
- A **scheduler** which performs efficient allocation, based on the channel quality, the mobile host's desired bandwidth, and, its priority.

For the predictor to work, either fast TDD, or a feedback/control channel is required, since the predictor needs to be fed with accurate measurments of the current channel conditions. For more details, see [5].

3 ADAPTIVE MODULATION

In traditional communications systems, channel variations are dealt with in a worst-case manner. For wireless systems this implies the use of a simple modulation scheme, and a complex error-correcting code. When the coding fails to compensate for temporary bad conditions, higher layers in the protocol will ensure that the information is correctly and completely transmitted, by requiring a retransmission of the erroneous data. We wish to avoid this by adapting our demands on the channel as it varies. By changing the modulation format as the channel SNIR (SIR) varies, we hope to accomplish less retransmissions.

The symbol alphabet is decided in advance, since we assume accurate predictions of the channel quality to be available. The decision is made on a frame-by-frame basis, each of which contains 48 time-slots. Each time-slot corresponds to one output value from the predictor, thus the channel is assumed to be constant during the time-slot.

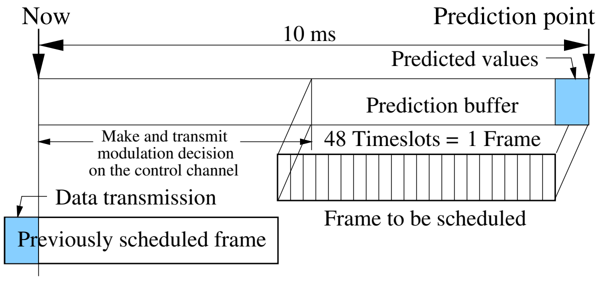

Figure 1: Based on measured channel SNIRs, predictions of future SNIR are made and stored in a buffer until a frame of $5ms$ is filled. The scheduling algorithm then has $5ms$ to make the decisions and transmit them to the other end of the link. Then, based on these decisions, the data transmission starts.

For each time-slot, a burst of 512 symbols is transmitted. Using a prediction horizon of $10ms$ we can collect the predictions during $5ms$, then use the remaining $5ms$ to make the decision and transmit it to the other side of the link, see Figure 1.

3.1 Finding the decision thresholds

For a given symbol error probability we can calculate the required SNIR for the different modulation formats used. Thus, we can decide the thresholds where we should change from one modulation format to another. A tight upper bound on the symbol error probability due to Gaussian noise for M-QAM modulation is given by [6]:

$$P_M \leq 1 - \left[1 - 2Q\left(\sqrt{\frac{3E_{av}}{(M-1)N_0}}\right)\right]^2. \quad (1)$$

Here the average symbol energy E_{av}, the noise power, N_0, and the modulation format, M, are assumed to be known. The Gaussian cumulative distribution function, $Q(x)$, can be calculated according to:

$$Q(x) = \frac{1 - \mathrm{erf}(\frac{x}{\sqrt{2}})}{2}, \quad (2)$$

where $\mathrm{erf}(x)$ is the error function:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \quad (3)$$

By solving (1) for $\frac{E_{av}}{N_0}$ and using (2) and (3), we obtain the SNIR required for a certain symbol error probability, P_M, and a given M:

$$\frac{E_{av}}{N_0} \geq \frac{2(M-1)}{3} \left[\mathrm{erf}^{-1}\left(\sqrt{1-P_M}\right)\right]^2. \quad (4)$$

In this investigation we use 64-QAM as the maximum modulation level, thus transmitting six bits per symbol when the channel is as its best. When the channel degrades, lower powers of two are used with BPSK being the lowest level. In Figure 2, the left hand diagram illustrates the SNIR-variation of a typical channel while the right hand part illustrates how the level of modulation can

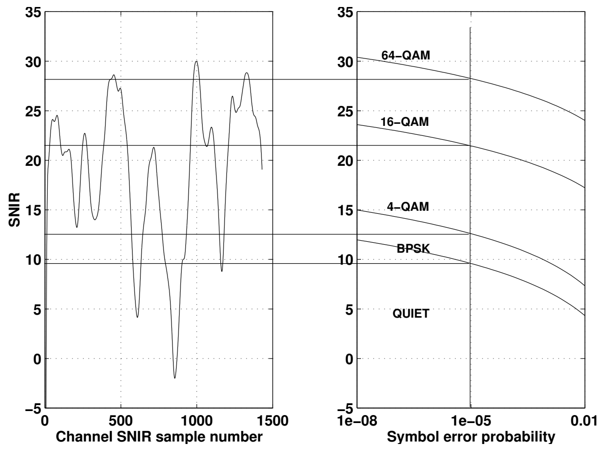

Figure 2: SNIR trend and modulation level related to the error probability.

be selected for a pre-specified symbol error probability. As an example we note that for an SNIR $\geq 22dB$ we can transmit during 150 time-slots (time-slot 390 to time-slot 540) with a modulation level of 16-QAM at a symbol error probability of $P_M \leq 10^{-5}$.

4 SCHEDULING

One way to make use of the channel predictions is to simply choose the modulation format for a user during the nearest future, to satisfy the demand of a low and constant bit error rate. On the other hand, it is a waste of time, and bandwidth, to choose a low modulation format (or to not transmit at all) when the channel condition for a specific link is poor: We could allocate that time-slot to another user, which has better conditions. Moreover, a user may not need all of its allocated bandwidth. Such inefficiencies can be avoided by the use of scheduling. Our scheduler works on a specific physical channel (such as a frequency band), where several links are maintained, and the users have to compete for the transmission time.

In order to efficiently distribute the channel bandwidth between different concurrent users on a TDD/TDMA channel, we make use of the predictor in a way that is a natural extension from the original adaptive modulation approach. The predicted SNIR values are now not only used for the selection of modulation format, but also for time-slot distribution among the users so that system throughput is maximized. One of many possible strategies for transmission scheduling will be investigated here.

4.1 Maximization of system throughput - Pass 1

First, the scheduling procedure allocates each time-slot to the user that can transmit most efficiently in that particular slot. This approach actually maximizes the system throughput (for a given error probability), but it may be a very unfair way of allocating time-slots among different users. Even in this case, a user may be allocated more bandwidth than is required, whereas at the same time, another user may not be allowed to transmit at all. To compensate for this unfairness, a re-distribution of the time-slots takes place.

4.2 Equalize user satisfaction - Pass 2

In most cases there will be users that have received more time-slots than they need, and users that have received less than they require. We call these "rich" and "poor" users, respectively. The re-distribution procedure starts by identifying the richest user, that is, the user with the largest over-allocation of bandwidth. The richest user offers its worst[1] time-slots to the scheduler for distribution among the poor users. The time-slots are given to the users that can use them best. This procedure is repeated as long as there still remains both rich and poor users. By not letting a previously rich user become poor, and vice versa, this second pass is guaranteed not to redistribute each time-slot more than once. The guarantee of limited run-time of each scheduling pass is an important advantage over other methods, such as linear programming.

This second pass will generally reduce the total allocated datarate as compared to pass 1. The reason for this is that the user receiving a time-slot in the first pass will most likely be able to use a higher modulation format than all other users in that time-slot. However, there is no reason to let a user occupy a time-slot when he doesn't really need it, just because he has good transmission quality.

The method outlined above has a number of parameters that need to be selected and adjusted in a particular system.

1. Allowed modulation format (as in Figure 2)
2. User priority (as in Table 1)
3. Predicted SNIR

These features are taken into account in the scheduling process by comparing them among the competing users, one feature at a time. If two users have the same value in the first feature, then the second one is compared to find a winning user. For users with equal modulation format and priority, the third feature to compare should be the measured/predicted SNIR. This will allocate the channel to the user that most probably will generate the fewest errors, all other things being equal.

The *order* in which the scheduler compares the allowed modulation format and priority of each user and time-slot will affect the way in which the scheduler allocates the bandwidth. For example, by comparing *user priority* before *modulation format*, the scheduler will always allocate the channel to the higher priority user as long as that user has something to transmit. This reconfiguration capability can be exploited to adapt the scheduler to different traffic situations: In normal (low-medium traffic) situations it can favor *user priority* higher than *modulation format*, whereas when traffic starts congesting, these two features can change places in order to achieve higher throughput, thus flushing out pending jobs from the queue.

Other methods to optimize the allocation decision can be considered, for example linear programming algorithms [9], and generalizations of existing router-scheduling algorithms to take the time-varying channel quality into account. The drawback of linear programming methods is that the procedure is iterative. Thus no

[1] In the sense of low transmission rate or low transmission quality

upper limit can be given for the number of operations required. Moreover, in the linear programming case, we would need to define an appropriate cost function that is to be minimized in order to optimize the scheduling decision.

5 EXPERIMENTS

To evaluate the proposed system solutions, a simulation series was conducted, assuming one base station transmitting to a number of mobile terminals. We will only consider the downlink here, since we expect it to be the bottleneck in future wireless communication systems. In this idealized scenario, we assume that the channel conditions are predicted without error for 10 milliseconds ahead in time. Moreover, perfect synchronization and transmission at a constant maximum amplitude, regardless of the symbol alphabet, is assumed. The experiment is applicable to both TDD and FDD systems, provided that accurate predictions of the channel conditions exist.

There is no implementation of Forward Error Correcting codes (FEC), nor Automatic Repeat reQuest (ARQ), in this experiment. These features are intended to be built on top of this proposed scheduling system.

For each prediction of the SNIR at the receiver, the modulation alphabet is selected for 512 consecutive symbols (one time-slot) for each user. This implies that the channel estimator and the predictor work at a rate of $\frac{bw}{512}$, where bw is the channel bandwidth. The data bit stream is then modulated and transmitted with a constant maximum amplitude over the noisy channel. White Gaussian noise with varying variance is added to simulate good and bad channel conditions. Thus, we use a one-tap fading channel, where the fading is simulated by varying the noise variance. At the receiver, the signal is demodulated and the obtained bit stream is compared to the original one. The number of errors is counted, as well as the number of transmitted bits.

The incoming traffic is generated using a Poisson distribution for the packet inter-arrival time, and a Pareto distributed packet size, except for the "VOICE" traffic class, which was chosen to have a fixed packet size. The packet size is generated in blocks of 512 bits. Thus the fixed packet size (=1) for the VOICE class means that each VOICE packet is 512 bits long. The Poisson cumulative distribution function is given by (5), and the Pareto cumulative distribution by (6).

$$F(t) = 1 - e^{-\lambda t} \tag{5}$$

$$F(t) = 1 - \left(\frac{k}{t}\right)^{\alpha}, \quad t \geq k \tag{6}$$

The parameters are λ, which is the inverse of the mean inter-arrival time, k is the minimum packet size in the Pareto distribution, and, α is a shape parameter[7, 8]. For $\alpha \leq 1$ the distribution has infinite mean, and for $\alpha \leq 2$ infinite variance.

In Figures 3-4 the outcome of two simulations are plotted. In Figure 3, the top graphs show the bit throughput represented by a vertical bar for each frame, and a different color/nuance of gray for each of 25 users. The bottom diagrams show the resulting bit error rate (BER). Each frame consists of 48 time slots. Each time slot corresponds to one output sample from the channel predictor. The scheduling is optimizing the transmission within each frame.

Class	P(err)	Prio	Inter-arrival	Packet size
VOICE	10^{-2}	5	$\lambda = 0.5$	Fixed, 1
MEDIA	10^{-3}	4	$\lambda = 0.5$	$k = 4, \alpha = 1.03$
DATA	10^{-4}	2	$\lambda = 0.05$	$k = 20, \alpha = 1.03$

Table 1: The different traffic classes used in the simulations and how their parameter values were chosen. A high value on k means large packets. Larger λ generates packets more often. A higher value on the priority means higher priority.

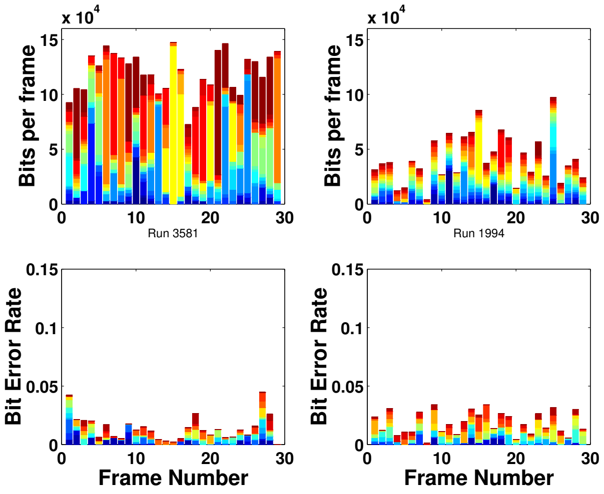

Figure 3: Scheduling and transmission result after running the scheduler on two equal sets of 25 users and their respective channels, but with different settings on the *order of comparison* of the parameters in the scheduling process. To the left, *modulation format* was compared before *user priority*. To the right, *user priority* was compared before *modulation format*. This is reflected in the resulting lower throughput at the right.

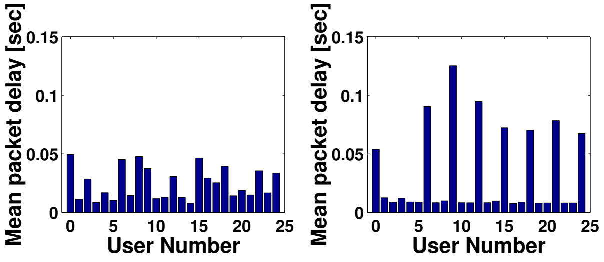

Figure 4: The delay profile for the 25 different users included in the simulation depicted in Figure 3, averaged over ten simulations. To the left, the average delay is lower, since *modulation format* is compared before *user priority*. To the right, we see a more differentiated delay profile, since higher priority users always get access to the channel before lower priority users. The users belong to the classes MEDIA, VOICE, VOICE, ..., MEDIA, VOICE, VOICE, MEDIA from left to right.

The scheduling algorithm was modified between the two simulations, by interchanging the parameters used for classification of the different users and channels. To the left, *modulation format* is compared before *user priority*, and, to the right, vice versa. As indicated by Figure 3, the parameters used in the scheduling process have a large impact on the performance of the scheduler.

The delay performance of the scheduler decisions is depicted in Figure 4. The delay is measured from the arrival of the packet, to the end of the time-frame in which the packet was completely transmitted. The absolute values of the delays should not be given too much importance, since they depend on the packet size, the packet inter-arrival time, the time-frame size, and the desired error probability. One should instead compare the two diagrams in Figure 4, and realize that the scheduler performance can be adapted to the traffic conditions.

6 CONCLUSIONS AND FUTURE WORK

Using a constant modulation format will result in error bursts due to fading, which the FEC codes cannot completely cope with. This is concluded in [5]. For data transmissions, errors will result in re-transmissions, invoked by an ARQ mechanism. The adaptive modulation approach provides a relatively constant error rate (Figure 3, bottom), which in turn provides an excellent basis for FEC codes, such as convolutional codes or block codes.

By introducing the adaptive modulation approach, we gain:

1. The error rate is kept at a constant level, thus feeding the FEC algorithms with manageable data.

2. Radio transmission is postponed when channel conditions are bad, thus reducing the interference affecting other terminals.

By adding the multiple access scheduler, keeping multiple links on a single frequency, we gain two more things:

1. System throughput can be maximized for a given frequency band (Figure 3, left). The more users we add, the more efficiently we use the frequency band.

2. User satisfaction becomes the central issue, rather than the allocation of some fixed number of time-slots in a varying environment (Figure 3, right).

Exploiting the flexibility of the scheduler by changing the order in which it compares the different users' parameters, we can adapt the scheduling performance to the current traffic situation:

1. Comparing *priority* before *modulation* results in a scheduling decision that more strictly obeys the priority demands (Figure 4, right). This is most convenient for low or medium traffic loads.

2. Doing it the other way, comparing *modulation* before *priority*, will generate a higher throughput in the system. This is convenient to alleviate the effects of a high system load.

The following topics will be investigated in the immediate future:

- Scheduling involving more than one frequency band at a time. High-priority users can choose between several independent channels on different frequencies.
- The performance gains as a function of prediction error levels and prediction horizon will be quantified.
- A deeper analysis of the required signalling overhead, the predictor initialization procedure, and the required hardware, will be carried out and included in the evaluations of the proposed systems.

7 ACKNOWLEDGEMENTS

This work was supported by the Foundation for Strategic Research, within the PCC-program, in Sweden.

The simulations were carried out using the IT++ package developed at the Communications Group, Department of Signals and Systems, Chalmers University of Technology, Sweden.

REFERENCES

[1] Toyoki Ue, Seiichi Sampei, and Norihiko Morinaga, "Adaptive Modulation Packet Radio Communication System using NP-CSMA/TDD Scheme", *IEEE Vehicular Technology Conference Proceedings*, pp. 416–420, May 1996.

[2] Soon-Ghee Chua and Andrea Goldsmith, "Variable-Rate Variable-Power MQAM for Fading Channels", *IEEE Vehicular Technology Conference Proceedings*, pp. 815–819, May 1996.

[3] Torbjörn Ekman and Gernot Kubin, "Nonlinear prediction of mobile radio channels: Measurements and MARS model designs", *IEEE International Conference on Acoustics Speech and Signal Processing*, Phoenix Arizona, March 1999.

[4] J.M. Torrance and L. Hanzo "Adaptive Modulation in a Slow Rayleigh Fading Channel", *PIMRC'96*, Oct. 1996, Taipei, Taiwan.

[5] Nilo Casimiro Ericsson, "Adaptive Modulation and Scheduling for Fading Channels", submitted to *IEEE Globecom Conference Proceedings*, Rio de Janeiro, December 1999.

[6] John G. Proakis, *Digital Communications*, 3rd ed. McGraw-Hill Book Co., Singapore 1995

[7] Shuang Deng, Alan R. Bugos, and Paul M. Hill "Design and Evaluation of an Ethernet-Based Residential Network", *IEEE Journal on Selected Areas in Communications*, pp. 1138–1150, vol. 14, no. 6, August 1996.

[8] Mark E. Crovella and Azer Bestavros "Explaining World Wide Web Traffic Self-Similarity", *Technical Report TR-95-015 - Revised*, Computer Science Department, Boston University, 1995

[9] Frederick S. Hillier and Gerald J. Lieberman *Introduction to Operations Research*, 5th ed. McGraw-Hill Book Co., Singapore 1990