mathematics

MDPI

Article

Analytical Model and Feedback Predictor Optimization for Combined Early-HARQ and HARQ

Tatiana Rykova [1,*], Barış Göktepe [1], Thomas Schierl [1], Konstantin Samouylov [2] and Cornelius Hellge [1]

[1] Video Communication and Applications Department, Fraunhofer Heinrich-Hertz-Institute, Einsteinufer 37, 10587 Berlin, Germany; baris.goektepe@hhi.fraunhofer.de (B.G.); thomas.schierl@hhi.fraunhofer.de (T.S.); cornelius.hellge@hhi.fraunhofer.de (C.H.)

[2] Applied Informatics and Probability Department, Peoples' Friendship University of Russia (RUDN University), Miklukho-Maklaya St. 6, 117198 Moscow, Russia; samuylov-ke@rudn.ru

* Correspondence: tatiana.rykova@hhi.fraunhofer.de

Abstract: In order to fulfill the stringent Ultra-Reliable Low Latency Communication (URLLC) requirements towards Fifth Generation (5G) mobile networks, early-Hybrid Automatic Repeat reQuest (e-HARQ) schemes have been introduced, aimed at providing faster feedback and thus earlier retransmission. The performance of e-HARQ prediction strongly depends on the classification mechanism, data length, threshold value. In this paper, we propose an analytical model that incorporates e-HARQ and Hybrid Automatic Repeat reQuest (HARQ) functionalities in terms of two phases in discrete time. The model implies a fast and accurate way to get the main performance measures, and apply optimization analysis to find the optimal values used in predictor's classification. We employ realistic data for transition probabilities obtained by means of 5G link-level simulations and conduct extensive experimental analysis. The results show that at false positive probability of 10^{-1}, the e-HARQ prediction with the found optimal parameters can achieve around 20% of gain over HARQ at False Negative (FN) of 10^{-1} and around 7.5% at FN of 10^{-3} in terms of a mean spending time before successful delivery.

Keywords: Markov chain model; stationary distribution; performance measures; 5G mobile communication; HARQ; early-HARQ

check for updates

Citation: Rykova, T.; Göktepe, B.; Schierl, T.; Samouylov, K.; Hellge, C. Analytical Model and Feedback Predictor Optimization for Combined Early-HARQ and HARQ. *Mathematics* **2021**, *9*, 2104. https://doi.org/10.3390/math9172104

Academic Editor: Daniel-Ioan Curiac

Received: 28 June 2021
Accepted: 23 August 2021
Published: 31 August 2021

1. Introduction

The Fifth Generation (5G) mobile networks hold enormous potential for innovations across vertical industries given its promised multi-Gbps speed, 1 ms latency, and massive connectivity. Among the three important use cases of 5G systems are: enhanced Mobile BroadBand (eMBB), massive Machine Type Communications (mMTC), and mission critical MTC (cMTC) [1], which include applications with Ultra-Reliable Low Latency Communication (URLLC) requirements with a target error rate of less than 10^{-5} and 1 ms end-to-end latency [2]. The cMTC exemplary scenarios that can only be enabled by URLLC [1]: are factory automation through network controlled systems, autonomous roads, platooning, haptic communications and many others. Despite the huge potential, according to the most recent 5G standard (3GPP Release-16 [3]) low latency and machine-type communication tasks are still in progress, which implies latency efficient and reliable protocols toward URLLC to be a relevant research topic. The researchers in [4] have performed a full-fledged, end-to-end measurement study on one of the first commercial 5G networks, focusing on network coverage, end-to end throughput and delay, according to which there is still a room for improvement, as the radio access network latency reduction between 4G and 5G networks is negligible: 2.19 ± 0.36 ms (5G) vs. 2.6 ± 0.24 ms (4G). Among the techniques standardized in 3GPP Release-15 aimed at achieving the strict requirements of reliability and latency are 5G flexible numerology, downlink pre-emption and uplink configured grant transmission [5,6]. Release 16 has brought a couple of new enhancements in physical layer design for URLLC, including Physical Downlink Control Channel (PDCCH) monitoring capability enhancements, new compact downlink control information format, Hybrid Automatic Repeat reQuest (HARQ) acknowledgement enhancement within a slot, and others [3]. There is plenty of research work that approaches the URLLC communications by enhancing the User Equipment (UE) feedback HARQ mechanism [7], which is a physical level protocol that sends ACK respond to the transmitter in case of successful decode of a packet and NACK in case of an error [8]. Its main limitation is Round TripTime (RTT), which is known to be a time interval between two transmissions. A plenty of research has been done on RTT reduction [9–12]. In [9], researchers suggest reducing the latency by performing short data transmissions that take one Orthogonal Frequency Division Multiplexing (OFDM) symbol period, which in its turn have imposed higher power and processing requirements on both the receiver and the transmitter. Another approach approved in Rel. 16 comprises autonomous transmission of redundant data required to meet the reliability target [10]. However, it might degrade the spectral efficiency by sending unnecessary retransmissions due to the processing and feedback delays. In [11] the optimal and sub-optimal resource allocations are derived for URLLC considering the reliabilities of both data and control channels. The authors, in [12], develop a scheme, which enables a better protection of the NACK signal and is based on the asymmetric feedback signal detection allowing up to one retransmission.

An early HARQ (e-HARQ) prediction—is another approach that has gained much attention recently due to its RTT reduction capabilities since it allows providing the feedback on the decodability of the received signal ahead of the end of the actual transmission process. The prediction is commonly performed by estimating the Bit Error Rate (BER) computed from the Log-Likelihood Ratios (LLRs) of bits with further hard thresholding as classification algorithms [13–15]. The researchers in [13] propose a method for on-line estimation of the BER during a turbo decoding process, in which LLRs are modeled as a mixture of two Gaussian random variables, after which the estimators for the mean and variance of these distributions are derived. In [14,15], prediction schemes in appliance with turbo decoder are presented and investigated from the point of false prediction rates. In [16], the authors introduce the prediction algorithm, which exploits the substructures of the Low-Density Parity-Check (LDPC) codes to start feedback calculation already on partially received codewords, and therefore, provides an excellent opportunity to send ACK/NACK report before the full reception of the codeword. The BER estimations are corrected by including the results of a few LDPC-based decoding operations performed on the partial codewords. In our previous work in [17], we address this problem by proposing an analytical model that studies e-HARQ prediction computed on half of the actual codeword and present comparative analysis of various performance measures. However, we did not investigate the role of a threshold parameter, which is known to be of critical importance for the given e-HARQ schemes. Moreover, by analyzing only the case in which the prediction is performed on half of the codeword, we could not get full picture of the given approach. Other research papers, such as [18,19], investigate ACK/NACK prediction schemes based on channel state estimation.

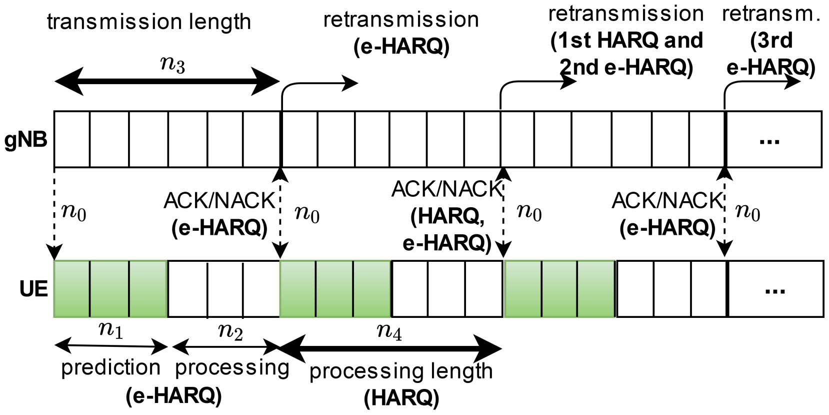
The contribution of this paper is twofold. First, we propose a new two-phase analytical model in discrete time, which evaluates the behavior of a e-HARQ prediction performed for any partially received codeword in a 5G transmission. The model allows obtaining the main performance measures, including the misprediction errors. Second, we formalize the optimization problem to find the optimal values used in a predictor’s classification, including a threshold value, a packet transmission length, and a length of the codeword used for prediction. By performing the substantial comparative analysis for various transmission and prediction lengths, we show the maximum gain that can be achieved in terms of performance measures by using e-HARQ prediction in comparison with the traditional HARQ approach. To the best of our knowledge, this is the first time an analytical model that covers parallel processing of e-HARQ and HARQ mechanisms has been proposed. Note that the given model can be applied also to other e-HARQ prediction schemes besides the LDPC-based one.

The paper is organized as follows: in Section 2 we firstly detail e-HARQ and HARQ mechanisms, introduce the basic model assumptions, and then present a system model with an incorporated e-HARQ and HARQ schemes, derive its balance equations and performance measures. Section 3 presents numerical results, including validation of the model and comparative analysis of schemes for various simulation assumptions. Moreover, we provide results of the optimization analysis, and the best achievable gains that can be found in terms of main performance measures in comparison with traditional HARQ approach without prediction. Finally, Section 4 concludes the paper.

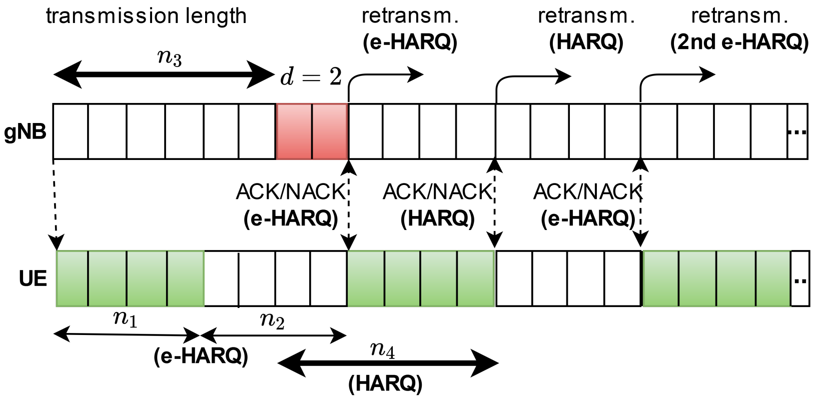
2. System Model

2.1. e-HARQ and HARQ Schemes

An e-HARQ prediction was introduced in order to reduce RTT in stringent URLLC transmissions. The general aim of an e-HARQ scheme is to provide an ACK/NACK feedback on the decodability of the received signal ahead of the end of the actual transmission process. Apparently, the performance of the prediction-based scheme directly depends on the selection of the algorithm that provides the feedback. In [16], the authors introduced the prediction algorithm, which exploits the substructures of the LDPC codes to start feedback calculation already on partially received codewords, and therefore, provides an excellent opportunity to send ACK/NACK report before the full reception of the codeword. These subcodes are formed by selecting a subset of rows and columns from the parity-check matrix. Let us study the main differences between the HARQ scheme and e-HARQ approach that exploits various LDPC codes lengths, shown in Figure 1.



(**a**) Prediction based on 3 OFDM symbols, $d = 0$



(**b**) Prediction based on 4 OFDM symbols, $d = 2$

Figure 1. Time diagram comparison between LDPC-based e-HARQ and HARQ schemes.

Here, we consider an example of a packet transmission length N equal to 6 OFDM symbols. Let us assume n_0 to be a number of propagation time units between the gNodeB (or gNB) and the UE, where a time unit aka microslot corresponds to the duration of an OFDM symbol. The e-HARQ scheme starts prediction process as soon as n_1 OFDM symbols are received at the UE and generates feedback respond in n_2 microslots. Consequently, the gNB will get it in $2n_0 + n_1 + n_2$ time units from the beginning of the transmission, and then the first retransmission may start. To preserve the general terms, we suppose that actual transmission length is n_3 OFDM symbols with the corresponding processing of n_4 microslots. So the retransmission based on HARQ scheme can be performed in $2n_0 + n_3 + n_4$ microslots, as shown in Figure 1.

The HARQ timeline is dominated by the processing time, which may scale with the Transmission Time interval (TTI) length [20], and the propagation time, which is independent of the TTI length. The given scalability principle is taken into consideration by equating $n_1 = n_2$ and $n_3 = n_4$. We suppose propagation time $n_0 = 0$, keeping in mind URLLC as a use case with UEs closely located to gNB. Furthermore, we assume various LDPC subcode lengths, e.g., in Figure 1a $n_1 + n_2 = n_3$, which means that at time moment n_3 we already have a e-HARQ FeedBack (eFB) message at gNB, and may initialize the retransmission process. In Figure 1b prediction is based on 4 OFDM symbols, which corresponds to $n_1 + n_2 = n_3 + d$, where $0 \leq d < N$ is a time offset, which is measured in OFDM symbols. By incrementing d the accuracy of the e-HARQ predictor obviously rises at the cost of a smaller number of possible retransmissions. Note that e-HARQ prediction can be performed on very small subcodes, even less than a half of the transmission length. However, due to the high e-HARQ prediction error, it is out of scope of this work.

2.2. Model Description

Let us consider a downlink 5G network scenario, in which a UE firstly predicts a feedback respond based on the partially received from the gNB codeword using the e-HARQ scheme at the first phase, and then process it with a traditional HARQ scheme at the second phase. We assume packets to be of the same transmission length, equal to N OFDM symbols. Taking into account all retransmissions required for the successful delivery, the packet has to be transmitted to the UE in T_N microslots, otherwise the packet is assumed to be lost.

The functioning of the system is given in discrete time with the length n equal to 1 OFDM symbol. In comparison with our previous work [17], where we studied only one codeword length used for prediction(Figure 1a) and performed analytical analysis on a time slot level equal to an actual transmission length; here, we have to operate on a microslot level of 1 OFDM symbol in order to achieve a fair comparison of the general systems with various N, d values, which substantially complicates the logic of mathematical expressions. We suppose that all the changes in the system occur at time moments $nh, h = 1, 2, \dots$

A packet arrives in the system with the probability $a, 0 < a \leq 1$. If the eFB, generated by the e-HARQ predictor at the first phase, is ACK, then the packet moves to the second phase with the probability $b_1(l)$, where l is a current number of deliveries of the given packet. Otherwise, when the eFB is NACK, the packet is to be retransmitted with the probability $\overline{b}_1(l)$. However, the processing of the previous negatively acknowledged packet based on the e-HARQ predictor is still being done at the second phase with a deterministic probability equal to 1. By doing this, we achieve parallel operation of e-HARQ and HARQ, but keep priority on the HARQ decision in order to avoid unnecessary retransmissions generated by the e-HARQ misprediction. The found regular HARQ FeedBack (rFB) at the second phase allows terminating the delivery process in case of ACK with the probability $b_i(l)$ or continuing the retransmission process in case of NACK with the probability $\overline{b}_i(l), i = 2, 3$. Here and further, $\overline{a, b} := \{a, a+1, \dots, b\}$ with $b > a$ is an interval of natural numbers between a and b. We clarify the difference between

$b_2(l)$ and $b_3(l)$ that can be expressed in terms of the feedback Random Variables (RVs) $\mathrm{eFB}, \mathrm{rFB} \in \{\mathrm{ACK}, \mathrm{NACK}\}$ as,

$$b_2(l) \quad := \quad P(\mathrm{rFB}_l = \mathrm{ACK} | \mathrm{eFB}_l = \mathrm{ACK}), \tag{1}$$

$$b_3(l) \quad := \quad P(\mathrm{rFB}_l = \mathrm{ACK} | \mathrm{eFB}_l = \mathrm{NACK}). \tag{2}$$

By setting the dependence between the feedback generated at the first phase and HARQ's decision at the second phase in terms of $b_2(l)$ and $b_3(l)$, we achieve a system model that can be used for analysis of existing e-HARQ prediction schemes.

If the packet reaches its delay constraint in terms of a maximum number of microslots T_N or a maximum number of transmissions $T_c = \lfloor \frac{T_N}{N} \rfloor$, it is assumed to be lost. It should be noted that a new packet may arrive only when the previous packet has been successfully transmitted or at an empty system. Figure 2 presents the structure of the proposed analytical model.

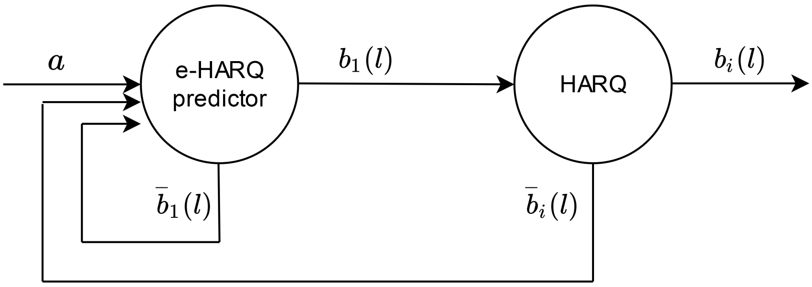


Figure 2. Structure of the model with two phases, corresponding to e-HARQ and HARQ, respectively. Here $i = \overline{2,3}$.

The functioning of the system is described by the homogeneous Markov chain ξ_n at time moments $nh, h = 1, 2\ldots$ with the state space $X = X_m^* \cup X_t$, where X_m^* defines the set of major states, at which the active events in the model take place and which will be further noted with the star, as

$$X_m^* = \{(0,0,0,0)^*, (1,1,1,t)^*, (s,l,v,t)^* : s \in \{1,2,3\}, \\ l = \overline{\lceil \frac{v}{2} \rceil + u(s,v), v-1}, v = \overline{2, T_c}, t \le T_N\}, \tag{3}$$

where v is a time slot number, and

$$u(s,v) = \begin{cases} 1, & \text{if } \{s = 1\} \cap \{v = 2n, n = 2,3,\ldots\}. \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$$t(l,v) = \begin{cases} vN + 2d(l - \lceil \frac{v}{2} \rceil), & \text{if } v = 2n, n = 1,2,\ldots. \\ vN + d(2(l - \lceil \frac{v}{2} \rceil) + 1), & \text{otherwise.} \end{cases} \tag{5}$$

However, since we are considering the model in microslots of the length n equal to a duration of 1 OFDM symbol, we have to define a set of transition states X_t, which allow reaching the major states

$$X_t = \{\{(s-1, l-1, v-1, t-j) : s = 1\}, \{\gamma(l,v)(s, l, v-1, \\ t-i) : s \in \{2,3\}\}, \{\delta(l,v)(s+2, l-1, v-1, t-j) : s \in \{2,3\}\}, \\ (s,l,v,t) \in X_m^*, i = \overline{1, N-d-1}, j = \overline{1, N+d-1}\}. \tag{6}$$

Here,

$$\gamma(l,v) = \begin{cases} 1, & \text{if } \{l < v-1\} \cup \{v = 2\}. \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

$$\delta(l,v) = \begin{cases} 1, & \text{if } \{l > \lceil \frac{v}{2} \rceil, v = 2n\} \cup \{v = 2n+1\}, n = 1,2,\ldots. \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

It can be seen in (3) and (6) that we utilize two parameters for time slot definition: v as a time slot number, which is measured in packet's transmission lengths, and t—a microslot number measured in OFDM symbols. We have to keep track of both of these parameters in order to compose a proper state space with the correct number of transmissions. Otherwise, by leaving out e.g., a time slot number v we may end up in the scenario where the states with the same number of microslots but varying number of transmissions will be erroneously presented as identical state. Moreover, the given system state definition allows clearer differentiation between the major states of the system.

The processing at the phases modeled by the parameter s can be considered in a binary case for better understanding, which is presented for the major state space set X^*_m in Table 1. Here, the position of one in a binary representation means the activity of the given phase.

Table 1. Description of the phase states.

s	Description	States Path
0 (0,0)	Empty system	
1 (1,0)	The partial codeword is being processed at phase 1	$0(N+d-1)$
2 (0,1)	The packet is processed at phase 2 due to ACK at phase 1	$2(N-d-1)$, $4(N+d-1)$
3 (1,1)	The packet is being retransm. At phase 1 because of NACK, and previous negatively acknowledged packet is being processed at phase 2	$3(N-d-1)$ $5(N+d-1)$

The phase states description in Table 1 refers only to major states, in which the actual change in the system takes place. However, the state path clarifies, which and how many transition states should we go through to reach the major ones: e.g., in order to get to $s = 1$, the packet has to go through $N+d-1$ transition 0- states. We draw a graph shown in Figure 3 that clearly demonstrates a whole picture of transition probabilities between the major phase states.

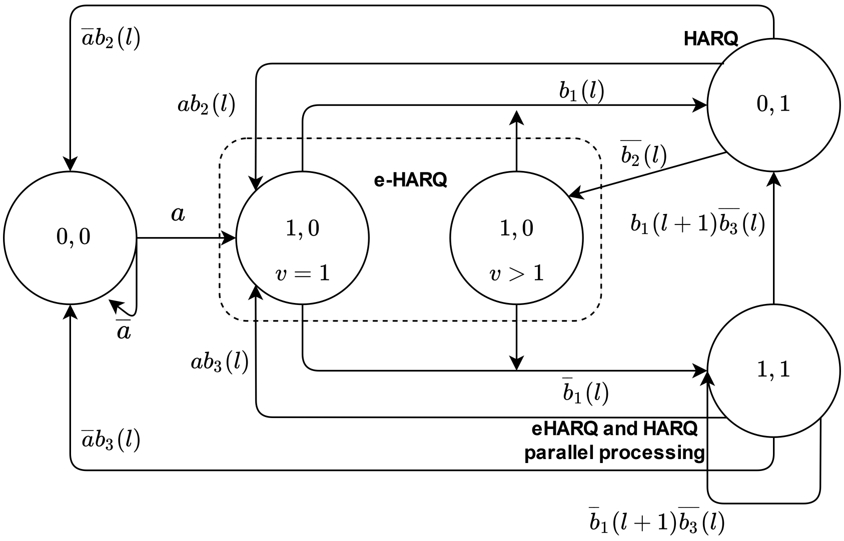


Figure 3. The graph of transition probabilities, taking into account only the major states.

The case $b_1(l) = 1$ corresponds to regular HARQ operation, since no retransmissions based on e-HARQ prediction take place.

2.3. Balance Equations

If $a, 0 < a \leq 1$ the Markov chain ξ_n is aperiodic, and there exists a stationary probability distribution $[\, s, l, v, t]\,, (s, l, v, t) \in X$ which is found from the balance equations:

$$a[0,0,0,0]^* = \overline{a}(\sum_{s=2}^{3} \sum_{(s,l,v,t) \in X_m^* \setminus X_{\text{end}}} b_s(l)[s,l,v,t]^* + \sum_{(s,l,v,t) \in X_{\text{end}}} [s,l,v,t]^*), \tag{9}$$

$$\begin{aligned}[0,0,0,1]^* = a([0,0,0,0] + \sum_{s=2}^{3} \sum_{(s,l,v,t) \in X_m^* \setminus X_{\text{end}}} b_s(l)[s,l,v,t]^* + \\ \sum_{(s,l,v,t) \in X_{\text{end}}} [s,l,v,t]^*),\end{aligned} \tag{10}$$

$$[1,1,1,t]^* = [0,0,0,t-1] = \ldots = [0,0,0,1], \tag{11}$$

$$\begin{aligned}[1,l,v,t]^* = [0,l-1,v-1,t-1] = [0,l-1,v-1,t-2] = \ldots \\ = [0,l-1,v-1,t-(N+d-1)] = \overline{b}_2(l-1) \\ [2,l-1,v-1,t-(N+d)]^*, v = \overline{3,T_c},\end{aligned} \tag{12}$$

$$[s,l,v,t]^* = \gamma(l,v)[s,l,v-1,t-1] + \delta(l,v)[s+2,l-1,v-1,t-1], \tag{13}$$

where $s = \overline{2,3}$ and

$$\begin{aligned}[2,l,v-1,t-1] = [2,l,v-1,t-2] = \ldots = [2,l,v-1,t- \\ (N-d-1)] = b_1(l)[1,l,v-1,t-(N-d)]^*,\end{aligned} \tag{14}$$

$$\begin{aligned}[3,l,v-1,t-1] = [3,l,v-1,t-2] = \ldots = [3,l,v-1,t- \\ (N-d-1)] = \overline{b}_1(l)[1,l,v-1,t-(N-d)]^*,\end{aligned} \tag{15}$$

$$\begin{aligned}[4,l-1,v-1,t-1] = [4,l-1,v-1,t-2] = \ldots \\ = [4,l-1,v-1,t-(N+d-1)] = b_1(l)\overline{b}_3(l-1) \\ \delta(l,v)[3,l-1,v-1,t-(N+d)]^*,\end{aligned} \tag{16}$$

and

$$\begin{aligned}[5,l-1,v-1,t-1] = [5,l-1,v-1,t-2] = \ldots \\ = [5,l-1,v-1,t-(N+d-1)] = \overline{b}_1(l)\overline{b}_3(l-1) \\ \delta(l,v)[3,l-1,v-1,t-(N+d)]^*,\end{aligned} \tag{17}$$

where $(s,l,v,t) \in X$.

$$X_{\text{end}} = X_{\text{end}}(1) \cup X_{\text{end}}(2) \cup X_{\text{end}}(3) \subset X_m^*, \tag{18}$$

presents a set of termination system states, from which we have to leave the system due to the reaching the T_N limit, where

$$X_{\text{end}}(1) = \{\mu(t(l,v+1) - T_N) \cdot (1,l,v,t(l,v))^*\}, \tag{19}$$

$$X_{\text{end}}(2) = \{\mu(t(l+1,v+1) - T_N) \cdot (2,l,v,t(l,v))^*\}, \tag{20}$$

$$X_{\text{end}}(3) = \{\mu(t(l+1,v+1) - T_N) \cdot (3,l,v,t(l,v))^*\}, \tag{21}$$

$$\mu(x) = \begin{cases} 1, & \text{if } x > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

The normalizing equation is defined as

$$\sum_{(s,l,v,t)\in X} [s,l,v,t] = 1. \tag{23}$$

One of the typical challenges of the systems that model technical systems in discrete time is to properly define the state space. In its turn, the cardinality of the state space has a direct impact on the feasibility of finding the stationary probability distribution. The cardinality for the given model's state space is equal to $\mid X \mid = \mid X_m^* \mid + \mid X_t \mid$, where $\mid X_m^* \mid$—is an overall number of all the major states, and is defined as

$$\begin{aligned} \mid X_m^* \mid = \mid & [0,0,0,0] + [1,1,1,t(1,1)] + \\ & \sum_{v=3}^{\lfloor \frac{T_N}{N} \rfloor} \sum_{l=\lceil \frac{v}{2} \rceil + u(1,v)}^{v-1} (1 - \mu(t(l,v+1) - T_N)) * [1,l,v,t(l,v)] + \\ & \sum_{s=2}^{3} \sum_{v=2}^{\lfloor \frac{T_N}{N} \rfloor} \sum_{l=\lceil \frac{v}{2} \rceil}^{v-1} (1 - \mu(t(l+1,v+1) - T_N))[s,l,v,t(l,v)] \mid, \end{aligned} \tag{24}$$

and $\mid X_t \mid$—is an overall number of all the transition states, and is computed as

$$\begin{aligned} \mid X_t \mid = \mid & \sum_{(1,l,v,t)\in X_m^*} \sum_{j=1}^{N+d-1} [0,l-1,v-1,t-j] + \\ & \sum_{s=2}^{3} \sum_{(s,l,v,t)\in X_m^*} \sum_{j=1}^{N-d-1} \gamma(l,v)[s,l,v-1,t-i] + \\ & \sum_{s=2}^{3} \sum_{(s,l,v,t)\in X_m^*} \sum_{j=1}^{N+d-1} \delta(l,v)[s+2,l-1,v-1,t-j] \mid . \end{aligned} \tag{25}$$

For example, for $N = 4, d = 1$ and $T_N = 28$ the cardinality set is equal to 122 states, out of which 26 belong to the major ones, and 96 states to the transition ones.

2.4. Performance Measures

The stationary probability distribution $[s,l,v,t], (s,l,v,t) \in X$ is computed from the balance equations along with the normalizing condition by solving a set of linear equations. In order to derive the main performance measures, we firstly define X_{quit} as the set of states, in which a packet can quit the system either successfully or unsuccessfully. Hence, it is defined as

$$\begin{aligned} X_{\text{quit}} := \{ & (s,l,v,t)^* : s \in \{2,3\}, l = \overline{\lceil \frac{v}{2} \rceil, v-1}, v = \overline{2, T_c}, \\ & t \leq T_N\} \cup \{(1,l,T_c,t) : l = \overline{\lceil \frac{T_c}{2} \rceil + u(1,T_c), T_c - 1}, t \leq T_N\}. \end{aligned} \tag{26}$$

Then a blocking probability, which is known to be a probability of a packet not being serviced during the predefined period T_N, is given as

$$\pi = \sum_{x \in X_{\text{quit}}} P_{\text{quit}}(x)(1 - B(x)), \tag{27}$$

where $B(x), x \in X_{\text{quit}}$ is defined as

$$B(x) := \begin{cases} 0, & \text{if } s_x = 1. \\ b_2(l_x), & \text{if } s_x = 2 \text{ and } x \in X_{\text{end}}(2). \\ b_3(l_x), & \text{if } s_x = 3 \text{ and } x \in X_{\text{end}}(3). \\ 1, & \text{otherwise.} \end{cases} \tag{28}$$

$P_{\text{quit}} : X_{\text{quit}} \to [0,1]$—is a conditional probability that a packet no matter how, successfully or not, quits the system at state $x, x \in X_{\text{quit}}$:

$$P_{\text{quit}}(x) := \frac{b_{\text{quit}}(x)[x]}{\sum_{x \in X_{\text{quit}}} b_{\text{quit}}(x)[x]}, \tag{29}$$

where

$$b_{\text{quit}}(x) := \begin{cases} b_2(l_x), & \text{if } s_x = 2 \text{ and } x \notin X_{\text{end}}. \\ b_3(l_x), & \text{if } s_x = 3 \text{ and } x \notin X_{\text{end}}. \\ 1, & \text{otherwise.} \end{cases} \tag{30}$$

The mean number of deliveries is calculated analogously

$$M = \sum_{x \in X_{\text{quit}}} P_{\text{quit}}(x) L(x), \tag{31}$$

where $L(x)$ the number of transmissions at state x is presented as

$$L(x) := \begin{cases} l_x - 1 & \text{if } s_x = 1. \\ l_x & \text{if } s_x = 2 \text{ or } s_x = 3. \end{cases} \tag{32}$$

In its turn, the mean spending time of a packet in the system, which has been successfully transmitted, is calculated as

$$T = \sum_{x \in X_{\text{quit}}} P_{\text{success}}(x) v_x, \tag{33}$$

where $P_{\text{success}}(x) : X_{\text{quit}} \to [0,1]$ is defined as $P_{\text{quit}}(x)$ with b_{quit} being replaced by b_{success}:

$$b_{\text{success}}(x) := \begin{cases} b_2(l_x), & \text{if } s_x = 2. \\ b_3(l_x), & \text{if } s_x = 3. \\ 0, & \text{otherwise.} \end{cases} \tag{34}$$

The probability of the system being idle is:

$$P_0 = [0,0,0,0]. \tag{35}$$

The e-HARQ prediction schemes are characterized by two types of errors: False Negative (FN) errors (e-HARQ predicts NACK, whereas HARQ sends ACK), and False Positive (FP) ones (e-HARQ predicts ACK, whereas HARQ sends NACK respond). FN decisions lead to redundant transmissions, which may result in spectral efficiency degradation but have no impact on latency and Block Error Rate (BLER). Hence, FN errors can be tolerated up to a certain limit. In its turn, FP errors indicate predictor's failure. As was already mentioned, common e-HARQ predictors use a threshold value θ to classify between ACK and NACK. Obviously, the selection of θ has a direct relation to performance of the e-HARQ mechanism. The FN probability of an e-HARQ predictor can be found as

$$P_{FN} = \sum_{v=2}^{T_c} \sum_{l=\lceil \frac{v}{2} \rceil}^{v-1} b_3(l)[3,l,v,t], \tag{36}$$

whereas the FP probability of an e-HARQ predictor is computed as

$$P_{FP} = \sum_{v=2}^{T_c} \sum_{l=\lceil \frac{v}{2} \rceil}^{v-1} \overline{b}_2(l)[2,l,v,t]. \tag{37}$$

Figure 4 demonstrates a block diagram of a studied communication system on a high-level perspective, including e-HARQ prediction, HARQ and configuration of the e-HARQ prediction based on the proposed analytical model. As input parameters, the model gets the transition probabilities $b_i, i = \overline{1,3}$ estimated from the e-HARQ prediction and HARQ mechanisms, and tunes the parameters of e-HARQ prediction by solving the optimization problem further described in Section 3. For the purpose of consistency, we have estimated the complexity of the model's stationary distribution computation, which is of the order $O(NT_c log T_c)$. The given complexity change corresponds to the worst-case scenario, when the analytical model together with e-HARQ is not able to reduce the amount of retransmissions based on the found optimal parameters. However, it has to be noted that e-HARQ configuration based on the proposed analytical model can be performed only once prior to the actual transmission, which obviously does not bring complexity changes to the actual transmission process demonstrated in the structure diagram above.

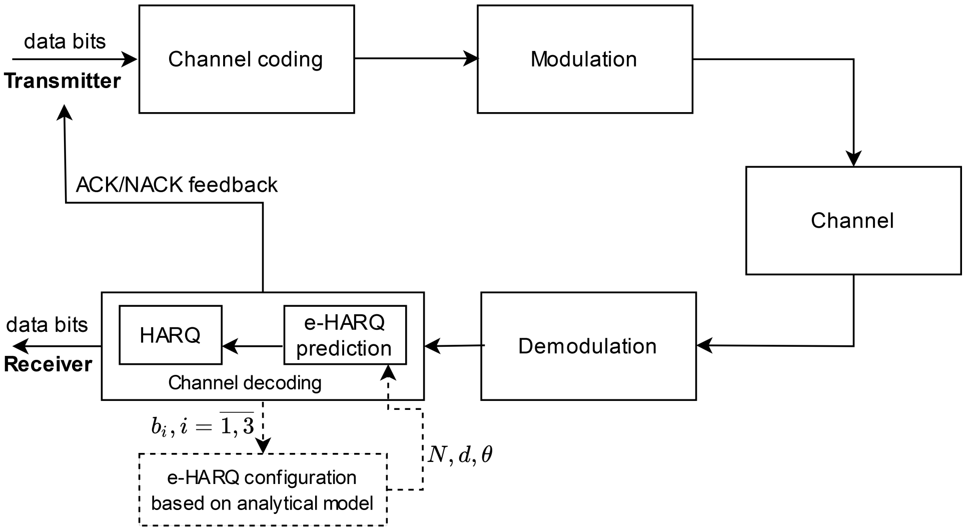


Figure 4. Block diagram of a considered communication system on a high-level perspective.

3. Experimental Analysis

3.1. Link-Level Simulation Setup

We perform experimental analysis with the transition probabilities, i.e., $b_1(l)$, $b_2(l)$ and $b_3(l)$ extracted from 5G-compliant Link-Level Simulations (LLS). The simulation parameters are summarized in Table 2. A packet of 500 bits have been encoded using 5G-compliant LDPC codes and mapped to an 1.08 MHz OFDM transmission using a 64-QAM modulation. The encoded signal has been sent over a TDL-C fading channel and processed at the receiver using frequency domain MMSE and an LDPC min-sum decoder. We use a logistic regressions (LR) predictor as an e-HARQ mechanism, where the log-odds of a binary output variable are modelled as a linear combination of the classifier's input variable [21]. At smaller $b_1(l)$ LR predictor operates more conservatively with sending ACKs whereas at high $b_1(l)$ it results in the increase of FP probability. Monte-Carlo simulations with 1.7 M iterations have been performed to estimate the probabilities, as described below. We varied the SNR between 5 dB and 12 dB and considered various subcode lengths between 1/2 and 5/6. The transition probabilities have been approximated as

$$b_2(l) = P(\mathrm{rFB}_l = \mathrm{ACK}|\mathrm{eFB}_l = \mathrm{ACK}, \mathrm{rFB}_{l-1} = \mathrm{NACK}) \approx 1 - \frac{P(\mathrm{rFB}_l = \mathrm{NACK}|\mathrm{eFB}_l = \mathrm{ACK})}{P(\mathrm{rFB}_{l-1} = \mathrm{NACK}|\mathrm{eFB}_{l-1} = \mathrm{ACK})}, \quad (38)$$

$$b_3(l) = P(\mathrm{rFB}_l = \mathrm{ACK}|\mathrm{eFB}_l = \mathrm{NACK}, \mathrm{rFB}_{l-1} = \mathrm{NACK}) \approx 1 - \frac{P(\mathrm{rFB}_l = \mathrm{NACK}|\mathrm{eFB}_l = \mathrm{NACK})}{P(\mathrm{rFB}_{l-1} = \mathrm{NACK}|\mathrm{eFB}_{l-1} = \mathrm{NACK})}. \quad (39)$$

The $b_1(l)$ values have been initialized to the corresponding decoding probabilities for $\theta = 0$ defined as

$$b_1(l) := P(\mathrm{rFB}_l = \mathrm{ACK}|\mathrm{rFB}_{l-1} = \mathrm{NACK}), \quad (40)$$

and have been increased by a step size of 0.1 for each l, respectively. Furthermore, the transition probabilities $b_1(l)$ have been capped at 1. Therefore, resulting to $b_1(l) = 1$ for all l values at threshold $\theta = 9$.

Figure 5 demonstrates an example of transition probabilities extracted from 5G-compliant link-level simulations with SNR = 10 dB and a packet transmission length of 14 OFDM symbols, which we further utilize to check the accuracy of the analytical model in Section 3.2. The plots show the curves for the first transmission only and encompass behavior for all subcode lengths: $d = \overline{0, 12}$ used for e-HARQ prediction. The threshold values θ correspond to the monotonous increase from 0 to 1 in terms of b_1, where 0 corresponds to permanent NACK generated by e-HARQ prediction, and 1—no prediction at all. When performing prediction based on a larger subcode length, i.e., $d = 12$, we can see that b_1 rises, starting from 0.5 due to its higher accuracy and the given quality of the channel.

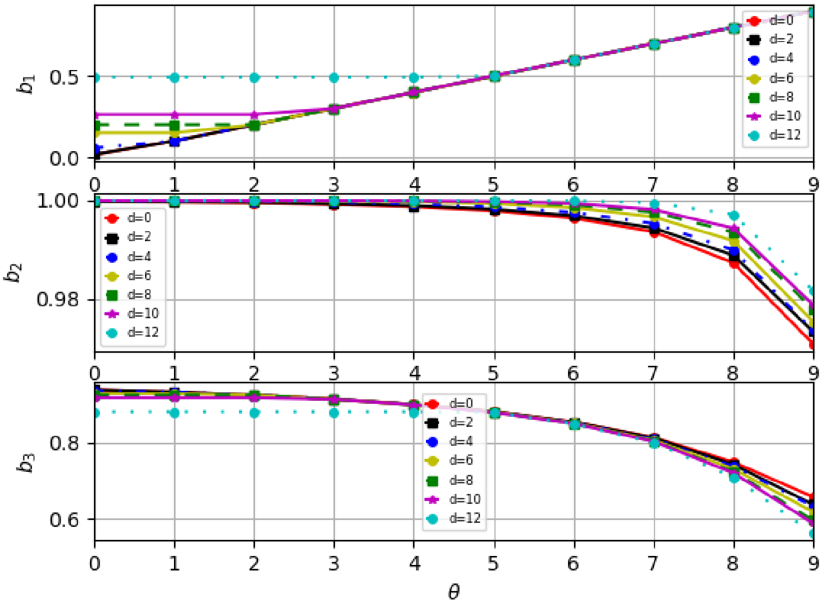


Figure 5. Transition probabilities extracted from the 5G-compliant link-level simulations.

Table 2. LLS assumptions for training and test set generation.

Transport block size (N_{Bits})	500
Transmission bandwidth	1.08 MHz (6 RBs)
Channel Code	Rate-1/5 LDPC [22]
Modulation order and algorithm	64-QAM, Approx. LLR
Power allocation	Constant E_b/N_0
Waveform	3GPP OFDM, normal cyclic-prefix, 15 kHz spacing
Channel type	1 Tx 1 Rx, TDL-C 100 ns, 7 GHz, 3 km/h
Equalizer	Freq. domain MMSE
SNR	5.0 dB:12.0 dB
Decoder type	Min-Sum (50 iter.)
Predictor type	Logistic regressions [23] (5 iter.)

3.2. Validation of the Model

We start experimental analysis by verifying the accuracy of the proposed analytical model. Since the matrix of transition probabilities due to the big cardinality of the state space is quite complicated for clear verification, we have conducted an event-based imitation analysis, in which we have simulated multiple times the packet transmission flow through the states of the model until the probability distribution converges. The given imitation model allows obtaining the probability distribution along with the performance measures based on the collected statistical data. We utilize the transition probabilities demonstrated in Figure 5 and do not differentiate them in terms of l number of deliveries. The arrival rate is considered to be 0.2 for a transmission length of 14 OFDM symbols given the fact that modelling is done on a microslot level, which corresponds to the use cases that experience a transmission flow of packets. The mean spending time measured in time slots for a set of threshold values θ and all possible d for both analytical and imitation models is shown in Figure 6. The imitation of the given generalized model with a microslot processing is very computationally intensive and requires much more computational efforts until the system converges compared to the given analytical modelling, where the required performance measures are computed right away. Due to this, we have provided only a few experimental points for comparative analysis between two types of modelling; however, it is enough to verify the correctness of the matrix of transition probabilities.

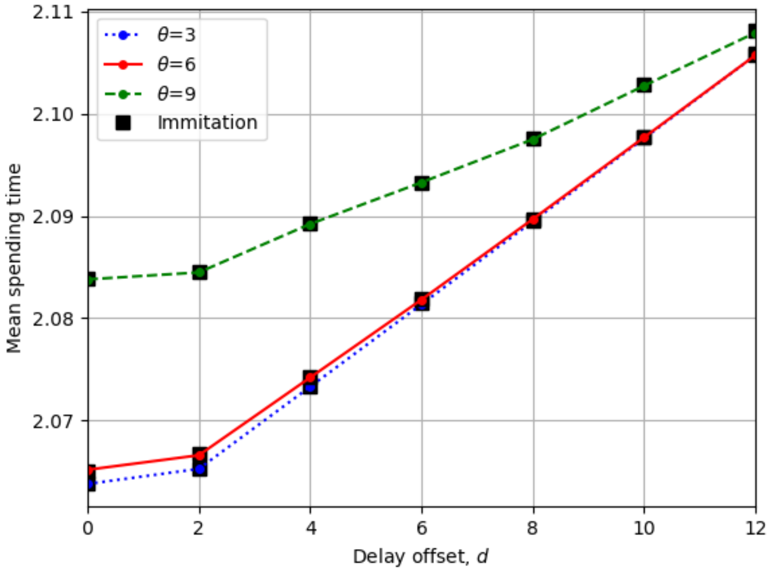


Figure 6. Analytical and imitation models comparison in terms of mean spending time.

3.3. Analysis of Model's Tendencies

We have conducted an extensive comparative analysis of the given analytical model for various SNR values, transmission lengths N, subcode lengths with time shift d and threshold values θ. For the experiments in this subsection, we have fixed SNR = 7.0 dB. The maximum transmission time in all the experiments $T_N = 84$ microslots, or 6 ms, so each of the transmission length scenarios may achieve its own maximum number of transmissions $T_c = \lfloor T_N/N \rfloor$. Figures 7 and 8 demonstrate probability of a system being idle and mean spending time for family of curves with various transmission lengths. The legend is consisted of two parameters: N(indicated by color), and d (indicated by marker), respectively, which is defined for a specific curve at the plot. The probability of a system being idle is one of the main characteristics that shows how fast the packet can be transmitted and the system goes back in idle state until the next packet arrives. The higher is the probability, the more efficiently the system copes with the packet transmission. In Figure 7 we can see that the system with the smallest transmission length and fastest feedback exchange $d = 0$ shows the best behavior with a slight degradation when increasing d. In general, the given tendency can be visualized for all transmission lengths, which converges to the scenario without e-HARQ prediction ($\Theta = 9$). In Figure 8 we can see the mean time spending of packet measured in microslots. Indeed, smaller transmission lengths encompasses a higher number of possible transmissions in the given time interval, which results in a faster successful delivery process given the quality of the channel. It is interesting to mention that at small SNRs, the systems with transmission length N and $d = 0$ outperforms the ones with $N - 1$ and maximum d. This tendency fades away with the improvement of network conditions, e.g., at SNR = 10 dB the curves show the similar behavior, however do not intersect.

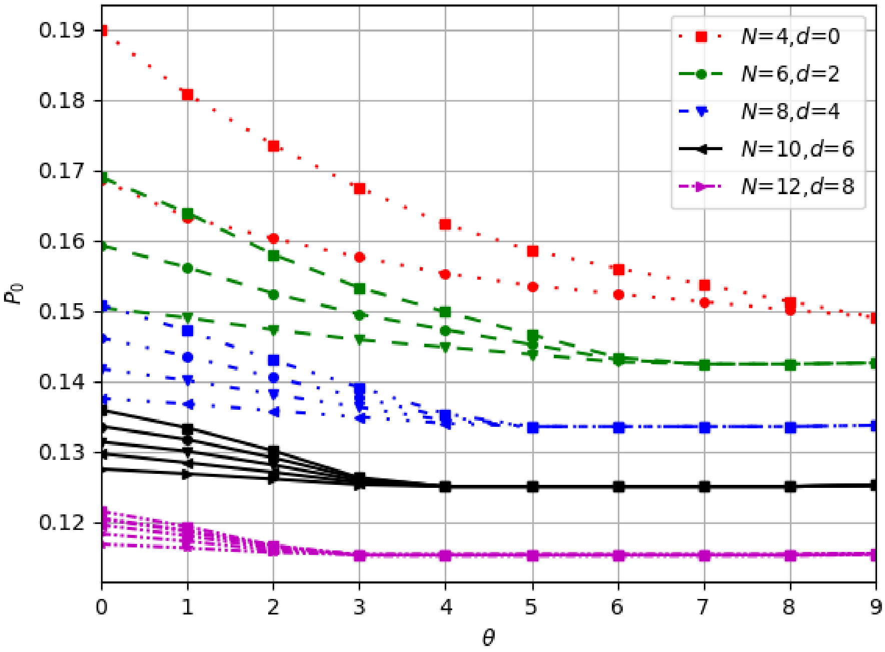


Figure 7. Probability of a system being idle for families of curves with two parameters: N (indicated by color), and d (indicated by marker).

Figure 9 gives us a clear view on prediction errors, such as false negative and false positive ones for $d = 0$. Larger transmission lengths obviously come along with higher FN probabilities in comparison with small N-systems at low θ values, whereas the false positive probabilities have an opposite behavior with smaller lengths deliveries, characterized by the highest prediction error at bigger θ. The given performance measures show us the tendencies that we exploit further when finding optimized parameters for e-HARQ prediction.

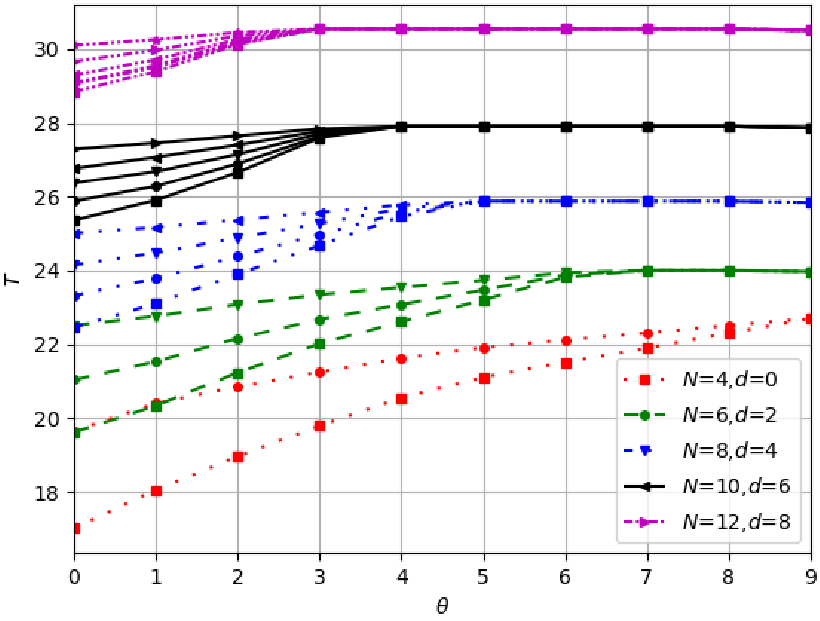


Figure 8. Mean spending time for families of curves with two parameters: N (indicated by color), and d (indicated by marker).

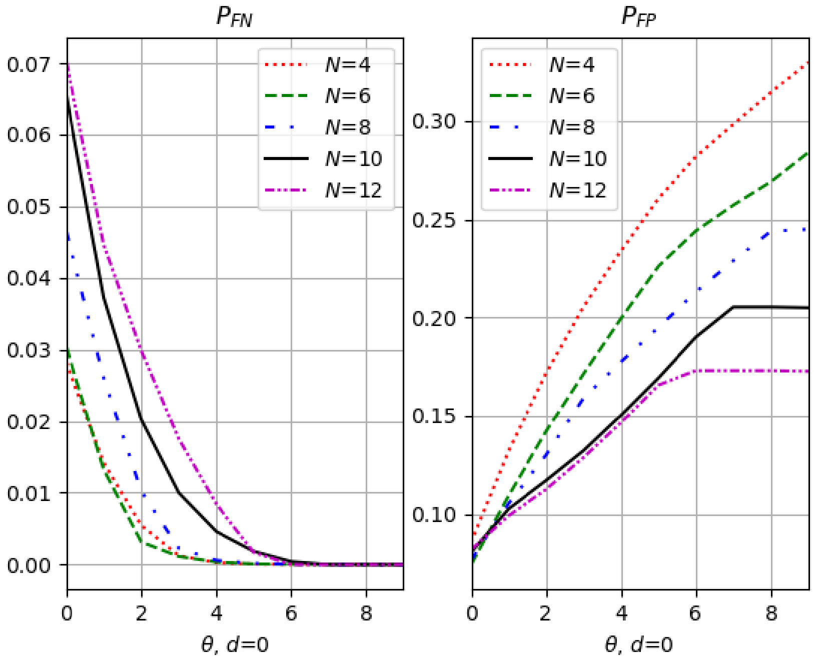


Figure 9. e-HARQ prediction errors: false negative probability and false positive probability from left to right, respectively.

3.4. Optimization Analysis

One of the main goals of the given generalized analytical model is to get a quick estimate of the performance measures in order to select a threshold value for e-HARQ prediction operation. Keeping in mind the stringent URLLC latency requirements, we might favor systems with the smallest transmission and subcode lengths given the tendencies above, however, we have to take into account false e-HARQ predictions. As was mentioned above, false-positive mispredictions increase the total BLER, whereas false-negative errors mainly cause unnecessary retransmissions, and hence degrade the spectral efficiency. Therefore, to find an optimal threshold parameter, the following cost function has to be minimized:

$$\min_{\theta}(w_1 P_{FN}(\theta) + w_2 T^*(\theta)), \tag{41}$$

where w_1 and w_2 are weight factors, $w_1 + w_2 = 1$ and $T^* = \frac{T}{T_N}$—is a normalized mean spending time until successful delivery.

Figure 10 shows dependence between the min cost function and the selected weight factor w_1 with a corresponding threshold value θ from left to right, respectively. It demonstrates in general higher min cost at lower w_1, which can be explained by the fact, that the normalized mean transmission time is significantly larger P_{FN}. So we see a reasonable monotonous decline of the min cost function for a family of curves N, d when incrementing the weight w_1 of $P_{FN}(\theta)$. This is consistent with expectations in the sense that this decline comes along with the rise of the threshold parameter θ as shown in Figure 10 (right), which leads to an unfavorable behavior in terms of the main performance measures (Figures 7 and 8). Therefore, given the target network characteristics, the weighted factor can be tuned; and, as can be seen in Figure 11, which shows cost functions with different weights and SNR values, it might be reasonable to take bigger weight value w_1, e.g., $w = [0.8, 0.2]$ in Figure 11a in worse network conditions to combat with higher FN probability. We also visualize a shift of local minimum to the left for lower weight w_1 values over a set of SNRs. The cost function in Figure 11d has a strict minimum at $\theta = 1$ due to quite a rapid decline of FN probability to zero, and further shifts the minimum fully to the right at $\theta = 0$ in better network conditions. It is worth mentioning that with the increase of subcode length, the cost function also rises.

Since we do not know what the actual cost function is, which is usually determined by the physical constraints, we try to find out here the way the system behaves under a range of cost functions. It should be noted that instead of a normalized mean spending time, other performance measures, e.g., blocking probability, could be included into the cost function evaluation. However, for the given initialization parameters, the blocking probability has shown insignificant variation over the whole range of θ, and therefore is not of practical interest. We further compare the performance measures and find the maximum gain achieved in comparison with the traditional HARQ approach given the constrained FN and FP probabilities, by formalizing and solving the given optimization problem:

$$\max_{N,d,\theta}(G(T(N,d,\theta),T_{HARQ}(N,d,\theta))), \tag{42}$$

with the constraints $P_{FN}(N,d,\theta) < FN_{\max}$ and $P_{FP}(N,d,\theta) < FP_{\max}$. Here,

$$G(x,y) = \frac{\mid x-y \mid}{y} * 100\%, \tag{43}$$

and $FN_{\max}, FP_{\max}$—are the maximum possible values of false negative and false positive probabilities, correspondingly. This approach allows us selecting optimal values N,d,θ for operation in a flexible 5G environment.The optimization problem is solved using direct search.

Figure 12 demonstrates the maximum achievable gain in percentage in terms of a mean spending time over the traditional HARQ approach for optimized parameters N,d,θ. It shows the gain curves for a number of FN limits and SNR equal to 12 dB. It is worth stressing that the higher is the $FN_{\max}$ probability, the more exponential is the rise of the gain curve over $FP_{\max}$. It can be seen that by constraining the FP probability to 0.15, we can achieve 20% of gain at FN = 0.1 and around 11% at FN = 0.001 in terms of a mean number of microslots required for successful delivery. Table 3 shows the gain in terms of a probability of a system being idle for a number of SNRs and $FN_{\max}$ values given a number of constrained $FP_{\max}$ values. In general, we see here a qualitatively consistent picture. By relaxing the $FN_{\max}$ constraint, the gain rises for all the cases of the constrained $FP_{\max}$.

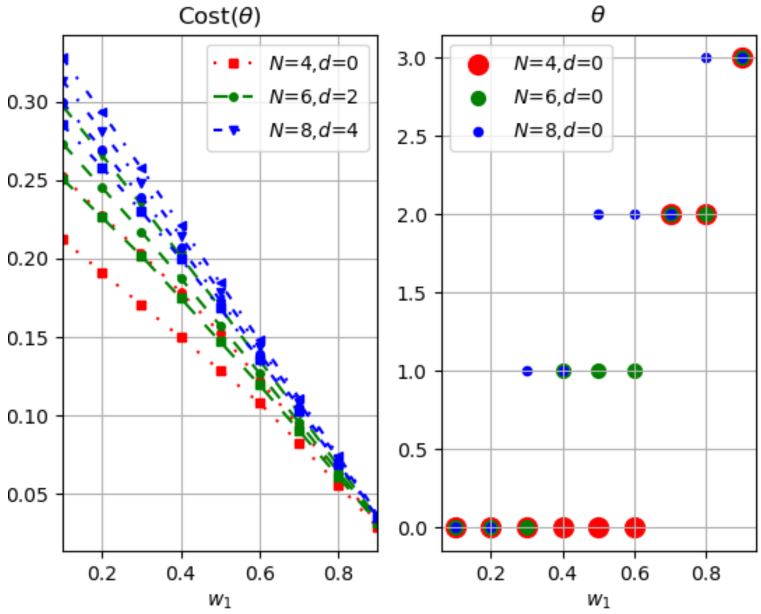
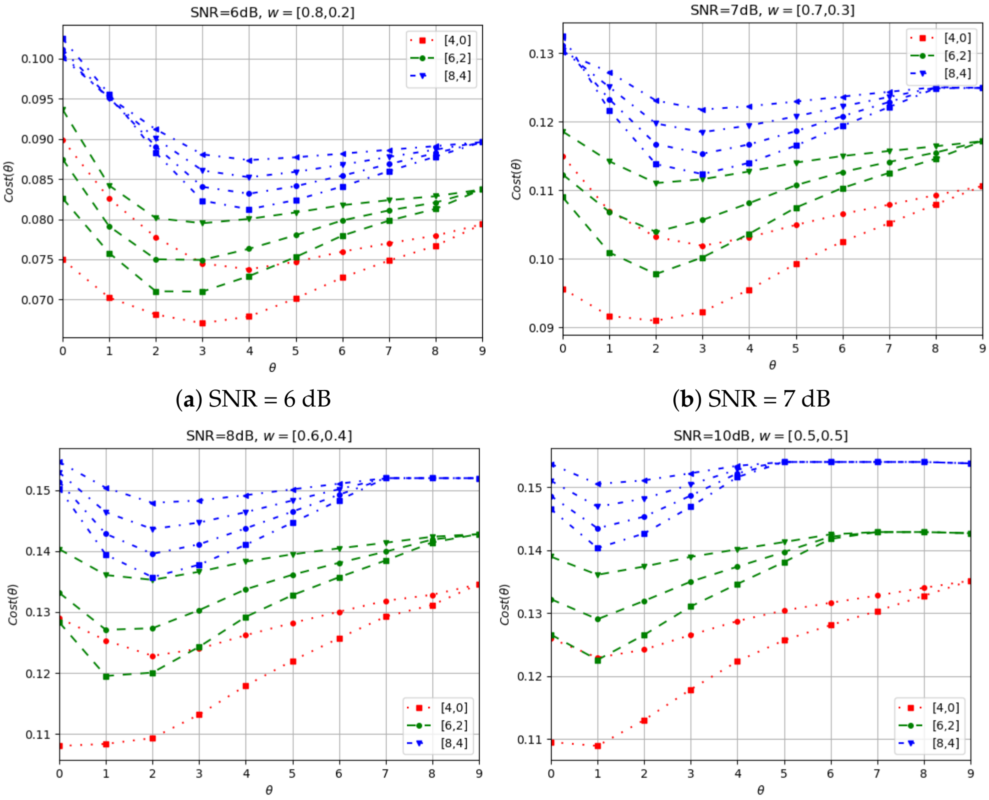


Figure 10. Min cost function versus weight factor, and corresponding θ for SNR = 8, $[N,d]$ from left to right, respectively.



(**a**) SNR = 6 dB (**b**) SNR = 7 dB

(**c**) SNR = 8 dB (**d**) SNR = 10 dB

Figure 11. Cost functions for various N (indicated by color), d (indicated by marker): (**a**) $w = [0.8, 0.2]$, (**b**) $w = [0.7, 0.3]$, (**c**) $w = [0.6, 0.4]$, (**d**) $w = [0.5, 0.5]$.

It is worth mentioning that at small $FN_{\max}$ values, e.g., $FN_{\max} = 10^{-4}$ the highest gain equal to 12.36% is achieved in network scenario of 10 dB, and in general rises in better SNRs up to a certain limit, after which it tends to lower down, which can be explained by the higher accuracy of the predictor and less erroneous transmission process. The latter behavior can be also visualized when relaxing the conditions of $FN_{\max}$. Also, the better channel conditions are, the faster the gain curve reaches its maximum. All in all, these results clearly demonstrate the positive gain of the e-HARQ prediction operation over the whole range of the FN and FP probabilities. In this paper we investigated a downlink network scenario with only one UE in terms of analytical modelling. In a general case, we assume that the gNB schedules the resources among the UEs without collisions, which brings us to consideration that the collective gain scales up given the number of users.

Table 3. Gain in % in terms of a probability of a system being idle.

	$FP_{\max} = 0.05$				$FP_{\max} = 0.1$				$FP_{\max} = 0.2$			
	$FN_{\max}$				$FN_{\max}$				$FN_{\max}$			
SNR	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
5	1.32	4.41	5.23	6.17	1.60	5.62	8.17	40.80	2.48	9.52	28.51	40.80
6	1.12	2.51	5.68	6.86	1.97	3.58	7.8	39.6	3.23	9.3	26.3	39.67
7	1.27	3.12	5.72	6.76	1.98	4.65	15.04	39.09	4.75	9.89	23.48	39.09
8	0.65	2.43	4.62	5.81	2.45	5.2	21.58	37.43	5.77	8.41	26.43	37.43
9	0.55	1.67	3.38	6.02	2.87	8.78	18.07	32.33	6.4	18.70	23.70	32.33
10	0.3	1.47	2.87	2.87	3.3	7.51	21.3	27.37	12.36	16.46	21.3	27.37
11	0.71	1.42	2.08	9.85	4.73	8.10	18.14	22.78	9.84	13.63	18.14	22.78
12	0.68	2.29	7.62	7.62	5.32	7.15	15.51	19.53	8.35	11.40	15.51	19.53

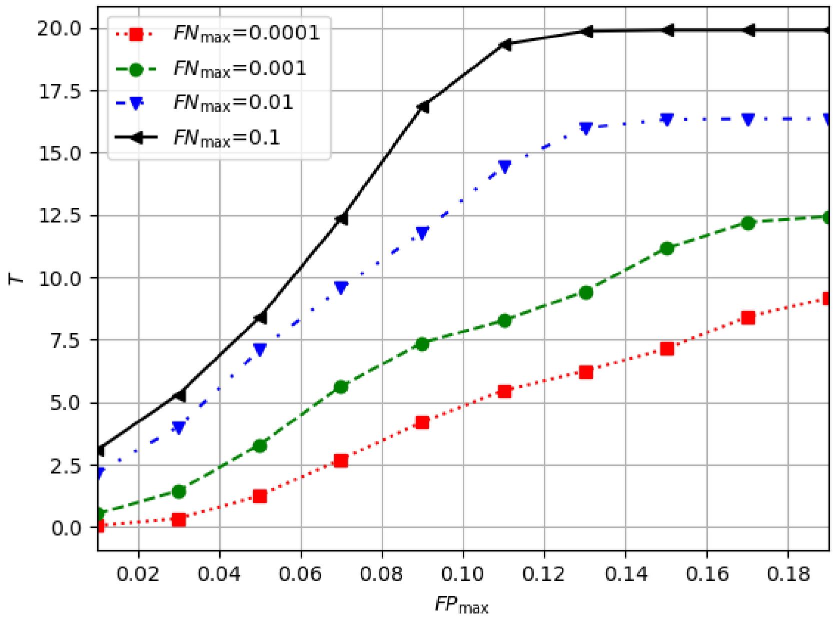


Figure 12. Mean spending time gain [%].

4. Discussion

In this paper, we have proposed an analytical model that can be used to evaluate the behavior of an e-HARQ prediction based on LDPC subcodes of various lengths and other LLR-based predictors used in 5G transmissions. The model covers parallel processes of e-HARQ and HARQ, and clearly demonstrates the benefit of the combined e-HARQ and HARQ mechanism in comparison with the traditional HARQ in terms of main performance measures, e.g., by constraining the FP and FN probabilities to 10^{-1}, we can achieve around 20% and 7.5% of gain at FN = 10^{-3} in terms of a mean spending time until successful delivery. It also provides an efficient and fast tool to analyze the way the system behaves under a range of cost functions aimed at finding optimal parameters for an e-HARQ prediction operation. The given analytical model in compliance with the optimization analysis can be efficiently used by network designers to get a quick estimate of the threshold value used for classification when designing an e-HARQ mechanism without the need to spend considerable time for simulation studies or expensive experimental setup. This work does not take into account scheduling effects due to the presence of multiple users in the network, which is currently under investigation.

Author Contributions: Conceptualization, C.H., B.G., K.S., T.S. and T.R.; methodology, T.R.; software, T.R., B.G.; validation, T.R., B.G. and C.H.; formal analysis, T.R.; investigation, T.R.; writing—original draft preparation, T.R.; writing—review and editing, B.G., T.R.; visualization, T.R.; supervision, C.H. and K.S.; project administration, C.H.; funding acquisition, C.H. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors are sincerely grateful to the reviewers for their helpful comments, which have significantly improved the quality of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tullberg, H.; Popovski, P.; Li, Z.; Uusitalo, M.A.; Hoglund, A.; Bulakci, O.; Fallgren, M.; Monserrat, J.F. The METIS 5G System Concept: Meeting the 5G Requirements. *IEEE Commun. Mag.* **2016**, *54*, 132–139. [CrossRef]
2. Bertenyi, B.; Nagata, S.; Kooropaty, H.; Zhou, X.; Chen, W.; Kim, Y.; Dai, X.; Xu, X. 5G NR Radio Interface. *J. ICT Stand.* **2018**, *6*, 31–58. [CrossRef]
3. 3GPP. 3GPP Release 16. Technical Report. 2020. Available online: https://www.3gpp.org/release-16 (accessed on 20 June 2021).

4. Xu, D.; Zhou, A.; Zhang, X.; Wang, G.; Liu, X.; An, C.; Shi, Y.; Liu, L.; Ma, H. Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption. In Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication, Online, 10–14 August 2020; ACM: New York, NY, USA, 2020; pp. 479–494. [CrossRef]
5. 3GPP TS 38.211. Physical Channels and Modulation. v15.7.0. Available online: https://www.3gpp.org/ftp//Specs/archive/38_series/38.211/38211-f70.zip (accessed on 20 June 2021).
6. 3GPP TS 38.214. Physical Layer Procedures for Data. v15.7.0. Available online: https://www.3gpp.org/ftp//Specs/archive/38_series/38.214/38214-f70.zip (accessed on 20 June 2021).
7. Nokia Shanghai Bell. Enhanced Industrial Internet of Things (IoT) and URLLC Support. Technical Report RP-193233, 3GPP. 2019. Available online: https://www.3gpp.org/ftp/TSG_RAN/TSG_RAN/TSGR_86/Docs/RP-193233.zip (accessed on 18 June 2021).
8. Frederiksen, F.; Kolding, T.E. Performance and modeling of WCDMA/HSDPA transmission/H-ARQ schemes. In Proceedings of the IEEE 56th Vehicular Technology Conference, Vancouver, BC, Canada, 24–28 September 2002; Volume 1, pp. 472–476. [CrossRef]
9. Mahmood, N.H.; Abreu, R.; Bohnke, R.; Schubert, M.; Berardinelli, G.; Jacobsen, T.H. Uplink Grant-Free Access Solutions for URLLC services in 5G New Radio. In Proceedings of the 2019 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, Finland, 27–30 August 2019; pp. 607–612. [CrossRef]
10. Liu, Y.; Deng, Y.; Elkashlan, M.; Nallanathan, A.; Karagiannidis, G.K. Analyzing Grant-Free Access for URLLC Service. *arXiv* **2020**, arXiv:2002.07842.
11. Shariatmadari, H.; Duan, R.; Iraji, S.; Li, Z.; Uusitalo, M.; Jäntti, R. Resource Allocations for Ultra-Reliable Low-Latency Communications. *Int. J. Wirel. Inf. Netw.* **2017**, *24*, 317–327. [CrossRef]
12. Shariatmadari, H.; Duan, R.; Iraji, S.; Jäntti, R.; Li, Z.; Uusitalo, M.A. Asymmetric ACK/NACK Detection for Ultra—Reliable Low—Latency Communications. In Proceedings of the 2018 European Conference on Networks and Communications (EuCNC), Ljubljana, Slovenia, 18–21 June 2018; pp. 1–166. [CrossRef]
13. Letzepis, N.; Grant, A. Bit Error Estimation for Turbo Decoding. In Proceedings of the IEEE International Symposium on Information Theory, Yokohama, Japan, 29 June–4 July 2003. [CrossRef]
14. Berardinelli, G.; Khosravirad, S.R.; Pedersen, K.I.; Frederiksen, F.; Mogensen, P. Enabling Early HARQ Feedback in 5G Networks. In Proceedings of the 83rd IEEE Vehicular Technology Conference (VTC Spring), Nanjing, China, 15–18 May 2016; pp. 1–5. [CrossRef]
15. Berardinelli, G.; Khosravirad, S.R.; Pedersen, K.I.; Frederiksen, F.; Mogensen, P. On the benefits of early HARQ feedback with non-ideal prediction in 5G networks. In Proceedings of the International Symposium on Wireless Communication Systems (ISWCS), Poznan, Poland, 20–23 September 2016; pp. 11–15.
16. Göktepe, B.; Fähse, S.; Thiele, L.; Schierl, T.; Hellge, C. Subcode-based Early HARQ for 5G. In Proceedings of the IEEE International Conference on Communications (ICC) Workshops, Kansas City, MO, USA, 20–24 May 2018. [CrossRef]
17. Rykova, T.; Göktepe, B.; Schierl, T.; Hellge, C. Analytical Model of Early HARQ Feedback Prediction. In *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*; Springer International Publishing: New York, NY, USA, 2020; pp. 222–239.
18. Makki, B.; Svensson, T.; Caire, G.; Zorzi, M. Fast HARQ Over Finite Blocklength Codes: A Technique for Low-Latency Reliable Communication. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 194–209. [CrossRef]
19. Hou, Z.; She, C.; Li, Y.; Zhuo, L.; Vucetic, B. Prediction and Communication Co-Design for Ultra-Reliable and Low-Latency Communications. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 1196–1209. [CrossRef]
20. Ericsson. *Way Forward on Processing Timing Reduction for sTTI*; Technical Report R1-165854; 3GPP. 2016. Available online: https://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_85/Docs/R1-165854.zip (accessed on 18 June 2021).
21. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2001.
22. MCC Support. 3GPP TS 38.212 v16.0.0. Technical Report. 3GPP. 2020. pp. 19–30. Available online: https://www.3gpp.org/ftp/Specs/archive/38_series/38.212/38212-g00.zip (accessed on 20 June 2021).
23. Strodthoff, N.; Göktepe, B.; Schierl, T.; Hellge, C.; Samek, W. Enhanced Machine Learning Techniques for Early HARQ Feedback Prediction in 5G. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2573–2587. [CrossRef]