Optimal BLER-Aware HARQ Redundancy Distribution under 5G Physical Layer Constraints

Julius Herrmann
Saarland Informatics Campus
Saarbrücken, Germany
jherrmann@ethz.ch
0009-0000-1889-5269

Marlene Böhmer
Saarland Informatics Campus
Saarbrücken, Germany
boehmer@nt.uni-saarland.de
0009-0002-5802-3443

Thorsten Herfet
Saarland Informatics Campus
Saarbrücken, Germany
herfet@nt.uni-saarland.de
0000-0002-3746-7638

***Abstract*—To enable reliable communication, 5G networks employ a Hybrid Automatic Repeat Request (HARQ) mechanism. Existing implementations typically operate with a fixed initial transmission target Block Error Rate (BLER) and equally sized HARQ transmission rounds. This paper proposes a redundancy distribution algorithm that, under a target BLER constraint, improves HARQ resource efficiency by jointly optimizing the initial target BLER and retransmission sizes, while satisfying physical layer constraints imposed by the 5G specification. Evaluation results for a representative scenario, obtained by simulation using a calibrated finite-blocklength channel model, show a resource reduction of up to $26\,\%$, with an average of $9.5\,\%$, compared to conventional fixed-size HARQ implementations in 5G networks and a more continuous resource consumption over the signal-to-noise ratio.**

***Index Terms*—5G mobile communication, Hybrid automatic repeat request, Finite blocklength, QoS, Resource distribution**

I. Introduction

The Hybrid Automatic Repeat Request (HARQ) mechanism is a central component of reliability in 5G wireless systems. Data channels use Incremental-Redundancy-HARQ (IR-HARQ) combining FEC with ARQ: LDPC-encoded transport blocks are split into partially overlapping Redundancy Versions (RVs), where each transmission adds redundancy information. The Medium Access Control (MAC) scheduler assigns resources to every transmission round and selects the Modulation and Coding Scheme (MCS). Prior work on HARQ has shown that performance can be significantly improved by carefully distributing redundancy information across HARQ transmission rounds [1]. In the context of 5G, exploiting this potential requires flexible assignment of resources to the transmission rounds and selecting the right MCS.

Typical initial-transmission BLER targets follow from the reported Channel Quality Indicator (CQI), which indicates suitable modulation and coding schemes at a BLER of at most $10\,\%$ (CQI tables 1 and 2) or 10^{-5} (CQI table 3). Transmissions are also typically of equal size[1]. However, many practical scenarios require different BLER targets, e.g. due to the packet error rate constraint of a 5G Quality of Service Identifier (5QI). Notably, 5QIs constrain the residual error rate after all transmissions, leaving freedom in how to achieve it.

[1]See default MAC scheduler implementations in 5G-Lena, MATLAB 5G Toolbox, and Open Air Interface.

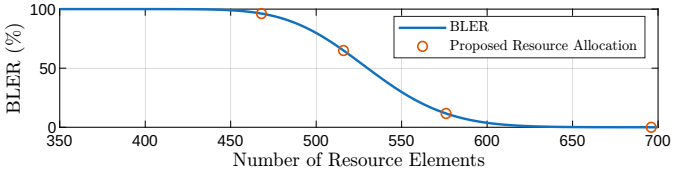


Fig. 1. Progression of the Block Error Rate (BLER) with increasing number of resource elements for QPSK modulation at SNR = 0 dB. The circles depict the proposed resource allocation for 4 transmissions with $\epsilon_{\text{target}} = 10^{-4}$.

We propose to configure the HARQ redundancy distribution for the target BLER required by the application, rather than the default $10\,\%$ initial-transmission BLER. This is complementary to upper-layer mechanisms (RLC mode, MCS table selection), which target specific failure scenarios such as residual radio link errors or mobility. Because LDPC-codes exhibit a waterfall shaped BLER curve over the number of assigned Resource Elements (REs) (see Figure 1), the proposed mechanism minimizes average required resource elements by entering the waterfall region with the initial transmission and only slightly increasing the allocation in subsequent rounds, enabled by a more flexible resource allocation. This mechanism is complementary to conventional MAC scheduling, which allocates resources among multiple UEs.

The contributions of our approach are as follows:

- adaptive HARQ redundancy distribution to meet transmission constraints in 5G
- added flexibility to target a resulting loss constraint after all HARQ transmission rounds
- reduced average resource consumption
- smoother performance across Signal-to-Noise Ratio (SNR)

II. Related Work

HARQ optimization has attracted considerable interest in the literature such as [1], which investigates the potential efficiency improvements of truncated HARQ transmissions by adjusting transmission lengths, and [2], which schedules HARQ transmissions under additional delay constraints. [1] provides a method to derive the optimal redundancy distribution for a given scenario. [3] investigates optimization opportunities by adjusting the modulation scheme of transmission

rounds. Although not the focus of this work, the authors highlight potential further improvements to be considered in mature systems such as 5G. Central to the optimization approaches is a sufficiently accurate BLER model based on finite-blocklength theory, especially when considering Ultra-Reliable Low-Latency Communication (URLLC) traffic [4]. [5], [6] set out to provide such a finite-blocklength error model based on the normal approximation of [7]. The channel model captures implementation losses of practical codes and is calibrated to fit the specific error code implementation under consideration. Although originally conceived for turbo codes, it can also be applied to LDPC codes as used in 5G, demonstrated by [3]. To the best of our knowledge, no prior work jointly considers optimal HARQ redundancy distribution and the practical constraints imposed by the 5G physical layer.

III. 5G HARQ Redundancy Distribution

To optimally distribute HARQ redundancy in 5G, the proposed algorithm produces a 5G-feasible resource allocation $\vec{RE}^{*}_{\text{conf}}$, expressed as a column vector of REs per transmission round, that is then scheduled by the MAC scheduler. We define the following optimization problem:

$$\vec{RE}^{*}_{\text{conf}} = \arg\min_{\vec{RE}_{\text{conf}}} \left(\overline{\text{RE}}(k, \vec{RE}_{\text{conf}}, \Omega) \right) \tag{1}$$

such that:

$$\epsilon(k, \|\vec{RE}_{\text{conf}}\|_1, \Omega) \leq \epsilon_{\text{target}} \tag{2}$$

$$\dim(\vec{RE}_{\text{conf}}) = N_C; \quad N_C \in \mathbb{N} \tag{3}$$

$$\eta(k, \vec{RE}_{\text{conf}}[1]) \in S_{\text{MCS}}; \quad k \in \mathbb{N} \tag{4}$$

$$12 \mid \vec{RE}_{\text{conf}}[c]; \quad \forall c \in [1, 2, \ldots, N_C] \tag{5}$$

$$\text{TBS}\left(\frac{\vec{RE}_{\text{conf}}[1]}{12}, \eta(k, \vec{RE}_{\text{conf}}[1]) \right) \geq k, \tag{6}$$

where k is the number of information bits, Ω is the SNR, and $\|\vec{RE}_{\text{conf}}\|_1$ refers to the 1-norm of $\vec{RE}_{\text{conf}}$. Control-plane overhead is left for future work. To give some intuition, (2) ensures that the target block error rate ϵ_{target} is achieved. (3) constrains the maximum number of transmissions to N_C, resulting in a truncated HARQ scheme. Any HARQ resource allocation for 5G networks has to adhere to a set of physical layer constraints imposed by the 5G specification:

(4): The spectral efficiency η used for the initial transmission must be selected from the standardized MCS tables. The set of possible spectral efficiencies is denoted as S_{MCS}. Subsequent transmissions can use the reserved entries in the MCS tables, which define only the modulation scheme, thereby allowing variable length transmissions.

(5): The minimum allocatable frequency-time unit is a Resource Block (RB), consisting of 12 subcarriers across one symbol, i.e., 12 REs. The resource allocation can therefore only consist of transmission rounds with an integer multiple of 12 REs, abstracting from the concrete resource mapping.

(6): Based on the initial resource allocation and selected spectral efficiency, the User Equipment (UE) calculates the Transport Block Size (TBS). Therefore, the algorithm must select a combination that yields a sufficiently large TBS. In our implementation, we select the next-closest viable TBS.

A. Channel Model

To fulfill (2), an error model that links k, the number of channel accesses N (REs in 5G), and Ω is required. Based on the normal approximation in [7], [5] proposes a calibrated maximum rate model that accounts for implementation losses. We reuse the calibration variables $\triangle \text{C}$ and γ from [5] but change their dependency from the SNR to the transmission rate R and N. The reformulations $\triangle \text{C}_M(R, N)$ and $\gamma_M(R, N)$ allow for better performance when predicting the BLER instead of the maximum achievable rate as in [5]. Using [6, Eq. 16] and adding the calibration variables, the expected BLER for an M-QAM modulated transmission over a complex Additive White Gaussian Noise (AWGN) channel can then be expressed as

$$\epsilon_M(k, N, \Omega) = \text{Q}\left(\frac{\text{C}_M(\Omega) - \triangle \text{C}_M(R, N) - R + \frac{\log_2(N)}{2N}}{\gamma_M(R, N) \cdot \sqrt{\frac{\text{V}_M(\Omega)}{N}}} \right), \tag{7}$$

where $R = \frac{k}{N}$ and $\text{Q}(\cdot)$ is the Gaussian Q-function. $\text{C}_M(\Omega)$ and $\text{V}_M(\Omega)$ refer to the channel capacity and dispersion of an AWGN channel with SNR Ω with M-QAM modulated transmission (see [8, Eq. 5] and [9, Eq. 4.4] respectively). To capture the implementation loss at large block lengths, $\triangle C_M$ is calibrated for block lengths of 1000, 5000, and 10000. A target BLER of $50\,\%$ is chosen as it corresponds to the inflection point of the BLER predicted by the normal approximation [1]. To capture small to medium block lengths, γ_M is calibrated for block lengths of 400, 500, 600, and 700. We choose two target BLERs of $70\,\%$ and $10\,\%$ to capture high and low error rate regions. The target BLERs and block lengths may be adjusted to fit the specific application scenario at hand. Both values are calibrated with link-level simulations following the methodology presented by [5], using the spectral efficiencies defined in MCS table 1 [10, Tab. 5.1.3.1-1] as simulation targets. The MATLAB 5G Toolbox is used as a standard-compliant 5G link-level simulator. The two-dimensional calibration variables are approximated using a planar approximation, and the calibration is performed separately for each relevant modulation order, i.e., QPSK, 16-QAM, and 64-QAM.

B. Redundancy Distribution

To find an optimal resource distribution, we use [1] as an intermediate step and then adapt the resource allocation for 5G. [1] finds an optimal distribution by setting up a Lagrangian function. As proposed in [1] we relax the problem by dropping the conditions (4), (5), and (6) — which will be ensured at a later step — and allow the distribution to take continuous values.

C. Algorithm and Adaption to 5G

Fig. 2. Algorithm calculates the optimal resource allocation for one UE. HARQCONFIGUREUE$(k, \Omega, \epsilon_{\text{target}}, N_C)$

1: $k' \leftarrow k + 24$
2: $N_{\min} \leftarrow \mathrm{N}_M(k', \Omega, \epsilon_{\text{target}})$
3: $\vec{RE}_{\text{conf}} \leftarrow \text{DISTRIBUTE}_M(\Omega, k', N_{\min}, N_C)$
4: $\eta \leftarrow \arg\min_{\eta \in S_{\text{MCS}}} \left(\left| \frac{k'}{\vec{RE}_{\text{conf}}[1]} - \eta \right| \right)$
5: $\vec{RE}_{\text{conf}} \leftarrow \text{VALIDATEALLOCATION}(\vec{RE}_{\text{conf}}, \eta)$
6: **return** $(\vec{RE}_{\text{conf}}, \eta)$

Fig. 3. Algorithm alters the given resource allocation such that it adheres to the 5G specification, i.e., the resource allocation fulfils constraints (4), (5), and (6) of the problem description. Variables k, k', Ω, ϵ_{target} and N_C are assumed to be known from the calling function HARQCONFIGUREUE. VALIDATEALLOCATION$(\vec{RE}_{\text{conf}}, \eta)$

1: $RB_{C=1} = \arg\min_{RB} \left(\text{TBS}(RB, \eta) \geq k \right)$
2: $\vec{RE}_{\text{conf}}[1] \leftarrow 12 \cdot RB_{C=1}$
3: **for** $i = 2$ to N_C **do**
4: $\quad N_{\triangle} \leftarrow N_{\min} - \|\vec{RE}_{\text{conf}}\|_1$
5: $\quad \vec{RE}_{\text{conf}}[i] \leftarrow 12 \cdot \left\lceil \frac{\vec{RE}_{\text{conf}}[i] \cdot \left(1 + \frac{N_{\triangle}}{\|\vec{RE}_{\text{conf}}[i \ldots N_C]\|_1}\right)}{12} \right\rceil$
6: **end for**
7: **return** $\vec{RE}_{\text{conf}}$

The HARQCONFIGUREUE Algorithm (Fig. 2) computes the resource allocation and adapts it for use in 5G with the help of VALIDATEALLOCATION (Fig. 3). A fixed CRC overhead of $24\,\text{bit}$ is assumed (line 1). In 5G, the CRC overhead may vary depending on the transport block structure, e.g., when multiple code blocks are used. Given the minimum number of resources required to achieve the target error rate determined by the channel model (line 2), the optimal allocation for each transmission round is computed (line 3). DISTRIBUTE_M implements the bisection method for the resource allocation of the initial transmission to minimize the average needed resources as developed in [1]. The resulting optimal resource allocation is continuous, thus not directly usable in 5G. To fulfil condition (4), a spectral efficiency which is offered by the selected 5G MCS table has to be chosen for the initial transmission. The continuous resource allocation indicates a target spectral efficiency $\frac{k'}{\vec{RE}_{\text{conf}}[1]}$, which we approach heuristically by the closest available spectral efficiency (line 4).

The VALIDATEALLOCATION Algorithm ensures the resource allocation is valid for 5G use, at the cost of a possible deviation from the continuous optimum. Given the chosen spectral efficiency, the minimum number of RBs is determined such that the calculated TBS can carry at least k bits. The first round then fulfils conditions (5) and (6). The resulting difference in totally distributed resources from $N_{\min}$ is then distributed to the remaining rounds that are quantized to integer multiples of RBs, maintaining their relative size. In cases where no combination of RB and η results in a TBS of exactly k, the next larger TBS is selected and $N_{\min}$ is recalculated accordingly. The difference from $N_{\min}$ is recalculated for each round as quantisation of the previous round impacts this difference for the next round.

These Algorithms only apply to one fixed modulation scheme; thus, to use them with the 5G MCS tables, which contain 3 to 4 modulation schemes, they are executed for each modulation scheme with selectable spectral efficiencies corresponding to that modulation scheme, choosing the result that minimizes the average needed resources.

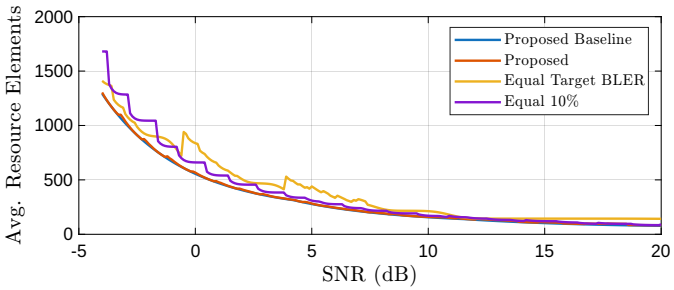
IV. EVALUATION

Evaluations are performed in MATLAB and compare the resulting resource allocations from our proposed algorithms to a baseline and to allocations that keep the retransmission length equal to the initial transmission. All results are based on the same channel model that was validated against MATLAB 5G toolbox link-level simulations.[2] The *Proposed* allocation is calculated as described in Section III-C. The *Proposed Baseline* allocation is calculated by using the same continuous allocation, but not adapting it to 5G; thus, it fulfils the conditions on the target BLER (2) and number of transmission rounds (3), but does not fulfil the conditions (4), (5), and (6). The *Equal 10%* allocation targets an initial BLER of 10% and keeps the resource allocation constant for all transmission rounds, emulating the use of an MCS index, as indicated by a CQI. Retransmissions comprise the same length as the initial transmission. The *Equal Target BLER* allocation also allocates the same resources for each transmission round, but instead of targeting a 10% BLER for the initial transmission, achieves the target BLER ϵ_{target} after all transmissions (condition (2)). Both fixed-size allocations also fulfil all 5G-specific constraints.

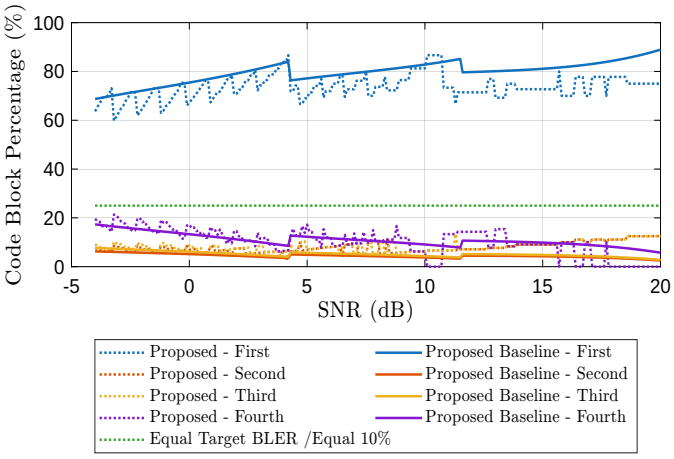
The presented evaluation results are based on the same basic parameters. The number of transmission rounds is set to $N_C = 4$ and the overall target BLER after the complete HARQ mechanism is $\epsilon_{\text{target}} = 10^{-4}$. The parameters are chosen to allow a fair comparison to the *Equal 10%* allocation in the way that, with 4 HARQ transmissions, this allocation also meets the target BLER calculated by $N_C = \lceil \log_{0.1}(\epsilon_{\text{target}}) \rceil$, and in fact undershoots by orders of magnitude. The set of available spectral efficiencies is based on MCS table 1 for PDSCH and the range of evaluated SNRs from $-4\,\text{dB}$ to $20\,\text{dB}$ spans the operating ranges of all three modulation schemes with 2, 4 and 6 bits per RE, defined in MCS table 1.

The average number of REs in Fig. 4a shows that the *Proposed* allocation reduces the required resources by $9.5\,\%$ on average over the SNR (up to $26\,\%$) compared to the *Equal 10%* allocation and by $23\,\%$ compared to the *Equal Target BLER* allocation, while remaining within $3.2\,\%$ additional resources of the optimal unconstrained *Proposed Baseline*, i.e., adhering to the 5G constraints costs little. The *Proposed* allocation achieves a first-transmission BLER of $83\,\%$ and requires 2.2 transmission rounds on average. It also smooths the fluctuations at transitions between spectral efficiencies observed

[2]Implementation: https://git.nt.uni-saarland.de/research-projects/independent/5GHARQdistribution



(a) Average resource requirements.



(b) Allocation per transmission round.

Fig. 4. Comparison of resource allocations for $k = 408$, $N_C = 4$, and $\epsilon_{\text{target}} = 10^{-4}$.

with the equal-length schemes. With this flexibility in resource allocation, the proposed allocation can both achieve reasonable spectral efficiency and maintain low average resource usage while fulfilling the BLER target. Fig. 4b shows that it achieves this by assigning approximately $80\,\%$ of the total resources to the initial transmission. The following transmissions contain only a small share of the overall resources — large enough to decrease the residual BLER (Fig. 1) but barely increasing resource usage in case of early termination due to successful decoding. The last transmission contains a slightly larger share of resources, which pushes the BLER below the target.

V. Discussion and Future Work

The proposed algorithm and resulting resource allocation comply with the 5G specification constraints and can therefore be implemented without modifying existing 5G networks beyond the scheduler's resource allocation logic. The algorithm requires a target BLER, selected based on the configured 5QI's reliability requirements. Furthermore, the algorithm requires knowledge of the current channel quality. This can be achieved either by mapping the reported CQI to estimated SNR values or by explicitly configuring the Channel State Information (CSI) to report Signal-to-Interference Noise Ratio (SINR) measurements. For a single UE, the computation reduces to a root search for ($N_{\min}$), the solution of the (N_C) optimality conditions, and one quantisation pass, repeated for at most three or four modulation orders, depending on the selected MCS table, where infeasible modulation orders are discarded beforehand by simple checks. For many-UE low-latency scenarios, the pre-calculated fixed allocation of [1] remains applicable, as the relative allocations in Fig. 4b depend only weakly on SNR.

Future work includes the delay aspect of the HARQ mechanism and how a resource allocation can additionally fulfil a delay constraint to support both reliability and latency requirements as in [2]. System-level simulations are planned to evaluate the accuracy and practical feasibility and to determine the interaction between HARQ resource allocation and UE scheduling. Furthermore, the accuracy of the applied channel model is to be verified using real-world deployments. Another direction of interest is adaptive redundancy allocation under time-varying channel conditions.

VI. Conclusion

This paper investigated the optimal distribution of redundancy for truncated HARQ transmission in 5G networks. We considered practical 5G physical layer constraints and imposed a target BLER requirement. We formulated a constraint-aware optimization problem and, in conjunction with a calibrated finite-blocklength error model, developed a HARQ redundancy distribution algorithm to minimize overall resource requirements. Evaluation results obtained so far indicate that explicitly optimizing HARQ transmission sizes reduces average resource requirements compared to a fixed-size 5G HARQ implementation. Overall, this work demonstrates that optimal HARQ redundancy distribution can complement existing 5G mechanisms to further improve resource efficiency when operating under a target BLER constraint.

References

[1] Y. Fu, A. Lozano, and H. Yang, "Variable-Rate Incremental-Redundancy HARQ for Finite Blocklengths," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 1413–1422, 2025.

[2] K. Vogelgesang, P. G. Pereira, and T. Herfet, "SHARQ: Scheduled HARQ for Time- and Loss-Rate-Sensitive Networks," in *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, 2023, pp. 640–643.

[3] M. Jiang, Y. Wang, F. Ding, and Q. Xu, "Finite-Blocklength Analysis of Coded Modulation with Retransmission," *Entropy*, vol. 26, no. 10:863, 2024.

[4] T.-H. Vu, M. Zeng, S. Kim, H. Vincent Poor, and Q.-V. Pham, "Short-Packet Communications: Recent Advances and Research Challenges," *IEEE Network*, vol. 40, no. 4, pp. 268–275, 2026.

[5] E. C. Song and G. Yue, "Finite Blocklength Analysis for Coded Modulation with Applications to Link Adaptation," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–7.

[6] ——, "Estimate BLER for Coded Modulation Based on Finite Block Coding," in *32nd Wireless and Optical Communications Conference, WOCC 2023, Newark, NJ, USA, May 5-6, 2023*. IEEE, 2023, pp. 1–6.

[7] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[8] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inf. Theory*, vol. 28, no. 1, pp. 55–66, Jan. 1982.

[9] Ebrahim MolavianJazi, "A unified approach to gaussian channels with finite blocklength," Ph.D. dissertation, University of Notre Dame, 2014.

[10] 3GPP, "5G; NR; physical layer procedures for data," 3GPP, Tech. Rep. TS 38.214 V19.1.0 (2025-09), 2025.