Shariatmadari, Hamidreza; Iraji, Sassan; Li, Zexian; Uusitalo, Mikko A.; Jäntti, Riku

Optimized Transmission and Resource Allocation Strategies for Ultra-Reliable Communications

Published in:
2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)

DOI:
10.1109/PIMRC.2016.7794801

Published: 22/12/2016

Document Version
Early version, also known as pre-print

Please cite the original version:
https://doi.org/10.1109/PIMRC.2016.7794801
Optimized Transmission and Resource Allocation Strategies for Ultra-Reliable Communications

Hamidreza Shariatmadari∗, Sassan Iraji†, Zexian Li‡, Mikko A. Uusitalo†, and Riku Jäntti∗
∗Department of Communications and Networking, Aalto University, {firstname.lastname}@aalto.fi
†Nokia Bell Labs, {firstname.lastname}@nokia.com

Abstract—Fifth generation (5G) wireless systems will provide connectivity for a wide range of new applications with diverse requirements. In part, the network needs to support ultra-reliable communications with low-latency (URLLC) for mission-critical applications. For these applications, the generated data should be delivered with a limited number of transmission attempts with high success probability. This paper considers the optimal transmission and resource allocations for URLLC in cellular systems. The resource allocations are derived for the fixed and adaptive transmission attempt assignments. The analysis results reveal that both fixed and adaptive transmission assignments, applicable to automatic repeat request (ARQ) and hybrid ARQ (HARQ) schemes, can reduce the required resources compared to the equal transmission assignment.

Index Terms—5G, Ultra-reliable communications, low-latency, resource allocation, transmission allocation, ARQ, HARQ, finite-blocklength.

I. INTRODUCTION

The early evolution of cellular systems mainly results in providing high data transmission rates for mobile users. This trend has been changing and fifth generation (5G) wireless systems should provide connectivity for all set of new applications, including those which require very high data rates. The majority of new applications are related to machine-type communications (MTC). In order to enable MTC, future cellular networks need to encounter new features which are not supported in the current cellular systems [1].

MTC has a wide breadth of applications with a diverse range of requirements. For instance, mission-critical applications need ultra-reliable communications with low-latency (URLLC) [2][3]. This indicates that the generated messages should be delivered within a short time period with very high success probability. Examples of this type of applications include industrial automation, remote surgery, augmented reality, and vehicular communications. The required level of reliability and latency vary according to the applications. The Third Generation Partnership Project (3GPP) targets at providing reliability up to $10^{-5}$ within 1 millisecond (ms) in the future cellular systems [4].

Building URLLC over cellular systems has its own challenges. These challenges include transmissions efficiency, optimum resource allocations, accuracy of feedback information, and low-latency. Hence, various network enhancements have been proposed to enable URLLC in the future cellular networks. For instance, the over-the-air transmission latency can be reduced by employing short transmission time interval (TTI). Supporting device-to-device (D2D) communications between nearby devices can also reduce the delay by providing a direct communication link [5]. Several potential solutions exist for improving the link quality, such as: utilizing massive multiple-input-multiple-output (MIMO) antennas, or implementing relay nodes.

This paper studies transmission and resource allocations for URLLC. The data communication is considered between two devices communicating through a cellular network. Data transmissions is performed over two links, i.e. uplink (link between the transmitter device and a base station) and downlink (link between a base station and the receiver device). In order to meet the latency constraint, the maximum number of transmissions is limited for each payload transmission. For example, less than 4 transmission attempts can be assumed to meet 1 ms latency according to the proposed 5G frame structure with TTI around 0.125 ms [6][7]. We introduce two different schemes to efficiently assign the number of transmissions for uplink and downlink. In the fixed transmission allocation scheme, the maximum number of transmission rounds for each link is set prior to the actual data transmission. The appropriate resource allocation is determined according to the link qualities and the total number of transmission realizations. In the adaptive transmission allocation scheme, the number of transmission rounds for uplink and downlink can vary during the data transmission. The resource allocation is also adapted during the data transmissions. The optimization problems are derived for the mentioned transmission schemes, utilizing automatic repeat request (ARQ) and incremental redundancy hybrid ARQ (IR-HARQ) schemes. In addition, sub-optimal solutions are provided to reduce the complexity of the optimizations. The efficiency of these methods are assessed for finite-blocklength codes.

The rest of this paper is organized as follows. Section II describes the considered system model and assumptions. Section III and Section IV present the resource allocations for the fixed and the adaptive transmission schemes, respectively. Finally, conclusions are drawn in Section V.

II. SYSTEM MODEL

We consider a reliable communication link with low-latency between two devices in a cellular system. The two communicating devices may be served by a single base station or two base stations separately. For the case of serving devices by two different base stations, it is assumed that the base stations can
Interchange data flawlessly with negligible delay. Hence, the communication reliability in this case is similar to the case of serving both devices with a single base station. Without loss of generality, we can consider the reliable data communication in one direction. Fig. 1 illustrates the system model for delivering data from a transmitter device to a receiver device. In order to transmit a message, first the transmitter sends the message to the base station in uplink. When the base station receives the message correctly, it will forward it to the receiver device in downlink. To provide the communication reliability, a channel coding scheme is applied for data transmissions. An adaptive coding scheme is utilized separately for uplink and downlink. This means that the message is encoded and decoded independently in both links. A high level of reliability entails encoding the messages with relatively low rates. This requires allocating much radio resources, particularly, for a link with low signal-to-interference-plus-noise ratio (SINR). To improve the radio resource utilization, an error-control scheme is also applied. This allows transmitting data with a higher rate and retransmitting the data upon error [8].

As the low-latency is essential for data communication, we consider the latency constraint as a limit on the maximum number of transmissions. It is assumed that a maximum of \( M \) transmission attempts are allowed for each payload transmission over both links. This value can be determined according to the desired latency, TTI, and processing time for encoding and decoding data. Denoting the maximum transmission attempts in uplink by \( m_1 \) and in downlink by \( m_2 \), the delay constraint entails:

\[
m_1 + m_2 \leq M. \tag{1}
\]

Let us assume that the desired end-to-end communication reliability corresponds to BLER of \( \epsilon \). A message is considered as erroneous if it is not delivered correctly within the latency budget, i.e., performing \( M \) transmission attempts. We consider the constant BLER constraint for all payload transmissions to provide a uniform reliability regardless of link qualities [9]. Denote the residual BLER of receiving payloads at the base station after performing \( m_1 \) transmissions by \( \epsilon_1 \) and the residual BLER of receiving payload messages at the receiver device after performing \( m_2 \) transmissions by \( \epsilon_2 \). The desired reliability is achievable if:

\[
1 - (1 - \epsilon_1)(1 - \epsilon_2) \leq \epsilon. \tag{2}
\]

This indicates a message can be delivered successfully if it is passed through both links without error. Note that to guarantee the reliability target, both links should provide a higher level of reliability compared to the desired end-to-end reliability \((\epsilon_1 \leq \epsilon, \epsilon_2 \leq \epsilon)\).

We study the data transmission in quasi-static conditions, where the channel coefficients remain constant during a payload transmission, including retransmission attempts, and then changes independently to other values [10]. In addition, it is assumed that the base station(s) accesses to the instantaneous channel state information (CSI) of both links. The base station(s) should assign transmissions and radio resources for both links. The resource allocation is performed according to the quality of links and the reliability target. In our analysis, we use the results of [11], [12]. In particular, the achievable rate for a blocklength of \( n \) under a block error probability of \( \epsilon \) is tightly approximated:

\[
r = C - \sqrt{\frac{V}{n}} Q^{-1}(\epsilon) + \frac{1}{n} O(\log_2(n)), \tag{3}
\]

where \( C \) is the Shannon capacity, \( V \) is the channel dispersion, and \( Q(\cdot) \) denotes the Gaussian Q-function. For fast fading channel with SINR of \( \gamma \), we have:

\[
C = \log_2(1 + \gamma), \tag{4}
\]

\[
V = \frac{\gamma}{2(\gamma + 1)^2} \log_2(e). \tag{5}
\]

In this paper, we use the following expression to derive a lower bound on the expected BLER for delivering \( k \) \((k \geq 100)\) bits of information using \( n \) channel uses at the SINR of \( \gamma \) [8][13]:

\[
E(n, k, \gamma) = Q(\frac{nC - k}{\sqrt{nV}}). \tag{6}
\]

III. OPTIMAL RESOURCE ALLOCATION WITH FIXED TRANSMISSION ROUNDS

This section considers the resource optimization for the fixed transmission assignment. The main assumption here is that the number of transmission rounds cannot be changed for uplink and downlink during the data transmission. So, the optimal resource allocation consists of assigning the transmission rounds and allocating the channel uses for each round.

![Fig. 1: System model.](image-url)
for uplink and downlink prior to performing the actual data transmission.

A. ARQ

The ARQ scheme allows the receiver to ask for retransmitting the erroneous received data. For each payload, the data transmission is the same for the initial transmission and retransmissions. The retransmission for a payload continues until the receiver decodes a payload successfully or the maximum number of allowed retransmissions is reached. Let us first consider the data transmission over a single link $i \in \{1, 2\}$ (1: uplink, 2: downlink). Denote the SINR at the receiver by $\gamma_i$. Assume that a payload with size of $k$ bits is encoded into $n_i$ channel uses. In addition, the payload can be transmitted with the maximum of $m_i$ transmission rounds. The expected channel uses for transmitting the payload can be represented as:

$$N_i(n_i, m_i, k, \gamma_i) = n_i \sum_{j=1}^{m_i} (E(n_i, k, \gamma_i))^{j-1}. \quad (7)$$

This indicates that the initial transmission is always performed, while a retransmission round is performed if decoding the message in all the previous transmissions was unsuccessful. The optimal transmission and resource allocations can be determined by minimizing the sum of expected channel uses in both links, achieving the reliability target:

$$\arg \min_{n_1, n_2, m_1} \{N_1(n_1, m_1, k, \gamma_1) + N_2(n_2, M - m_1, k, \gamma_2)\} \quad (8)$$

s.t.

$$1 - [1 - (E(n_1, k, \gamma_1))^{m_1}] [1 - (E(n_2, k, \gamma_2))^{M - m_1}] \leq \epsilon.$$

This optimization jointly considers the transmission and resource allocations for both links. Fig. 2 represents the optimal transmission allocations between two links for different reliability targets, allowing 4 transmission attempts. In each case, the figure is divided into different regions, indexed by $(i, j)$ where $i$ and $j$ represent the assigned transmission rounds for the link 1 and link 2, respectively. It is evident that the resource optimization generally results in assigning more transmission rounds to the link with lower SINR level. The assigning of more transmission attempts to the link with the lower SINR is boosted for the higher reliability target. The optimization (8) needs to determine three parameters, which might be considered complex. Hence, we propose a suboptimal solution for this problem. For this purpose, we assume that the same BLER target is set for both links. To achieve the reliability target, we set $\epsilon_1 = \epsilon_2 = 1 - \sqrt{\frac{1}{4} - \epsilon}$, which represents the residual BLER after performing the transmission attempts over both links. For the link $i$ with a maximum of $m_i$ transmissions, the BLER for each transmission round can be set as $\sqrt{\epsilon_i}$ which guarantees the residual BLER target after performing all transmission rounds. This value can be used to determine the required channel uses in each round. This suboptimal solution can be expressed as:

$$\arg \min_{m_1} \{N_1(n_1, m_1, k, \gamma_1) + N_2(n_2, M - m_1, k, \gamma_2)\}, \quad (9)$$

where

$$n_i = \min n : E(n, k, \gamma_i) \leq \sqrt{\epsilon_i}. \quad (10)$$

The complexity of this suboptimal solution is reduced as the transmission allocation is the main parameter and the required channel uses should be determined at discrete values of $m_i$.

Fig. 3 illustrates the channel use gain achieved considering optimization (8) and (9) compared to the equal transmission allocation ($m_1 = m_2$) when totally 4 transmissions are allowed.

$$\frac{(N_1 + N_2)_{\text{equal allocation}} - (N_1 + N_2)_{\text{optimal allocation}}}{(N_1 + N_2)_{\text{equal allocation}}}.$$  

It can be observed that the achieved gain for the optimal (8) and suboptimal (9) solutions are very close. In addition, the achieved gains are increased for the higher reliability target, also for larger difference between SINR levels.

B. IR-HARQ

IR-HARQ scheme enables sending segmented parts of a message incrementally and decoding the message using combined received information. For each payload transmission, the receiver asks for retransmission if it cannot decode the message successfully using the received data segments while
the maximum number of transmissions limits is not reached. Again, let us first consider a single link. For the link \( i \) with a maximum of \( m_i \) transmission attempts, we define \( n_i^j = \{n_{i,1}, \ldots, n_{i,m_i}\} \), where \( n_{i,j} \) denotes the channel uses in \( j \)th HARQ round. Assuming that the segmented data transmission can meet the lower bound of the BLER (6), the residual BLER after \( j \)th HARQ round can be expressed as [13]:

\[
E\left(\sum_{l=0}^{j} n_{i,l}, k, \gamma_i\right),
\]

where \( n_{i,0} = 0 \). This equation indicates that the \( j \)th HARQ round is performed if the receiver cannot decode the message successfully using all previous received data segments. So, the residual BLER after performing all the transmissions merely depends on the whole blocklength and not on its segmentation. The expected utilized channel uses for this link can be expressed as:

\[
\overline{N}_i(n_i^j, m_i, k, \gamma_i) = \sum_{j=0}^{m_i} n_{i,j} E\left(\sum_{l=0}^{j-1} n_{i,l}, k, \gamma_i\right).
\]

Now, we consider the transmission and resource allocations for both links. The following optimization targets in minimizing the sum of expected required channel uses in uplink and downlink, considering the reliability target:

\[
\arg\min_{n_1, n_2, m_1} \{\overline{N}_1(n_1^j, m_1, k, \gamma_1) + \overline{N}_2(n_2^j, M - m_1, k, \gamma_2)\}
\]

s.t.

\[
1 - \left[1 - E\left(\sum_{l=0}^{M-m_1} n_{2,l}, k, \gamma_2\right)\right][1 - E\left(\sum_{l=0}^{M-m_1} n_{2,l}, k, \gamma_2\right)] \leq \epsilon.
\]

Fig. 4 represents the optimal transmission allocations between two links for different reliability targets, allowing 4 transmission rounds. Similar to optimal transmission for ARQ scheme, more transmission rounds are allocated to the link with the lower SINR level. In contrast, when the reliability target is increased, the unequal transmission allocation is occurred at higher SINR level differences. This is due to the fact that for optimal channel uses of HARQ scheme allowing two transmission attempts, the initial transmission is performed with a moderate BLER performance (around 10%) which results in utilizing less channel uses compared to the ARQ scheme [8][14]. In addition, when one transmission round is reduced from the link with the higher SINR level and assigned to the other link, the expected channel uses for the former link is increased significantly, while is decreased for the latter link. Considering these facts, for HARQ case, very high gain from allocating extra transmission should be achieved to compensate the loss from limiting the transmission to only one round. Again, in order to simplify the optimization, we can consider the same BLER target for both links as \( \epsilon_1 = \epsilon_2 = 1 - \sqrt{1 - \epsilon} \). This results in a suboptimal solution as:

\[
\arg\min_{m_1} \{\overline{N}_1(n_1^j, m_1, k, \gamma_1) + \overline{N}_2(n_2^j, M - m_1, k, \gamma_2)\},
\]

where

\[
\sum(n_i^j) = \min n : (n, k, \gamma_1) \leq 1 - \sqrt{1 - \epsilon},
\]

\[
n_1^j = \arg\min_{\vec{n}} \{\overline{N}_1(\vec{n}, m_1, k, \gamma_1)\},
\]

\[
\sum(n_2^j) = \min n : E(n, k, \gamma_2) \leq 1 - \sqrt{1 - \epsilon},
\]

\[
n_2^j = \arg\min_{\vec{n}} \{\overline{N}_2(\vec{n}, M - m_1, k, \gamma_2)\}.
\]

The sum(\(\vec{n}\)) represents the sum of all the vector elements. For this suboptimal solution, \( m_1 \) is the main parameter. For each link, the whole blocklength is initially determined regardless of the number of transmission rounds. Then, the optimal data segmentations are determined at discrete value of \( m_1 \). This approach separates the process of transmission and resource allocations, requiring less computations. Fig. 5 illustrates the achieved channel use gain considering optimization (13) and (14) compared to the equal transmission allocation \( m_1 = m_2 \) where totally 4 transmissions are allowed. It can be observed that the gain increases for a higher reliability target and higher SINR difference between two links. However, the achieved gains are less compared to the ARQ case (compared with Fig. 3) as HARQ scheme inherently provides better spectral efficiency.
IV. OPTIMAL RESOURCE ALLOCATION WITH ADAPTIVE TRANSMISSION ROUNDS

This section considers the resource allocation for the adaptive transmission scheme. As data transmissions are occurred in uplink and downlink successively, the resource allocation for the downlink can be performed adaptively after the uplink data transmission. In this way, for a payload transmission if the message is delivered in the uplink utilizing \( j \) transmission rounds, the radio resource for downlink can be allocated knowing that \( M - j \) transmission rounds are available.

A. ARQ

For the adaptive transmission allocation, the number of transmission rounds are not known before performing data transmission. For uplink, we define \( n_1 \) as the channel uses in each ARQ round. The data may be delivered to the base utilizing 1 to \( M - 1 \) transmission rounds. For downlink, the channel uses depend on the remaining transmission rounds. So, we define \( n_2 \) as the number of channel uses for each transmission round when \( j \in \{1, \ldots, M - 1\} \) transmission rounds are available. In a similar manner utilized in the previous section, the following optimization is considered to minimize the sum of expected required channel uses in uplink and downlink, guaranteeing the reliability target:

\[
\text{arg min}_{n_1, n_2^1, \ldots, n_2^{M-1}} \left\{ \sum_{j=1}^{M-1} \left( E(n_1, k, \gamma_1) \right)^{j-1} \times \left[ n_1 + N_2(n_2^{M-j}, M - j, k, \gamma_2) \right] \right\}
\]

\[
s.t.
1 - \sum_{j=1}^{M-1} \left( E(n_1, k, \gamma_1) \right)^{j-1} [1 - E(n_1, k, \gamma_1)]
\times \left[ 1 - (E(n_2^{M-j}, k, \gamma_2))^{M-j} \right] \leq \epsilon.
\]

In this optimization, there are \( M \) parameters that should be determined which improve the computation complexity compared to the fixed transmission allocation. To simplify the optimization, we propose a suboptimal solution requiring less computation. We set the downlink transmission reliability as \( 1 - \sqrt{1 - \epsilon} \) regardless of remained transmission rounds. This transmission reliability allows determining the resource allocation for the downlink according to the remaining transmission rounds. The suboptimal resource allocations are derived as:

\[
\text{arg min}_{n_1} \left\{ \sum_{j=1}^{M-1} \left( E(n_1, k, \gamma_1) \right)^{j-1} \times \left[ n_1 + N_2(n_2^{M-j}, M - j, k, \gamma_2) \right] \right\}
\]

\[
s.t.
\sum_{j=1}^{M-1} \left( E(n_1, k, \gamma_1) \right)^{j-1} [1 - E(n_1, k, \gamma_1)] \geq \sqrt{1 - \epsilon},
\]

where

\[
n_2^{M-j} = \min n : E(n, k, \gamma_2) \leq \frac{M-j}{\sqrt{1 - \sqrt{1 - \epsilon}}}.
\]

B. IR-HARQ

As mentioned earlier, in the adaptive resource allocation, the number of transmissions for downlink and uplink can vary from 1 to \( M \). So, for uplink we define \( \vec{n}_1 = \{n_{1,1}, \ldots, n_{1,M-1}\} \) where \( n_{1,j} \) denotes the number of channel uses in \( j \)th HARQ round. The HARQ transmissions in uplink stops if the base station decodes the message correctly or maximum of \( M - 1 \) rounds is reached. The channel uses for the downlink depend on the remainder transmission rounds. We define \( \vec{n}_2 = \{n_{2,1}, \ldots, n_{2,M-1}\} \) represents the channel use segmentation for the case that maximum of \( j \) transmission rounds can be realized. The following optimization can be considered to minimize the sum of expected channel uses in uplink and downlink:

\[
\text{arg min}_{\vec{n}_1, \vec{n}_2} \left\{ \sum_{j=1}^{M-1} E\left( \sum_{l=0}^{j} n_{1,l}, k, \gamma_1 \right) \times \left[ n_{1,j} + N_2(n_2^{M-j}, M - j, k, \gamma_2) \right] \right\}
\]

\[
s.t.
1 - \sum_{j=1}^{M-1} [1 - E\left( \sum_{l=0}^{j} n_{1,l}, k, \gamma_1 \right)][1 - E\left( \sum_{l=0}^{M-j} n_{2,l}, k, \gamma_2 \right)] \leq \epsilon.
\]

The complexity of this optimization is high as there are \( M \) vectors which should be determined. So, we propose a suboptimal solution to separate the uplink and downlink resource allocation to avoid the complexity.
The total number of transmission rounds are divided between uplink and downlink according to the link qualities. In the adaptive allocation scheme, the number of transmission rounds for each link is not fixed and can be changed during the data transmission. Both schemes offer the better resource utilization, in particular, for high reliability targets and asymmetric link qualities. These schemes can be utilized in the future cellular systems to better support URLLC.

**Fig. 7:** Channel use gain for adaptive transmission allocation applied to HARQ \((k = 100, \ SNR_1 = 0 \ dB)\).