Frequenz
60 (2006)
5-6

QoS compliance through cross-layer assisted resource allocation

By Benno Zerlin, Christian Hartmann, Jörg Eberspächer, Josef A. Nossek

Abstract – This article adresses the QoS compliant delivery of packet data flows through the MAC and PHY layers of a wirless communication system. The available degrees of freedom in the call admission control, the scheduling unit, the MAC protocol as well as in the signal processing units of the PHY layer are employed to maximize the resource efficiency or equivalently minimize the amount of resources required for a quality compliant service. In order to solve the resulting cross-layer problem we make three contributions: First, a generic approach derives an optimum solution to the problem of jointly optimizing the radio parameters in a system that is constrained by upper layer requirements. Second, QoS aware scheduling algorithms are proposed that access the cross-layer information provided by the above scheme. Based upon these scheduling techniques a third part targets the problem of call admission control. Basing its decision on cross-layer knowledge gathered in the above schemes the admission control can be designed to maximize the number of flows that can be served QoS compliantly. The proposed compound of cross-layer techniques is evaluated numerically.

Index Terms – Scheduling, cross-layer optimization

1. Introduction

Design and optimization of communication systems are conventionally based upon a vertically layered system description. Different layers within a stack only employ the functionalities of their lower neighbor and export functionalities exclusively to the next higher layer. This way the mutually decoupled investigation of different layers is enabled. Although this approach has served well in the design of a wide number of communication systems it results in clearly suboptimum system configurations and is not capable of solving local tradeoffs in a globally optimum sense. To overcome these drawbacks cross-layer optimization has evolved. It aims at improving the interactions of different layers by introducing a limited additional information exchange and by allowing for a certain degree of cooperation among the layers.

This article presents a cross-layer technique to provide a required set of *quality of service* (QoS) parameters in a wireless communication system. Under these constraints the framework is capable of maximizing the system efficiency through the cross-layer optimum allocation of the system's resources which implies the joint optimization of all radio parameters. This functionality is obtained through a two stage procedure. At first a closed analytic expression for the regarded class of QoS parameters provides the means to optimize all radio parameters for a given scheduling setting. These considerations in Section 2 explicitely include the coupling of different flows through interference on the broadcast channel, the mechanisms of *forward error correction* (FEC) coding and the effects of *hybrid automatic repeat request* (HARQ) protocols. With the optimization of parameters contained in these units the scheduling unit can access system resources by posing QoS constraints to the lower layers. Moreover it is granted access to information about the feasibility of a scheduling constellation which results from the lower layer optimization.

This low complexity representation of lower layers is employed by a novel MAC scheduling scheme introduced in Section 3. It bases on an introduced urgency level of each user in relation to the urgency level in the cell. Urgency within this context is defined through the so called residual time, which denotes the time a flow can wait before a service with a certain throughput will lead to a QoS violation. The flows with the highest urgency level are then scheduled to the PHY layer. The corresponding values for delay and throughput are then passed to the derived cross-layer optimization as requirements to the lower layers which verifies the feasibility of this constellation and determines the optimum PHY resource allocation.

The overall compound is completed by a corresponding call admission control in Section 4. Matching the urgency considerations of the proposed MAC scheduling scheme, the admission control is formulated through the introduced residual time as well and decides whether a new stream can be admitted or not.

Including all regarded layers and sublayers of the system we have evaluated our proposal in combined PHY-MAC layer simulations. The resulting performance in Section 5 shows the superiority and the significant gains in servable system load that can be achieved through an QoS aware cross-layer resource allocation.

2. Optimization of radio parameters

This section first introduces the theoretical means to express the QoS parameters throughput and delay in an HSDPA like system. The mathematical properties of this model allow for the formulation of a cross-layer optimum mode selection and resource allocation scheme. Subsection 2.3 finally reviews the cross-layer interface that is available to the MAC procedures in Section 3.

2.1 System model

Aiming for a mode dependent mapping of resources to QoS parameters, we introduce the mode of operation $M_k = \{A_k, R_k, N_k\}$ of user $k = 1, ..., K$ as an element of the Cartesian product $\{4,16\}\times[0;1]\times\{1,2,...,15\}$ that contains the cardinality of the chosen modulation alphabet A_k, the code rate R_k, and the number of CDMA code channels N_k. Depending upon this mode the following derivations formulate an expression for the QoS parameters throughput ρ_k and delay τ_k. The latter is defined in an outage sense as the time which in 98% of all cases suffices to transmit a packet error free over the lower layers.

2.1.1 Broadcast channels

We model the multiple access channel through the SINR as visible to the decoder input:

$$\gamma_k = \frac{\chi r_k P_k}{\sum_{i=1}^{K} \upsilon r_k P_i + P_\eta}. \qquad (1)$$

Assuming MRC receivers this SINR can base upon the sum r_k of squared absolute values of the channel coefficients. In the case of Rayleigh fading channels the corresponding distribution can be obtained through Lemma 4.3.b.1 from [1]. The product of this amplification with the corresponding transmit power P_k is multiplied by the spreading factor χ. In parallel the interference expression in the denominator is formulated through the abstract calculus of orthogonality factors υ. Because of the use of a long scrambling code and because the spreading codes are orthogonal

in synchronous use υ is constant for all i, including $i = k$. Moreover P_η denotes the receiver noise. As the channel coefficients in general are time variant so is the SINR. However, the remainder of this section uses a constant SINR $\gamma_k^{(\mathrm{rq})}$ in combination with the outage probability:

$$\pi_{\mathrm{out},k} = \Pr\left(\gamma_k < \gamma_k^{(\mathrm{rq})}\right). \tag{2}$$

2.1.2 Channel coding

Based on the above model of broadcast channels, the cutoff rate theorem allows us to formulate an upper bound for the code word error probability of block codes. Through a linearization of the Gallager error exponent, the analytic modeling of the relation between decoder SINR and error probability in coded transmission systems is enabled. Moreover and in contrast to capacity based approaches, it includes the complete mode dependency, i.e. the influence of modulation alphabets with finite cardinality A_k and finite block lengths beside the binary code rate R_k ld A_k. Aiming for the employment of these very favorable properties within the cross-layer system model, [2] introduces the cutoff rate theorem, that bounds the error probability of a block code with block length B and code rate R_k ld $A_k < R_0(\gamma_k)$ in bits per channel use by:

$$\pi_{\mathrm{pe}} < 2^{-B\left(R_0(\gamma_k) - R_k \,\mathrm{ld}(A_k)\right)}. \tag{3}$$

The cutoff rate $R_0(\gamma_k)$ is defined as a function of the conditional probability density of obtaining a channel output given a certain channel input. For discrete channels with SINR γ_k and a modulation alphabet of cardinality ld A_k the computation of the cutoff-rate can be found in [3].

2.1.3 Hybrid ARQ protocols

The following finds an analytical treatment for HARQ mechanisms in terms of the probability $f_m[m]$ of necessary HARQ transmissions m.

Sparing *incremental redundancy* methods we assume that retransmissions take place with identical parity information, i.e. Chase combining is used to superimpose the soft values of multiply received packets. As the packets in the different transmissions of this mode do not differ and all face independent noise realizations on the channel, a soft combining of these packets superimposes noise components incoherently, resulting in a cumulative SINR increase $\Delta\gamma_k[m]$. Together with the means to quantize this SINR enhancement [4], which depends on the employed modulation alphabet and the specifics of the FEC code, the packet error probability after m transmissions results as:

$$\pi_{\mathrm{pe}}[m] < 2^{-B\left(R_0(\gamma_k[m]) - R_k \,\mathrm{ld}(A_k)\right)}. \tag{4}$$

This equation allows us to formulate the probability, that it takes m transmissions to decode a packet error free, as the product of the probability of loosing m-1 consecutive packets and successfully transmitting the mth. The probability $f_m[m]$ thus is given by:

$$f_m[m] = \left(\prod_{m'=1}^{m-1} \pi_{\mathrm{pe}}[m']\right)\left(1 - \pi_{\mathrm{pe}}[m]\right). \tag{5}$$

Employing a proof from [5] which holds for large block sizes B the remainder of this article assumes $f_m[m]=\delta[m-m^*]$, where m^* is defined as the smallest m for which $\pi_{\mathrm{pe}}[m] < \varepsilon$.

2.1.4 QoS expressions

The above considerations base upon the assumption of a constant SINR γ_k. Including the outage considerations from the broadcast section we can express the probability that it takes n slots of length T until a packet is transmitted error free as:

$$f_n[n] = \pi_{out}^{n-m^*}\left(1-\pi_{out}\right)^{m^*}\binom{n-1}{m^*-1}. \tag{6}$$

This very general QoS description comprises the targetted parameters throughput and outage delay which can be obtained as:

$$\rho_k = \frac{m^*}{1-\pi_{out}}\frac{B}{T}. \tag{7}$$

$$\tau_k = \arg\min_{\tau'} \tau' \;\mathrm{s.t.}: \sum_{n=1}^{\lfloor \tau'/T \rfloor} f_n[n] > 98\% \tag{8}$$

2.2 Optimization

Publications like [6] have introduced a generic approach for the solution of the problem:

$$\left\{P_1, M_1, \ldots, P_K, M_K\right\}^{\mathrm{opt}} = \arg\min_{P_1, M_1, \ldots, P_K, M_K} \sum_{k=1}^{K} P_k \;\; s.\mathrm{t.}: \begin{cases} \rho_k \geq \rho_k^{(\mathrm{rq})} \\ \tau_k \leq \tau_k^{(\mathrm{rq})} \end{cases}. \tag{9}$$

These generic considerations can be applied to provide cross-layer optimum resource allocation and mode selection schemes on the background given in this article as well. The iterative optimization procedure consists of three core steps:

- Computation of equivalent requirements based upon an assumption on the outage probability along:

$$\tilde{\rho}_k^{(\mathrm{rq})} = \frac{1}{1-\hat{\pi}_{\mathrm{out}}}\rho_k^{(\mathrm{rq})}; \quad \tilde{\tau}_k^{(\mathrm{rq})} = \left\lfloor (1-\hat{\pi}_{\mathrm{out}})\frac{\tau_k^{(\mathrm{rq})}}{T} \right\rfloor T. \tag{10}$$

These are the requirements that have to be met in a zero-outage single user system in order to fulfill $\rho_k^{(\mathrm{rq})}$ and $\tau_k^{(\mathrm{rq})}$ in the original system.

- Mode optimzation along:

$$\gamma_k^{(\mathrm{rq})}, M_k^{(\mathrm{opt})} = \arg\min_{\gamma_k, M_k} \gamma_k \;\mathrm{s.t.}: \begin{cases} \tilde{\rho}_k \geq \tilde{\rho}_k^{(\mathrm{rq})} \\ \tilde{\tau}_k \leq \tilde{\tau}_k^{(\mathrm{rq})} \end{cases}. \tag{11}$$

These problems due to the outage based formulation of the link can be solved through a lookup table.

- Computation of resulting outage probability and update along the rule:

$$\hat{\pi}_{\mathrm{out}}[i] = \pi_{\mathrm{out}}[i-1]. \tag{12}$$

For a proof of convergence and optimality of this iterative procedure we refer to [6]. If the optimization is performed - frequently enough to assume a constant channel the approach collapses to a single table based iteration. Finally the downlink power control problem remains to be sovled, cf. [6].

2.3 Crosslayer interface

The above handling allows for a compact representation of the lower MAC and the PHY layer in the compound of our cross-layer resource allocation scheme. These lower sublayers can be accessed by a set of user demands on throughput and latency to which it responds with a feasibility flag and possibly with the amount of spare resources, i.e. an indication of the fractional system load. As the mode selection and the power allocation scheme both are optimum in the formulated sense, the compound acts at its economic maximum.

3. Urgency-feasibility scheduling

The task of the packet scheduler in the described system is to choose the next flow or group of flows to be served by the physical layer. A fundamental trade-off in packet scheduling is one between fairness and throughput [7,8]. For instance the *Round Robin* (RR) and *Weighted Round Robin* (WRR) schedulers provide for equal time share and equal throughput, respectively, while performing relatively poor in terms of cell throughput. On the other hand, the *Maximum Rate* (MR) scheduler maximizes the throughput while not considering any notion of fairness.

The *Proportional Fair* (PF) scheduler, which has been proposed for instance for use in HSDPA systems, attempts to balance the described trade-off by considering a metric which is the ratio between the feasible rate and the average throughput of a flow [9].

However, none of these schedulers is able to deal with explicit QoS requirements. Therefore we propose a scheduling approach, which takes into account the individual QoS constraints, the transmission history, as well as the current packet queue of each flow. Considering these inputs at the scheduling instant, the scheduler selects a flow or a group of flows, which is then assigned resources by the physical layer. The proposed scheduler, namely *Urgency Feasibility* (UF) scheduler is described in the sequel.

The functionality of UF scheduling consists of three steps:
1. Determination of the urgency of each flow.
2. Determination of the urgency level of the cell.
3. Selection of a set of scheduling candidates to be assigned resources on the physical layer.

The three steps are briefly described in the following paragraphs:

3.1 Flow urgency – residual time

Our goal is to serve each flow, such that its respective QoS requirements are met. For this purpose we want to determine the urgency of each flow to be served at the current instant. The challenge here is to find an urgency measure that enables the comparison among flows which might have very different sets of QoS requirements. For this purpose we define the *Residual Time* t_R as an urgency measure of a given flow as follows:

Given the current packet queue and transmission history of a flow, how long can the flow wait to be served, until transmitting with its mean data rate will just in time avoid a fatal QoS violation.

Analytical expressions have been derived for t_R for a number of QoS constraints, like minimum mean data rate, receive buffer continuity, maximum mean packet delay, and maximum mean packet loss rate [10].

Note that if a flow has more than one QoS constraint, the residual time is computed for each constraint separately and then the minimum is considered to be the current residual time of that flow. As an example, the residual time is computed for the receive buffer continuity in the sequel.

We assume, the transmitted data is buffered at the receiver and $R(t)$ denotes the receive buffer level. The QoS constraint of receive buffer continuity can thus be expressed as:

$$R(t) > 0 \qquad \forall t. \tag{13}$$

With the receive buffer outflow rate ρ_0 and $D(t)$ denoting the data that is successfully transmitted in $[0, t]$, we can write:

$$R(t) = D(t) - \rho_0 t. \tag{14}$$

At some time instant $t + \Delta t$ we have:

$$R(t + \Delta t) = D(t + \Delta t) - \rho_0 \cdot (t + \Delta t). \tag{15}$$

For the determination of t_R, we assume that no data is transferred in $[t, t + \Delta t]$, and thus $D(t + \Delta t) = D(t)$. If $t + \Delta t$ is the time instant for which $R(t + \Delta t) = 0$, we can set Δt to t_R and have:

$$\Delta t = \frac{D(t)}{\rho_0} - t. \tag{16}$$

As soon as a packet is transmitted successfully, the receive buffer level is increased. Therefore we account for the transmission time of the first packet in the transmission queue when determining t_R and with the current mean transmission data rate of the user ρ_{TX}, we obtain:

$$t_R = \frac{D(t)}{\rho_0} - \frac{d_1}{\rho_{TX}(t)} - t, \tag{17}$$

with d_1 denoting the size of the first packet in the transmission queue. If the receive buffer level $R(t)$ is available for the computation of t_R, we can write:

$$t_R = \frac{R(t)}{\rho_0} - \frac{d_1}{\rho_{TX}(t)}. \tag{18}$$

3.2 Cell urgency level

In the previous section, we described how we can quantify the urgency of each individual flow. In addition to this, we define the *Cell Urgency Level* (CUL) as a measure for the overall urgency of transmissions within the cell. A simple but efficient definition for the CUL making use of the residual time of individual flows is based on the most urgent flow in the cell:

$$CUL = \frac{1}{\min_{i=1,\ldots,K}(t_R^{(i)})}, \tag{19}$$

where K is the number of flows currently in the system.

In the following section we describe, how the CUL can be used to adaptively tune the trade-off between maximum throughput and QoS fairness depending on the load situation.

3.3 Selection of scheduling candidates

Using the above defined measures, our packet scheduler determines the set of candidate flows U as follows:

$$U = \left\{ flow^{(i)} \mid t_R^{(i)} \le t_R^* \right\}. \tag{20}$$

The adaptive behaviour of the scheduler is due to the fact, that the threshold residual time t_R* is not fixed but a function of the CUL, which is defined as:

$$t_R^* = \begin{cases} t_x + t_y + \left[\min\left(t_R^{(i)}\right) + t_0 - t_x \right]^n - t_0^n & for \quad \min\left(t_R^{(i)}\right) \geq t_x \\ \min\left(t_R^{(i)}\right) & else \end{cases}$$

with $t_0 = (1/n)^{1/(n-1)}$. The parameters t_x, t_y, and n can be chosen to achieve different characteristic behaviors of the scheduler. For instance by choosing $t_y = 0$, $n = 1$, and arbitrary t_x, the UF scheduler can be forced to behave like the *Earliest Deadline First* (EDF) scheduler, implying that the deadlines are considered to be the computed values for t_R. Another extreme case can be achieved by choosing $t_x = 0$, $t_y \to \infty$, and arbitrary n, which yields the MR scheduler. However, to benefit from the adaptability of the UF scheduler, it is most appropriate to choose intermediate values, such that as the urgency level (the CUL) changes, the scheduling behavior gradually shifts between the two extreme cases.

4. Admission control

By applying advanced scheduling methods, we can increase the probability that all flows in the system achieve their required QoS. However, in order to guarantee the users in the system a success probability of, say 99%, we need to limit the number of flows in the system. For this purpose an Admission Control mechanism is required, which decides if a new flow generated at time t can be admitted into the system or not. It is obvious that such a mechanism should be matched with the scheduler that the system uses. Therefore, for our system using the UF scheduler, we use an admission control measure, which is derived from the residual time values. Specifically, we define the *Cell Load Level* (CLL) as follows:

$$CLL = \min_{k=0,\ldots,K-1} \left(t_R^{k+1} - \sum_{j=1}^{k} \frac{\sum_{i=1}^{I_j} d_i^{(j)}}{\rho_{TX}^{(j)}(t)} \right), \qquad (21)$$

where $d_i^{(j)}$ denotes the size of the ith packet in the transmission queue of flow j and I_j is the number of backlogged packets of flow j. In (21) the K flows currently in the system are sorted according to ascending residual time. We can interpret (21) as the time before transmitting the first packet of all queues in the order of ascending residual time, is just still feasible without QoS violations. As an efficient admission control mechanism, we propose to define an admission threshold value CLL_{Th}, which is compared to an exponentially smoothed average of the CLL to decide if a new flow can be admitted or not.

5. Cross-layer simulations

In order to assess the potential of our proposed cross-layer system, we have integrated all models described above into one simulation tool. Due to limited space, we can only roughly sketch the structure of our cross-layer simulation model as follows: Flows are generated according to a Poisson traffic model with different packet characteristics within each flow. Upon initialization, each flow is associated with a set of QoS requirements and a user position, uniformly chosen in the cell area. During the duration of the flow, the users move according to a random direction mobility model. In each scheduling interval the UF scheduler determines the set U among all flows in the system. The set U, containing the scheduling candidates in ascending order of their residual time is then passed to the physical layer module. Along with each user in U the scheduler passes to the physical layer module: the distance of the user to the base station, a delay requirement $\tau_k^{(rq)} = t_R$ equal to the residual time and a transmission rate requirement $\rho_k^{(rq)}$. The physical layer module applies a pathloss model with attenuation exponent 3.5 plus a Rayleigh fading model and determines the physical layer resources (power, time slots, codes) assigned to each user. Resources are assigned in ascending order of residual time, such that the delay and rate requirements for each flow are met as long as resources are available. A feedback informs the scheduler, which flows the physical layer module was able to accommodate, such that the packet queues can be updated and the simulation can proceed.

As a proof of concept we include an early simulation result comparing the proposed UF scheduling approach with the RR packet scheduler. In this case all flows are admitted into the system (no admission control applied). The results in Fig. 1 show the performance advantage of the cross-layer assisted UF resource allocation with respect to Round Robin scheduling.

6. Conclusions

We have presented a resource allocation scheme for the QoS compliant optimization of parameters within the RLC, MAC and PHY layer. It consists of the interaction of admission control, scheduling and a cross-layer handling for the lower sublayers. Through an analytical system model and the subsequent optimization of the lower layers a compact representation was obtained, that allowed the capacity enhancing formulation of scheduling and admission control techniques through the concept of urgency levels. Simulations back the made contributions and demonstrate the potential performance gains.

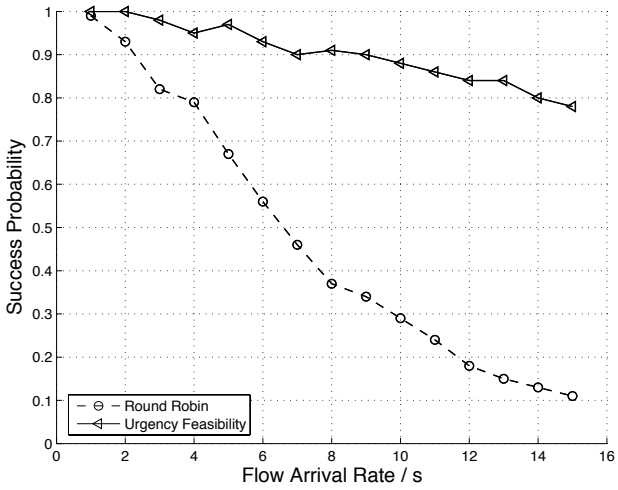

Fig. 1: Performance advantage of the cross-layer assisted UF resource allocation with respect to Round Robin scheduling.

References

[1] S. B. Provost, A. M. Mathai, *Quadratic Forms in Random Variables*. Marcel Dekker Inc., 1992

[2] R. G. Gallager, *Information Theory and Reliable Communications*. John Wiley and Son, 1968

[3] M. T. Ivrlac, T. P. Kurpjuhn, C. Brunner, W. Utschick, „Efficient Use of Fading Correlations in MIMO Systems“, *Proceedings of the 54th IEEE Vehicular Technology Conference*, 2001

[4] M. Dottling, T. Grundler, A. Seeger, „Incremental Redundancy and Bit-Mapping Techniques for High Speed Downlink Packet Access”, *Proceedings of the Global Telecommunication Conference*, 2003

[5] G. Caire, D. Tunietti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel", *IEEE Transactions on Information Theory*, vol. 47, pg. 1971-1988, 2001
[6] B. Zerlin, J. A. Nossek, "A Generic Approach to Cross-Layer Assisted Resource Allocation", *Proceedings of the ITG/IEEE Workshop on Smart Antennas*, 2006
[7] S. Borst, „User-level performance of channel-aware scheduling algorithms in wireless data networks", *Proceedings IEEE Infocom*, 2003
[8] P. Liu, R. Berry, M. L. Honig, „Delay-sensitive packet scheduling in wireless networks", *Proceedings IEEE WCNC*, 2003
[9] P. Viswanath, D. N. C. Tse, R. Laroia, „Opportunistic beamforming using dumb antennas", *IEEE Transactions on Information Theory*, 48(6):1277-1294, June 2002
[10] C. Hartmann, R Vilzmann, A. Schmitt-Nilson, J. Eberspächer, „Channel aware scheduling for user-individual QoS provisioning in wireless systems", *Proceedings of the IEEE Vehicular Technology Conference*, Fall 2004

Dipl.-Ing. Benno Zerlin
Lehrstuhl für Netzwerktheorie und Signalverarbeitung
Technische Universität München
80290 München
Germany
Fax: +49–89–289 28504
E-mail: benno.zerlin@tum.de

Dr.-Ing. Christian Hartmann
Lehrstuhl für Kommunikationsnetze
Technische Universität München
80290 München
Germany
Fax: +49–89–289 23523
E-mail: hartmann@tum.de

Prof. Dr.-Ing. Jörg Eberspächer
Lehrstuhl für Kommunikationsnetze
Technische Universität München
80290 München
Germany
Fax: +49–89–289 23523
E-mail: joerg.eberspaecher@tum.de

Prof. Dr. techn. Josef A. Nossek
Lehrstuhl für Netzwerktheorie und Signalverarbeitung
Technische Universität München
80290 München
Germany
Fax: +49–89–289 28504
E-mail: nossek@tum.de

(Received on March 24, 2006)
(Revised on March 29, 2006)