Adaptive Modulation and Coding with Hybrid-ARQ for Latency-constrained Networks

Tania Villa†, Ruben Merz*, Raymond Knopp†, Uday Takyar‡
† Eurecom
∗ Telekom Innovation Laboratories
‡ EPFL, School of Computer and Communication Sciences
tania.villa@eurecom.fr, ruben.merz@telekom.de, raymond.knopp@eurecom.fr, uday.takyar@epfl.ch

***Abstract*—Traffic generated by machine to machine (M2M) communication or online gaming will be a large and integral part of the traffic transported by LTE-advanced and beyond networks. This traffic is characterized by sporadic and low-throughput packet arrivals. It must be scheduled under a latency constraint. Sporadic traffic creates environments where the channel-quality information (CQI) is outdated or unavailable. Fast-fading, the non-stationarity of inter-cell interference and the heterogeneity of Rel-10 networks further exacerbates this issue. However, current LTE-Advanced schedulers and resource allocation schemes are not optimized for these particular scenarios. In this paper, we propose a scheduling and resource allocation mechanism for latency-constrained operation. Our solution significantly improves the spectral efficiency of delay-constrained networks by optimizing a joint hybrid-ARQ and adaptive modulation and coding (AMC) policy that changes the number of dimensions (physical resources) used in each round. With only one bit of feedback, obtained causally from hybrid-ARQ, we achieve a performance close to the ergodic capacity.**

I. Introduction

High-performance online gaming, machine-to-machine (M2M) and sensor data communications are emerging massive applications for cellular networks. A typical example of M2M applications in mobile environments is sensors connected to public transport vehicles, to trains or to equipment in factories. M2M communications are part of the Internet of Things (IoT) revolution. M2M is expected to create an increasing number of connected devices, which will exceed human-to-human communications over the following years (50 billions machines against seven billion people for 2011) [1], [2]. A large class of the traffic generated by these emerging applications can require low-latency [3], [4]. For example, for online gaming application, the low-latency is critical to offer the best game experience as possible [5]. Large portions of M2M applications are expected to produce sparse traffic with low-delay constraints. The two main reasons being power reduction through discontinuous reception (DRX) and transport of small sporadic packets to M2M devices. Concretely, a user-equipment (UE) terminal emerging from an idle state to deliver a small packet to the network should reconnect for the smallest amount of time possible to conserve power. Moreover, the paradigm of many sparsely connected UEs transmitting sporadic traffic poses interesting problems related to resource allocation policies in a scheduled-access MAC protocol such as that of 3GPP Long Term Evolution (LTE).

It is predicted that these applications, in addition to voice and Internet traffic, will be an integral part of the traffic transported by LTE [6] and LTE-Advanced [7] networks. M2M is expected to account for a considerable amount of the traffic of such networks [2].

Although significant headway in latency reduction has been made thanks to the development of LTE, the LTE architecture and protocols can still be improved for these application areas in order provide efficient spectral usage for M2M services [4] or to satisfy low-latency requirements. The extremely large number of devices connected to the network, the expected reliability of the service regardless of the operation environment, and low-latency requirements from applications such as emergency messages, video surveillance or health-care will require some enhancements to the network including link adaptation protocols, modulation and coding, and hybrid automatic repeat and request (hybrid-ARQ or simply HARQ) schemes [8], [9]. All these network optimizations will be included as a part of the LTE-Advanced standard since M2M communication is one of the main focuses in LTE-Advanced [2].

In this paper, we concentrate on the HARQ, adaptive modulation and coding (AMC) and physical resource allocation mechanisms. We consider optimized rate-adaptation policies in the case of sparse and latency-constrained traffic (See Figure 1). Namely, we optimize the HARQ and AMC mechanism for these specific traffic characteristics. Our optimized policies are applicable for both downlink (DL) and uplink (UL) data.

The optimization is carried out by adapting the number of dimensions (physical resources e.g. sub-carriers or resource blocks in LTE) used in each HARQ round. Furthermore, because of the sparse traffic characteristic, of moderate to high mobility, of insufficient uplink CQI periodicity or of inter-cell interference, we investigate cases where the UL channel-quality information (CQI) is outdated or unavailable. In such cases, the scheduler must operate blindly with respect to AMC and can only benefit from binary feedback after the first HARQ transmission round (in the form of ACK/NACK signaling [6]).

Our contributions are the following:

- We derive analytical expressions, based on mutual information modeling, that capture the throughput performance of latency-constrained networks.
- We develop an optimized rate adaptation policy. This policy is based on the dynamic adaptation of the number of dimensions (resource blocks) used by each HARQ

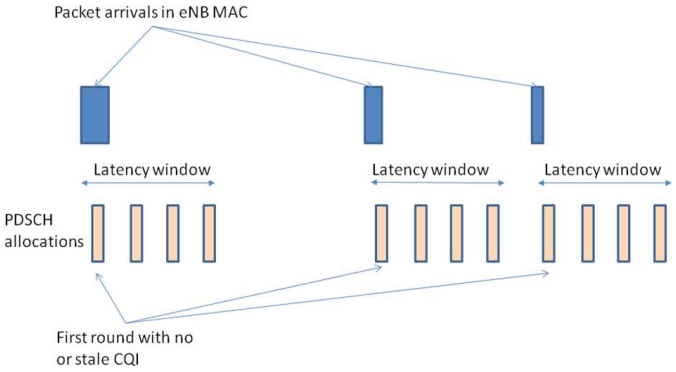

Fig. 1. Sparse traffic with delay-constrained scenario. Traffic arrivals in the eNB medium access control (MAC) layer are sparse as depicted in blue (there are three of them). The latency constrain is four slots, i.e. there are up to four possible PDSCH channel (see Section III) allocations. Because of the sparse traffic, channel-quality information (CQI) is outdated or unavailable on the first slot.

round which is a feature of the Rel-8/10 LTE coding and modulation subsystem.

- Surprisingly, this policy can operate with only one bit of feedback from the HARQ process. We also show that additional performance improvements are obtained when outdated channel-state information becomes available.

The remainder of this paper is organized as follows. Additional related work is presented in Section II and detail on the resource allocation mechanism in LTE systems in Section III. Our optimized joint HARQ and AMC mechanism for delay-constraint environment is exposed in Section IV. Its performance is evaluated in Section V. Finally, we detail its application to LTE in Section VI and conclude in Section VII.

II. Related Work

HARQ is a combination of traditional ARQ and error correction codes [10] and a widely deployed building block of current mobile communication systems. It generally achieves a better performance than ordinary ARQ techniques. With AMC, the so-called modulation and coding scheme (MCS) can be adapted dynamically to optimize the throughput while keeping a target error rate. Link-adaptation mechanisms such as HARQ and AMC are necessary to achieve the high peak data-rates of LTE-A and beyond networks [6]. However, all these techniques typically require sophisticated channel estimation schemes in order to obtain channel state information (CSI) and to provide CQI feedback for the adaptation mechanisms. The transport of CQI can also result in a significant amount of control signaling overhead. Hence, because of resources constraints, prior work on M2M communications for 3GPP networks has suggested the use of fixed and non-adaptive mechanisms [11]. But as we show in this paper, even in cases without or just outdated feedback, it is possible to obtain significant spectral efficiency and outage probability improvements with appropriately designed mechanisms.

M2M communication will also need to address issues related to large number of users accessing the network. The work in [11], [12] propose to group users with similar quality of service (QoS) characteristics and requirements into access clusters. However, this solution only addresses medium access control (MAC) with large number of users and different QoS requirements. Besides, jitter (defined as the difference in time between two successive packet arrivals or departures) was the only performance metric considered. However, error probability, throughput or spectral efficiency cannot be ignored. In [13], [14], a resource allocation scheme specific to healthcare applications employs admission control for traffic prioritization and delay minimization. This scheme is well-suited for voice-communication over a cellular network. However, our data traffic requirements are fairly different as mentioned in Section I and this scheme is not applicable to our scenario.

In a more general setting, [15] proposes a rate and power adaptation scheme, based on perfect CSI, for a fading channel scenario. Queuing is also taken into account and the trade-off between average delay and average transmission power is investigated. Similarly, [16] derives a rate and power mechanisms to optimize delay for single-user system with queuing. No retransmission protocol is taken into account and lower and upper asymptotic bounds are obtained. However perfect CSI is always assumed at the receiver. The effect of imperfect CSI on ARQ and HARQ is investigated in [17], [18]. In particular, [17] concentrates on optimizing the amount of energy devoted to channel estimation in order to maximize the throughput. For instance, allocating more power for channel estimation leaves less energy for data transmission, resulting in higher error rate. Unlike previous work, power control across HARQ rounds is explored in [19], namely to minimize the packet error rate under an average transmit power constraint. This mechanism requires feedback at the receiver and it might be difficult to use for M2M devices. Indeed, they usually have no or limited access to power sources because of their low power consumption [2]. Finally, [20] considers the trade-off between coding and ARQ in a fading channel. It shows that choosing a high error probability (10% or higher) maximizes the throughput for a HARQ protocol with a fixed outage probability. It is also shows that an optimal rate exists for each round.

Compared to the related work present in this Section, our work differs significantly. We address cases with no or only outdated CQI. Furthermore, we do not perform any power control because it is impractical or simply not feasible for M2M scenarios. Rather, we develop an optimized rate adaptation policy that changes the number of physical resources (i.e. dimension) across rounds. Our optimization is based on the outage constraint when a retransmission protocol is used. With respect to [19], we show that only one bit of feedback (ACK or NACK) of the HARQ protocol is sufficient for significant improvements in packet error rate. Even without any CQI, our results show dramatic error-rate reductions and improvements of the spectral efficiency.

III. Resource Allocation in LTE Systems

LTE-advanced release 10 (and higher) is the evolution of the LTE standard [7], [6]. It provides higher peak data-rates (1 Gbps UL, 500 Mbps DL) and spectral efficiency, support for very flexible deployment scenarios including DL/UL asymmetric bandwidth allocations and non-contiguous spectrum

allocation. Applications such as online gaming and M2M systems are integral parts of LTE-advanced [6].

As explained earlier in Section II, LTE supports both HARQ and AMC for efficient resource allocation. But they are not the only mechanisms. LTE also supports the adaptation of the so-called transport-block size and the amount of physical resource blocks (PRB) used per transmission. A transport-block is the name given to a block of data at the MAC layer [21]. For LTE, HARQ is supported for the physical layer shared channels in both UL (PUSCH) and DL (PDSCH), and separate control channels are used to send the associated acknowledgment feedback. HARQ can be classified as either synchronous or asynchronous and the retransmissions can be adaptive or non-adaptive [6]. In a synchronous system, retransmissions occur at a predefined time. In an asynchronous system, the retransmissions can occur any time (and must be signaled). With adaptive HARQ, the MCS and other transmission attributes can be changed after each round. In a non-adaptive context, transmission attributes are fixed or pre-defined. In LTE, HARQ is asynchronous and adaptive in the DL and synchronous in the UL. Retransmissions can be adaptive or not in the UL [6]. Feedback for HARQ in LTE comprises a simple ACK/NACK signal. The HARQ and AMC algorithm is implemented and managed by the scheduler at the MAC layer. Based on the available CQI , it is the scheduler at the base-station (eNodeB or eNB in LTE) that can address the different quality and latency requirements of all the associated user equipments (UE).

An important factor for the eNB scheduling algorithm is the accuracy of the available CQI for the active UEs in the cell. In the DL, CQI is reported back by the UE. For the UL, the eNB can use sounding reference signals (SRS) or other signals transmitted by the UEs to estimate CQI [6]. On the UL, two channels are used to send CQI reports. Namely the physical uplink control channel (PUCCH) and the physical uplink shared channel (PUSCH). Reporting can be periodic or aperiodic. Periodic reports are normally transmitted on the PUCCH. However, the eNB can request the user equipment (UE) to send aperiodic CQI reports on the PUSCH because it is more appropriate to transmit large and detailed reports [22]. The key issue with respect to resolving channel quality is the ability to obtain accurate information in the presence of a large number of connected UEs (even idle) and the presence of sporadic interference (primarily on the uplink) in heterogeneous network deployments. The eNB can adjust the periodicity and granularity of the CQI feedback (down to 2 ms periodicity for both CQI and SRS in Rel-10 LTE [23]) allowing it to trade-off between the amount of overhead and the accuracy of the channel information. Of course, when a long delay occurs with respect to the scheduling time, the performance can be significantly affected. Short feedback periodicity is difficult to achieve in heavily loaded cells.

In summary, LTE offers a lot of flexibility in terms of resource allocation and, in particular, resource allocation algorithms can be tailored for a particular class of traffic with specific requirements. Nevertheless, work is still needed to exploit this flexibility efficiently for key emerging applications. To this end, in the remainder of this paper we develop and evaluate a resource allocation mechanism for sparse latency-constrained traffic. Furthermore, we consider cases where CQI feedback is either unavailable, or outdated.

IV. Optimized hybrid-ARQ and AMC Policies for Delay-Constrained Operation

Compared to ARQ, using HARQ increases spectral efficiency while ensuring reliability. But further improvements can be obtained by combining HARQ and AMC. When implementing AMC with HARQ, the MCS is adapted between each retransmission. Indeed, as our results show in Section V, adapting the rate assignment between retransmission rounds can significantly enhance the performance.

In the following, we develop and analyze a combined hybrid-ARQ and AMC policy for sparse and latency-constrained traffic scenarios. Packet arrivals are sporadic and must be scheduled under a latency constraint (see figure 1). In this context, CQI is typically outdated or unavailable. Note that outdated CQI also occurs because of moderate to high mobility, of insufficient uplink CQI periodicity or of non-stationary inter-cell interference. The latter will become more and more important with LTE release 10 networks and their inherent heterogeneity. Hence, the scheduler must operate blindly for AMC and can only benefit from feedback after the first HARQ transmission round in the form of ACK/NACK signaling.

A. Signal Model and Assumptions

In the following, we present the signal model and assumptions for the remaining of this Section. Without loss of generality, we consider OFDM signaling. The UL of an LTE system uses an SC-FDMA modulation. Our joint HARQ and AMC policy applies equally, but the signaling details differ. This is left for an extension of this work. Therefore, for a particular sub-carrier, let x denote the complex-valued transmitted symbol, z denote the additive white Gaussian noise (AWGN), and h denote the channel gain. Both z and h are modeled with a zero-mean and unit variance complex Gaussian random variable. Let l denote the discrete-time index i.e. $x[l]$ is the lth transmitted symbol. Using y to denote received symbols, the lth received symbol in a particular sub-carrier is

$$y[l] = h[l]x[l] + z[l], \qquad l = 1, 2, \ldots, N. \tag{1}$$

We consider a block-stationary Rayleigh fading channel model. Fading remains static for the duration of a HARQ round but varies between retransmissions. The HARQ feedback channel is assumed to be error-free. CQI can be received after each round. However, prior to the first round, CQI may or may not be available. As explained earlier, this can occur because of sparse traffic. Furthermore, because of fast-fading, (low) mobility or non-stationary inter-cell interference, CQI can at any round be simply unusable. Consequently, we analyze cases where, at the first transmission round, CQI is either outdated or simply unavailable. On the further rounds, we keep on assuming that CQI is not available. But, we take advantage of the one bit of ACK/NACK information given to the transmitter after each HARQ round.

For the outdated CQI case, it is assumed that the fading statistics are available to the transmitter. This assumption is reasonable because the eNB scheduler can maintain a database of channel measurements in its cell, allowing it to derive the fading statistics over time.

We do not consider interference created by neighboring transmitters. This case is left for further study. We focus on single antenna systems, although our model can be extended to multiple-input and multiple-output systems (MIMO).

B. Modeling and Optimization of a joint Hybrid-ARQ and AMC Policy

We consider a one-shot transmission model where one transport-block of size N_{TB} arrives in sub-frame n and must be served at maximum spectral-efficiency under a latency constraint. We denote by N_R the maximum number of transmission rounds. To characterize code performance and the effect of the channel, we use the instantaneous mutual information in each transmission round. This theoretic measure, although asymptotic, provides a very accurate indication of potential performance in LTE, whose coded-modulation subsystem performs close to asymptotic limits. Let H_i denote the vector of channel realizations in the ith transmission round. Then $I(H_i)$ denotes the corresponding instantaneous mutual information. Accordingly,

$$I(H_1, \ldots, H_{N_R})$$

defines the mutual information accumulated over N_R transmission rounds. In order to compute the mutual information, we assume Gaussian input signals (upper-bound on QAM modulation). For example, let us consider one sub-carrier of a SISO system without interference and let P denote the received power, $h_{0,i}$ is the channel response at round i and N_0 is the noise power, then

$$I(H_1, \ldots, H_{N_R}) = \sum_{i=1}^{N_R} \log_2 \left(1 + \frac{P|h_{0,i}|^2}{N_0}\right). \quad (2)$$

Generalizing the notation from [24], the probability of decoding a transport-block in round n with N_j as the number of dimensions used in round j is

$$\Pr\Big(I(H_1, \cdots, H_n) > (\sum_{j=1}^{n} N_j) R_n, \; I(H_1, \cdots, H_i) < (\sum_{j=1}^{i} N_j) R_i, \forall i < n\Big). \quad (3)$$

Let $\mu(n)$ denote the target transport-block error probability after n transmission rounds. The latency constraint is expressed by ensuring that the probability that the transport-block is not served after N_R transmission rounds is below $\mu(N_R)$. Under this framework, AMC is the optimization of the rate sequences R_i such that (1) the packet error probability remains below $\mu(N_R)$ after N_R transmission rounds and (2) the spectral-efficiency is maximized. The optimization is carried out as a function of the distribution of $I(H_1, \ldots, H_{N_R})$.

For simplicity, we consider at most two retransmission rounds (ARQ rounds), but our policy can also be applied for more than two. We consider three scenarios

1) Minimal-latency: a trivial case of serving the packet in one round which corresponds to the minimal-latency AMC policy.
2) Latency-constrained with no prior CQI: we consider two transmission rounds and no information about the channel.
3) Latency-constrained with outdated CQI: we consider again two transmission rounds, but unlike the previous case, we assume that we have outdated information about the channel with some correlation with the actual channel.

For simplicity and in the interest of obtaining semi-analytical results, we concentrate on one sub-carrier, i.e. that H_i is a scalar. In an upcoming full simulation study using fully-compliant LTE modem implementations, this is reconsidered using wideband 3GPP-SCM channel models.

C. Scenario Analysis: Minimal-latency

We consider first the trivial case of serving the transport-block in one round. This is the minimal-latency AMC policy. The rate allocation law for R_1 is given by the solution to

$$\Pr\big(I(H_1) < R_1\big) = \mu(1). \quad (4)$$

Without any a priori information regarding the channel statistics, this essentially says that the best that can be done is to transmit with the lowest spectral-efficiency coding scheme (i.e. lowest MCS) to minimize latency. With a priori information. the largest MCS such that the probability of channel realizations requiring a smaller MCS is still below the threshold is chosen.

Let H_0 denote the channel corresponding to outdated CQI. If stale CQI is available prior to transmission of the transport-block, then the rate should be chosen such that

$$\Pr\left(I(H_1) < R_1 | H_0\right) = \mu(1) \quad (5)$$

in order to take into account the outdated CQI.

D. Scenario Analysis: Latency-constrained with no Prior CQI

We now consider the case with two transmission rounds. Let B define the number of information bits to be transmitted. Let N_T denote the total number of dimensions available and let N_1 denote the number of dimensions used in the first round. Hence, the rate in the first round is $R_1 = \frac{B}{N_1}$, and the rate in the second round $R_2 = \frac{B}{N_T}$. We define

$$\lambda = \frac{N_1}{N_T}. \quad (6)$$

and we can relate R_1 to R_2 with $R_2 = \lambda R_1$. Let $\bar{R}$ denote the overall spectral efficiency. With $\mu(1)$ as the outage probability after the first round, we have

$$\begin{aligned} \bar{R} &= R_1(1 - \mu(1)) + \mu(1) R_2 \\ &= R_1(1 - \mu(1)) + \mu(1)\lambda R_1. \end{aligned} \quad (7)$$

We want to maximize $\bar{R}$ such that the probability of outage after the second round is below the given constraint $\mu(2)$. For the first round, there is no feedback information. The outage probability $\mu(1)$ is unknown but it depends on H_1 and the signal-to-noise ratio (SNR). We can relate R_1 to $\mu(1)$ as follows. From equation (4), we have

$$\Pr\left(I(H_1) < R_1\right) = \Pr\left(\log_2\left(1 + SNR|h_1|^2\right) < R_1\right) = \mu(1). \tag{8}$$

Consequently, we obtain

$$R_1 = \log_2\Big(1 - SNR\ln(1-\mu(1))\Big). \tag{9}$$

In the second round, feedback about the previous round is available. The outage probability is now given by

$$\Pr\Big(I(H_1, H_2) < R_2 | I(H_1) < R_1\Big) = \mu(2) \tag{10}$$

We can rewrite equation (10) as follows

$$\Pr\Big(I(H_1, H_2) < R_2 | I(H_1) < R_1\Big) = \frac{\Pr(I(H_1, H_2) < R_2, I(H_1) < R_1)}{\Pr(I(H_1) < R_1)} = \frac{\int_0^{\frac{2^{R-1}-1}{SNR}} e^{-|h_1|^2} d|h_1|^2}{\mu(1)} - \frac{\int_0^{\frac{2^{R-1}-1}{SNR}} e^{-a-|h_1|^2} d|h_1|^2}{\mu(1)} = \mu(2) \tag{11}$$

where

$$a = \left(\left(\frac{2^{R_1}}{1+SNR|h_1|^2}\right)^{\frac{\lambda}{1-\lambda}} \frac{1}{SNR}\right) - \frac{1}{SNR} \tag{12}$$

and the limits stem from the fact that if $I(H_1) < R_1$ then $|h_1|^2 < \frac{2^{R_1}-1}{SNR}$. The integrals in equation (11) are evaluated numerically.

To find the optimal value of R_1 in the first round, we perform an extensive exploration on $\mu(1)$, given that we want to maximize equation (7) and subject to the constraint $\mu(2)$ in equation (11).

E. Scenario Analysis: Latency-constrained with Outdated CQI

We now analyze the case when outdated CQI becomes available to the transmitter. We make the additional assumption that the channel remains constant over the two transmission rounds and let $h = h_1 = h_2$. Furthermore, we denote by h_0 the channel value that corresponds to the outdated CQI. In order to model a possible correlation between h_0 and h, we use the following model. Let ρ be the correlation parameter, then

$$h = \sqrt{\rho}h_0 + \sqrt{1-\rho}h'$$

where h_0 and h' are i.i.d. Gaussian-distributed random variables. Note that in this case,

$$\rho = \mathbb{E}\left[h_0 h^*\right].$$

In addition, $|h|^2$ is a non-central Chi-square random variable with two degrees of freedom. We follow the same general procedure to obtain the throughput and probability of outage than in the previous cases. However, the spectral efficiency is a function of the outdated CQI and we have to average over the distribution of $|h_0|^2$.

First, let $\gamma_1 = \sqrt{\frac{2^{R_1}-1}{SNR}}$ be the outage threshold in the first round and $\gamma_2 = \sqrt{\frac{2^{R_2}-1}{SNR}}$ be the outage threshold in the second round. Then, $\Pr\left(h > \gamma_1\right)$ represents the probability of having a successful transmission in the first round, $\Pr\left(h < \gamma_1, h > \gamma_2\right)$ is the probability of being unsuccessful in the first round but successful in the second round, and $\Pr(h < \gamma_2)$ gives the probability of being in outage. All these probabilities are a function of the cumulative distribution function (CDF) of $|h|^2$. The non-centrality parameter of $|h|^2$ is $s^2 = \rho|h_0|^2$. Let $F_\chi\left(x\right)$ denote the CDF of $|h|^2$. It can be expressed in terms of the Marcum Q-function [25] i.e.

$$F_\chi\left(x\right) = 1 - Q_1(s, x) \tag{13}$$

where $Q_M(a, b)$ is the Marcum Q-function with parameters M, a and b.

The overall spectral efficiency $\bar{R}$ over the two ARQ rounds can be written as

$$\bar{R} = \Pr\left(h > \gamma_1\right) R_1 + \Pr\left(h > \gamma_2, h < \gamma_1\right) R_2. \tag{14}$$

To find the optimal rates, we first obtain R_2 from the outage constraint $\mu(2)$. Since we know that $h < \gamma_2$ implies an outage, R_2 is given by solving equation (15) for R_2. Therefore

$$\mu(2) = \Pr\left(|h|^2 < \frac{2^{R_2}-1}{SNR}\right) = F_\chi\left(\gamma_2^2\right) \tag{15}$$

Next, to find the value of R_1 that will maximize the overall spectral efficiency, we first write (14) in terms of the Marcum Q-function. We have

$$\bar{R} = Q_1(a_1, b_1)R_1 + (Q_1(a_2, b_2) - Q_1(a_1, b_1))R_2 \tag{16}$$

where $a_1 = a_2 = s$, $b_1 = \gamma_1$, and $b_2 = \gamma_2$. We now take the derivative of equation (16) with respect to R_1, and we solve for R_1 when the derivative is zero. We obtain

$$\begin{aligned} \frac{\partial \bar{R}}{\partial R_1} &= \frac{\partial Q_1(a_1, b_1)}{\partial R_1} R_1 + Q_1(a_1, b_1) - \frac{\partial Q_1(a_1, b_1)}{\partial R_1} R_2 \\ &= \frac{\partial Q_1(a_1, b_1)}{\partial R_1}(R_2 - R_1) + Q_1(a_1, b_1) = 0. \end{aligned} \tag{17}$$

To find the derivative of the Marcum Q-function in (17), we used [26].

V. Performance Evaluation

In this section, we present numerical results in terms of (1) the probability of outage and (2) the achieved spectral efficiency. Remember that we assume Gaussian signaling to compute the mutual information. Throughout this section, we fix the spectral efficiency to 2 bits per channel use. A more comprehensive set of rates will be detailed in a further study. The maximum number of retransmissions is one round (at most two rounds).

Figure 2 presents the minimum SNR necessary to achieve a given outage probability $\mu(2)$. For a given value of $\mu(2)$, we calculate the corresponding SNR for our rate adaptation

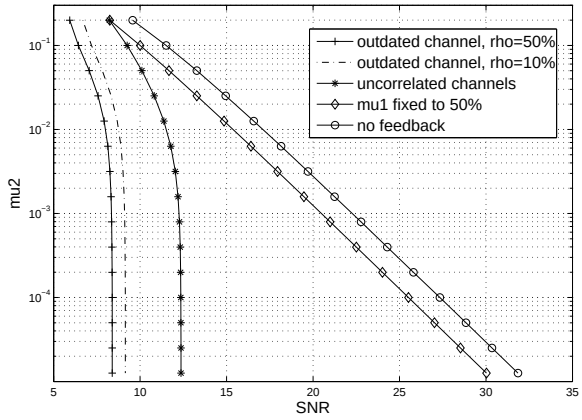

Fig. 2. For different values of the probability of outage after the second round μ_2, we calculate the corresponding SNR for the different scenarios. The symbol ρ is the correlation coefficient between the actual channel and the channel corresponding to outdated CQI information. We compare a correlation coefficient value of $50\%, 10\%$ and uncorrelated case. For comparison purposes, we also plot two more cases. First when no ACK/NACK feedback is available from the HARQ process. Second, when μ_1 is fixed to 50% with $\lambda = 0.5$ to make sure that 50% of the dimensions are used in each round.

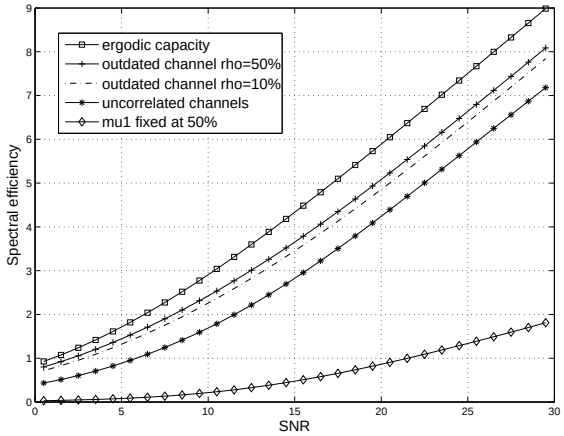

Fig. 3. Spectral efficiency versus SNR for the different scenarios. We set $\mu(2)$ to 1%. The symbol ρ is the correlation coefficient between the actual channel and the outdated/stale CQI information. We compare correlation coefficients of $50\%, 10\%$ and uncorrelated case. For comparison purposes we plot the curve for the ergodic capacity and μ_1 fixed to 50% with $\lambda = 0.5$ to make sure that 50% of the dimensions are used in each round.

policy. For the scenario (3) from Section IV, the outdated CQI has a correlation coefficient with the actual channel of $\rho = 10\%$ or $\rho = 50\%$. For comparison purposes, we consider two more cases in addition to the scenario (2) and (3) from Section IV. First we evaluate a case where we force the probability of outage after the first round to 50%, fixing $\lambda = 0.5$ to make sure that 50% of the dimensions are used in each round. Typically, while conventional systems try to ensure a 10% outage probability per slot, we observe from our results that a higher value gives, in fact, a higher overall spectral efficiency. Second, we evaluate a case where no feedback at all is available i.e. when we can not even receive ACK/NACK from the HARQ process. This highlights the significant gain from adapting the rate across rounds with only one bit of feedback, even in the case without any CQI information. The gain is even higher when only outdated CQI information is available. Our rate adaptation policy gives a zero probability of outage without the need of having a high SNR. From the results in Figure 2 we can observe that it does not make a difference to increase the SNR above 12.5 dB for the case without CQI. We show that adjusting the dimensions used in each round results in almost causal feedback performance. In our scenarios, the two rates are simply controlled by $\lambda = \frac{R_2}{R_1}$, which depends on the SNR and target outage probability $\mu(2)$. We only need one bit of feedback, which we get causally from HARQ. In fact, it is the state of the channel that actually chooses the code rate. By choosing the rate in the first round as high as possible, we can guarantee a probability of outage after the second round while maximizing the spectral efficiency.

Figure 3 presents the overall spectral efficiency obtained for a given SNR. We set $\mu(2)$ to 1%. For the outdated CQI case, we consider $\rho = 50\%$ and $\rho = 10\%$. For reference purpose, we also plot the ergodic capacity (Rayleigh channel capacity), i.e. for perfect rate adaptation. Finally, we again consider a scenario when the rate in the first round is chosen so that the probability of outage after the first round is fixed to 50%. This value is chosen because it gives the highest spectral efficiency. Fixing the probability of outage after the first round to more or less than 50% gives, in fact, a lower overall spectral efficiency.

From our results we see a significant improvement in spectral efficiency even in the case without CQI. When we can benefit from outdated CQI, we achieve a performance close to the ergodic capacity.

VI. Application in LTE Rate-Adaptation

The policies considered in Section IV require an incremental-redundancy coded-modulation system with the possibility of adjusting the code rate (physical resources) via puncturing and repetition (rate matching). This is possible on the LTE Rel-8/10 DL and UL where the allocation can be adapted across transmission rounds, the granularity of which depends on the transport block size. The only difference between the UL and DL allocation is that the modulation order must remain fixed on the UL for each round. It is likely that this restriction is insignificant with respect to performance. The slight penalty for inter-round rate adaptation is the requirement to send a new DL control information (DCI) packet with updated resources rather than an automatic retransmission. Let us now explain with an example the rate-adaptation policy for two-round transmission in an LTE context.

For illustration purposes, consider a DL transmission in transmission mode 1 (SISO transmission). The policy must choose the initial spectral efficiency, MCS_1 corresponding to the first transmission round and N_1 and N_2 corresponding to the number of allocated physical resource blocks. An additional constraint (on the downlink) is that these must be multiples of an integer P which is the allocation granularity dictated by the transmission bandwidth (for a 10MHz carrier, $P = 3$ [23].) Figures 4 and 5 show the result of an

optimized two-round protocol, specifically the rate R_1 which must be translated to MCS_1 and λ which provides the ratio of the physical dimensions used in the two rounds. This can now be transformed as a function of the target transport block size. Now consider an average SNR of 7 dB, this corresponds R_1 of 2 bits/dimension and $\lambda = 0.07$. Assume an allocation in the first round using 3 resource blocks (minimal allocation for a 10 MHz carrier). In a PDSCH-only normal prefix subframe with one PDCCH symbol, the total number of resource elements is $N_1 = 450$ (13 PDSCH symbols, 3 with 10 PDSCH resource elements per resource block and 10 with 12.) The target transport block size would therefore be around 900 bits, so the closest transport block size for 3 resource blocks is 904 bits ([23, Table 7.1.7.2.1-1]) with $\mathrm{MCS}_1 = 16$ (16QAM). The second-round dimensions would ideally be $N_2 = N_1(\frac{1-\lambda}{\lambda}) = 5978$ yielding 39.9 (39 or 42) resource blocks.

The previous example exposes the difficulty of applying such latency-constrained policies for large transport block sizes, at least under the constraints of the current LTE specifications. For instance, if we were to use the same operating point (7 dB average SNR) with twice the number of dimensions in the first round, corresponding to a transport block size of 1800 bits, the number of required resource blocks in the second round jumps to 78 or 81 which is only realizable on a 20 MHz carrier. This can be easily overcome by allowing a third transmission round, i.e. by increasing the acceptable latency of the transmission.

Another important consideration regarding application on the UL is that the nature of the resource allocation policy is fundamentally related to power control, since we are assuming a constant transmit energy per channel dimension. This is also the adopted policy in LTE (assuming power adjustements are not made during retransmission rounds). Basically, low power is used during the first transmission and significantly more power is used in the second transmission if required. In the numerical example above the power boost is $10\log_{10}\left(\frac{1-\lambda}{\lambda}\right) = 11.2\mathrm{dB}$. This, of course, requires that the UE has signalled sufficient *power headroom* [23] for the eNB to allow this allocation. This clearly shows that, on the UL, latency can be controlled through a combination of rate adaptation, HARQ and power control. An instance of this appearing in the literature in the case of MIMO transmission with HARQ can be found in [27], although in that example the number of dimensions across tranmsission rounds remained fixed and the energy per channel dimension increased across rounds.

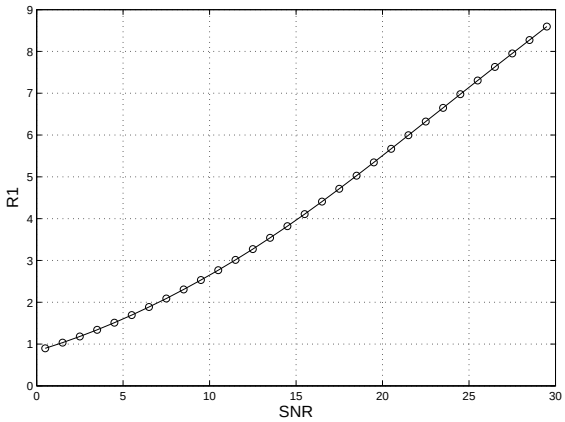

Fig. 4. We consider the scenario without CQI (uncorrelated channels), and we plot the rate in the first round (R1) for different values of SNR. We fix the probability of outage after the second round to 1%.

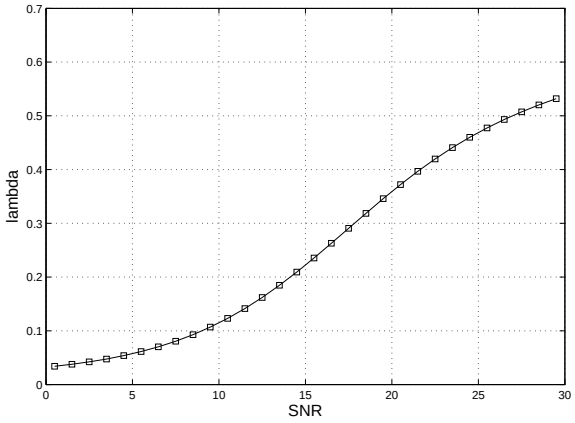

Fig. 5. For the scenario without CQI (uncorrelated channels), we plot the λ parameter against different values of SNR. We fix the probability of outage after the second round to 1%. λ determines the rate used in the second round according to equation [6]

VII. Conclusions

Two potential extensions to improve latency performance can be considered. First, more detailed feedback can be used to indicate the remaining amount of information required to decode based on the first transmission. Second, power-ramping can be used as an incremental power policy to accelerate decoding by increasing the power as rounds progress. For sparse packet arrivals this may increase aggregate spectral efficiency with respect to a constant power policy. More detailed feedback can be achieved through exploitation of CQI information (wideband and sub-band) after the first round of transmission, assuming that the eNB schedules PUSCH feedback to cover the ACK/NACK and CQI information. The CQI just represents the state of the channel after a particular transmission round.

For future work, several directions are available. First we will extend our model to consider more than two transmission rounds as well as the interaction with the scheduling of multiple UEs. We will also take into account interference. Finally, we plan to implement and evaluate our mechanism in simulation and on a software-defined radio platform.

References

[1] N. Nikaein and S. Krea, "Latency for real-time machine-to-machine communication in LTE-based system architecture," *Wireless Conference 2011 - Sustainable Wireless Technologies (European Wireless), 11th European*, pp. 1 –6, april 2011.

[2] Y. Chen and W. Wang, "Machine-to-machine communication in LTE-A," in *Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd*, sept. 2010, pp. 1 –4.

[3] IST-2003-507581 WINNER D1.3 version 1.0, "Final usage scenarios." [Online]. Available: http://www.ist-winner.org

[4] K. Zheng, F. Hu, W. Xiangy, M. Dohler, and W. Wang, "Radio resource allocation in LTE-Advanced cellular networks with M2M communications," *IEEE Communications Magazine, accepted for publication*, 2012. [Online]. Available: http://www.cttc.es/resources/doc/110330-m2m-ieeecomsmag-final-14964.pdf

[5] J. Manweiler, S. Agarwal, M. Zhang, R. Roy Choudhury, and P. Bahl, "Switchboard: a matchmaking system for multiplayer mobile games," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*, ser. MobiSys '11. New York, NY, USA: ACM, 2011, pp. 71–84. [Online]. Available: http://doi.acm.org/10.1145/1999995.2000003

[6] S. Sesia, I. Toufik, and M. Baker, *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley Publishing, 2009.

[7] S. Parkvall, A. Furuska andr, and E. Dahlman, "Evolution of LTE toward imt-advanced," *Communications Magazine, IEEE*, vol. 49, no. 2, pp. 84 –91, february 2011.

[8] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson, "M2m: From mobile to embedded internet," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 36 –43, april 2011.

[9] A. Lioumpas, A. Alexiou, C. Anton-Haro, and P. Navaratnam, "Expanding LTE for devices: Requirements, deployment phases and target scenarios," *Wireless Conference 2011 - Sustainable Wireless Technologies (European Wireless), 11th European*, pp. 1 –6, april 2011.

[10] S. Lin and D. J. C. Jr., *Error control coding - fundamentals and applications*, ser. Prentice Hall computer applications in electrical engineering series. Prentice Hall, 1983.

[11] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 66 –74, april 2011.

[12] S.-Y. Lien and K.-C. Chen, "Massive access management for qos guarantees in 3GPP machine-to-machine communications," *Communications Letters, IEEE*, vol. 15, no. 3, pp. 311 –313, march 2011.

[13] K.-D. Lee and A. V. Vasilakos, "Access stratum resource management for reliable u-healthcare service in LTE networks," *Wirel. Netw.*, vol. 17, pp. 1667–1678, oct 2011. [Online]. Available: http://dx.doi.org/10.1007/s11276-011-0371-6

[14] ——, "Managing resources for healthcare service calls in a cellular network with relays and machine-type communication nodes," in *Proceedings of the First ACM MobiHoc Workshop on Pervasive Wireless Healthcare*, ser. MobileHealth '11. New York, NY, USA: ACM, 2011, pp. 1:1–1:8. [Online]. Available: http://doi.acm.org/10.1145/2007036.2007038

[15] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *Information Theory, IEEE Transactions on*, vol. 48, no. 5, pp. 1135 –1149, may 2002.

[16] I. Bettesh and S. Shamai, "Optimal power and rate control for minimal average delay: The single-user case," *Information Theory, IEEE Transactions on*, vol. 52, no. 9, pp. 4115 –4141, sept. 2006.

[17] L. Cao, P. Y. Kam, and M. Tao, "Impact of imperfect channel state information on arq schemes over rayleigh fading channels," in *Communications, 2009. ICC '09. IEEE International Conference on*, june 2009, pp. 1 –5.

[18] L. Cao and P.-Y. Kam, "On the performance of packet ARQ schemes in rayleigh fading: The role of receiver channel state information and its accuracy," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 2, pp. 704 –709, feb. 2011.

[19] T. Chaitanya and E. Larsson, "Outage-optimal power allocation for hybrid arq with incremental redundancy," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 7, pp. 2069 –2074, july 2011.

[20] P. Wu and N. Jindal, "Coding versus arq in fading channels: How reliable should the phy be?" in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, 30 2009-dec. 4 2009, pp. 1 –6.

[21] 3rd Generation Partnership Project, 3GPP TS 36.211, "Technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); physical channels and modulation," December 2008.

[22] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE-Advanced*. John Wiley & Sons, 2011. [Online]. Available: http://books.google.fr/books?id=X9XwEOxYnAkC

[23] G. T. . 3rd Generation Partnership Project, "Technical specification group radio access network; evolved terrestrial radio access (e-utra); physical layer procedures," March 2011.

[24] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *Information Theory, IEEE Transactions on*, vol. 47, no. 5, pp. 1971 –1988, july 2001.

[25] J. G. Proakis, *Digital Communications*, 4th ed. McGraw–Hill, 2001.

[26] A. Annamalai and C. Tellambura, "A simple exponential integral representation of the generalized marcum q-function qm (a, b) for real-order m with applications," in *Military Communications Conference, 2008. MILCOM 2008. IEEE*, nov. 2008, pp. 1 –7.

[27] H. El Gamal, G. Caire, and M. O. Damen, "The mimo arq channel: Diversity-mutliplexing-delay tradeoff," *IEEE Transactions on Information Theory*, pp. 3601–3621, August 2006.