Hybrid Automatic Repeat Request (HARQ) in Wireless Communications Systems and Standards: A Contemporary Survey

Ashfaq Ahmed, *Senior Member, IEEE*, Arafat Al-Dweik, *Senior Member, IEEE*, Youssef Iraqi, *Senior Member, IEEE*, Hussam Mukhtar, *Member, IEEE*, Muhammad Naeem, *Senior Member, IEEE*, Ekram Hossain, *Fellow, IEEE*

***Abstract*—Automatic repeat request (ARQ) schemes, and in particular hybrid-ARQ (HARQ) schemes, which jointly adopt forward error correction (FEC) and ARQ, are essential to provide reliable data transmission in wireless communications systems. However, the feedback from the receiver to the transmitter and the retransmission process used in ARQ incurs significant cost in terms of power efficiency, throughput, computational power and delay. Unfortunately, such drawbacks can limit their applications to several current and emerging technologies. More specifically, the increasing number of wireless users has create spectrum scarcity, relying on small-size batteries create power constraints, deployment of real-time applications boost the demand for ultra-low delay networks, and the ultra-small low-cost internet of things (IoT) devices has limited signal processing and computation capabilities. Consequently, extensive research efforts have been dedicated to overcome the limitations inherent in HARQ. This survey paper provides an extensive literature review of the state-of-the-art HARQ techniques and discusses their integration in various wireless technologies. Moreover, it provides insights on advantages and disadvantages of particular ARQ types and discusses open problems and future directions.**

***Index Terms*—Wireless communications systems, Retransmission protocols, Automatic Repeat Request (ARQ), Hybrid ARQ (HARQ)**

I. Introduction

THE basic task of a communication system is to transfer a data unit from one point to another while satisfying certain quality of service (QoS) requirements and resources constraints. The data unit can be a bit, symbol, packet or frame. The QoS requirements may include the bit error rate (BER), packet error rate (PER), data transfer rate and delay. The resources required for a wireless communications system are typically power, energy, time, space, spectrum and hardware with certain specifications. The efficiency of a communications system depends on the amount of resources required to complete the transfer process successfully, and it is highly desirable to maximize the efficiency by minimizing the utilized resources. The QoS requirements are determined by the targeted application. For example, in certain internet of things (IoT) applications, the data rate can be as low as 1 bps [1] while it can be about 10 Gpbs, as in the case of full-color 3D holographic display system [2].

The international telecommunication union (ITU) classifies various applications based on their QoS requirements into 8 classes [3]. Table I shows the QoS requirements for each class in terms of packet time delay (PTD), packet drop rate (PDR) and packet error rate (PER). The bit rate (R_B) range is based on practical measurements performed in Germany. Classes with short PTD and small R_B requirements, such as Classes 0 and 1, target real-time jitter-sensitive interactive applications, such as voice over internet protocol (VoIP) and video conferences. Classes 2 and 3 target signaling traffic and interactive applications with high data transfer rates. Class-4 should support short transactions, video streaming or bulk data. Class-5 is unspecified (regarding all performance parameters) and is targeted to traditional best effort Internet applications. Classes 6 and 7 are provisional, and are needed for new emerging applications that require very low PTD and PER, but high R_B. fifth generation (5G) technologies are expected to provide throughput of 1-10 Gbps, and an end-to-end delay of 1 ms [4] and they are designed to support all QoS classes.

Due to noise, interference, channel fading, and other impairments, data transferred between two communicating entities are prone to errors [5]. Therefore, the underlying communications protocols should be designed to correct data errors in order to support a particular QoS class PER. Generally speaking, the wireless communications systems utilize protocols at different layers to correct data errors at the receiver. At the physical layer, error detection and/or correction are performed using forward error correction (FEC) coding [6]. At the radio link and the transport layers, error correction relies on retransmission-based methods such as automatic repeat request (ARQ). Also, at the application layer, retransmission methods can be applied to recover some lost or corrupted packets. Adopting one or more of the aforementioned methods depends on the system resources and QoS requirements. Nevertheless, FEC and ARQ are essential building blocks for most popular standards such as fourth generation (4G) wireless standard [7],

Ashfaq Ahmed, Arafat Al-Dweik, and Youssef Iraqi are with Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE (email: {ashfaq.ahmed, arafat.dweik, youssef.Iraqi}@ku.ac.ae).

Hussam Mukhtar is with the College of Engineering, University of Dubai, Dubai, UAE (email: hhadam@ud.ac.ae).

Muhammad Naeem is with the Department of Electrical & Computer Engineering, COMSATS University Islamabad, Wah campus, Pakistan (email: mnaeem@ciitwah.edu.pk).

Ekram Hossain is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada (email: Ekram.Hossain@umanitoba.ca). The work of E. Hossain was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

TABLE I: ITU QoS classes.

	PTD (ms)	**PDR**	**PER**	R_B **Range (Mbps)**
Class-0	100	10^{-3}	10^{-4}	< 2
Class-1	400	10^{-3}	10^{-4}	2 to 8
Class-2	100	10^{-3}	10^{-4}	8 to 18
Class-3	400	10^{-3}	10^{-4}	18 to 25
Class-4	1000	10^{-3}	10^{-4}	25 to 50
Class-5	-	-	-	50 to 100
Class-6	100	10^{-5}	10^{-6}	100 to 200
Class-7	400	10^{-5}	10^{-6}	200 to 500

[8], 5G [9], wireless fidelity (WiFi) [10], and narrow band (NB)-IoT [11]. A scheme that jointly adopts FEC and ARQ for error correction is denoted as hybrid-ARQ (HARQ).

Although using FEC and ARQ is indispensable for reliable communications, they also increase the end-to-end delay and computational complexity, reduce the throughput, spectral and energy efficiencies, particularly for systems that support several types of applications with various QoS requirements. Therefore, extensive research effort has been steered to alleviate the adverse effects of FEC and ARQ. However, ARQ has more degrees of freedom as compared to FEC, which makes it a fertile research area. This survey paper provides an extensive literature review of the research work that considers ARQ, highlights limitations, challenges and future opportunities.

A. Review of Existing Survey Literature

This survey provides a comprehensive literature review for the techniques and schemes adopted for retransmission in most common communications standards such as WiFi, Bluetooth, NB-IoT, 4G and 5G networks. Moreover, it also includes several state-of-the-art and emerging technologies such as non-orthogonal multiple access (NOMA), cooperative communications, and visible light communications. To the best of the authors' knowledge, there is no work reported in the literature that covers ARQ and HARQ in such a holistic manner. The main survey work that considers HARQ is summarized in Table II, and detailed as follows.

Ngo and Hanzo [12] studied HARQ in the context of cooperative wireless communications, summarized and compared the performance of several cooperative HARQ schemes. Moreover, the authors proposed a relay-switching scheme to increase the achievable throughput along with a general design procedure for HARQ-cooperative systems. The work in [13] has surveyed several HARQ papers, however, the main focus of the paper is dedicated for developing two new techniques for reducing the system complexity using the mutual information. In the first technique, an early stopping strategy is applied to reduce the complexity of iterative decoder. In the second approach, the iterative decoding processes is differed until the receiver is confident that it has sufficient information for successful decoding. Mukhtar *et al.* [6] compared the performance of various FEC schemes, and investigated the integration of TPCs and ARQ. The presented results show that TPCs are highly appropriate for HARQ due to their inherent error self-detection capabilities. A survey of different schemes reported in the literature for underwater acoustic networks from the data link to transport layers is presented in [14]. Several HARQ techniques have been investigated at different layers for reliable data transfer in underwater networks.

B. Motivation and Contribution

As can be noted from the examined survey work on HARQ, it can be realized that there is no work that considers HARQ in a holistic manner. In particular, HARQ has been integrated in several emerging technologies and adopted by new standards, which are not considered in the existing surveys. The main contributions of this article are as follows:

1) Presents a comprehensive summary of start-of-the-art HARQ advancements.
2) Surveys various error detection schemes used for HARQ.
3) Discusses the integration and performance of HARQ with several wireless technologies such as ultra reliable low latency communication (URLLC), cooperative communications, multiple-input multiple-output (MIMO), massive MIMO, NOMA, cognitive radio (CR), caching, simultaneous wireless information and power transmission (SWIPT), and unmanned aerial vehicle (UAV) assisted communications.
4) Summarizes the currently utilized HARQ in most common standards such as WiFi, 4G, 5G, wireless personal area networks (WPANs), and NB-IoT.
5) Investigates the main performance metrics of HARQ such as throughput, latency, energy efficiency, and complexity.
6) Presents some intelligent HARQ schemes for multimedia applications.

C. Paper Organization

In Section II, the fundamentals of the HARQ system are presented. The classification of HARQ on the basis of retransmission data is presented in Section III. In Section IV, various metrics to evaluate the performance of a HARQ system are presented. Advance HARQ techniques are provided in Section V. An overview of the HARQ employed by various wireless standards is given in Section VI, whereas, Section VII covers the integration of HARQ in emerging wireless technologies. The future directions are provided in Section VIII, and finally the conclusion is drawn in Section IX. The acronyms and the list of symbols used throughout this paper are shown in Appendix I and Appendix II, respectively.

II. Basics of HARQ

The typical HARQ consists of four main operations, FEC, error detection, combining, and retransmission. At the transmitter, HARQ systems generally follow the structure described in [15], which is shown in Fig. 1 for block codes, and can be described as follows:

1) A binary information block $\mathbf{d} = [d_1, d_2, \ldots, d_K]$ is divided into L equal parts, $\mathbf{d} = \left[\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \ldots, \mathbf{d}^{(L)}\right]$, where $\mathbf{d}^{(i)} = \left[d_1^{(i)}, d_2^{(i)}, \ldots, d_k^{(i)}\right]$, $d_i \in \{0, 1\}$, $i \in \{1, 2, \ldots, L\}$, and $k = K/L$.

TABLE II: Summary of existing HARQ survey papers, where A: application, T: Technology, and O: Optimization.

	Year	Scope	Orientation	Remarks
[12]	2014	Cooperative wireless communication	A	Review the state-of-art and investigate the performance of HARQ in the context of cooperative wireless communications. Proposed a general design guidelines for HARQ aided cooperative systems
[13]	2013	Turbo codes to be used with HARQ techniques	T	New iterative turbo codes techniques are presented to be used with HARQ in order to save the power consumption due to turbo codes iterative decoding.
[6]	2016	Turbo product codes (TPCs) in concatenation with HARQ	T	TPCs in terms of encoding, decoding, error performance, and complexity is surveyed. Also considers the advantages of integrating TPCs in hybrid automatic repeat request systems
[14]	2018	Reliable retransmissions in UnderWater Acoustic (UWA) networks	T	Surveyed on Reliable Data Transfer in Underwater Acoustic (UWA) Networks. Retransmission mechanism is studies for enhancements in data link layer.
This work	-	-	{A, T, O}	Overview of ARQ and HARQ schemes and their application to various wireless standards and emerging and state-of-the-art communications technologies.

2) Each of the $\mathbf{d}^{(i)}$ sequences is applied to an error detection encoder where l_{ed} bits are appended to $\mathbf{d}^{(i)}$ for error detection purposes at the receiver side. The error detection encoder output can be written as,

$$\mathbf{c}^{(i)} = \Big[\underbrace{c_1^{(i)}, c_2^{(i)}, \ldots, c_k^{(i)}}_{d_1^{(i)}, d_2^{(i)}, \ldots, d_k^{(i)}}, c_{k+1}^{(i)}, c_{k+2}^{(i)}, \ldots, c_{k+l_{\text{ed}}}^{(i)}\Big]. \tag{1}$$

3) Each of the sequences $\mathbf{c}^{(i)}$ is then applied to a channel encoder that appends l_{ec} bits to $\mathbf{c}^{(i)}$ to form a codeword $\mathbf{u}^{(i)}$. The number of bits applied to the channel encoder input is $k + l_{\text{ed}}$, and the number of output bits is n. The parity bits can be expressed as $\mathbf{p}^{(i)} = [p_1^{(i)}, p_2^{(i)}, \ldots, p_{l_{\text{ed}}}^{(i)}]$.

4) Several codewords are buffered to form a large block of bits $\mathbf{U} = [\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(L)}]$ is interleaved to mitigate the effect of burst errors. The interleaver output for a given codeword is denoted as $\mathbf{B} = \mathcal{I}(\mathbf{U})$.

5) The interleaver output $\mathbf{B}$ is modulated using a particular modulation scheme. The modulation type and order might be identical for the entire packet, or might be adaptively changed based on the channel conditions [16]–[18]. The modulated codeword $\mathbf{s}^{(i)} = [s_1^{(i)}, s_2^{(i)}, \ldots, s_Q^{(i)}]$, where Q is the total number of modulated symbols in the sequence.

6) To form a packet $\mathbf{z}^{(i)} = \left[z_1^{(i)}, z_2^{(i)}, \ldots, z_N^{(i)}\right]$, one or more modulated sequences are grouped, then the packet header is appended. The packet header consists of several fields, which are required for addressing, and providing other necessary information to enable the receiver extract the information bits correctly. Examples for the header fields are source port, destination port, sequence number, and priority indicator. The packet header size is a key factor that determines the transmission efficiency, and it depends on the considered protocol. For the IPv4, the header size is 192 bits while it is 320 bits for the IPV6. To simplify the presentation of the main concepts, we assume without loss of generality that each packet consists of one modulated sequence, unless it is mentioned otherwise, thus, $\mathbf{z} = \mathbf{s}$. It is also worth noting that one or more packets should be stored in a buffer for retransmission purposes.

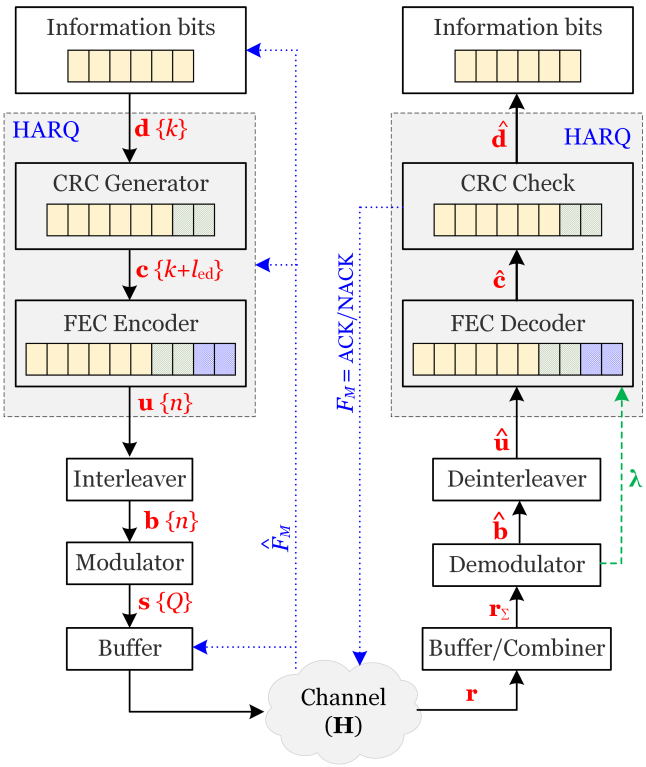

Fig. 1: Overview of HARQ process, including ARQ and FEC components.

At the receiver, the order of execution and required processing depend on the system design. Nevertheless, the typical sequence starts with demodulation, deinterleaving, FEC, error detection, and finally retransmission. Fig. 1 shows a basic diagram for the HARQ process, and brief description of each process is given in the following subsections. The processes are ordered according to the receiver side in Fig. 1, i.e., bottom-up approach. As can be noted from Fig. 1, the received sequence might be buffered to allow combining in the case that sequence is transmitted more than once. In Fig. 1, the numbers in curly brackets refer to the sequence size.

A. Forward Error Correction

The main task of FEC is to correct the received packets when they have relatively small number of errors, which may spare the need to retransmit such packets. FEC can be realized using various types of channel codes such as turbo codes, low density parity check (LDPC) codes, TPCs, or polar codes. Channel codes are generally characterized by:

1) Code rate (R): It is the ratio between the number of input and output bits of the encoder, $(k + l_{\mathrm{ed}})/n$, which indicates the additional bandwidth requirements due to the FEC process. The bandwidth expansion remains fixed unless the code rate is changed during transmission. It is worth noting that the error detection bits are also encoded by the channel encoder.
2) Coding gain: It is the indicator for the error correction capability of the channel code. The coding gain depends on the channel code type and the adopted decoding process. The main factors that determine the performance of the decoding process are the decoder type, i.e., soft decision decoder or hard decision decoder, number of iterations, and the decoding early stopping criterion [19]–[22].
3) Complexity: The decoding process is usually more computationally involved than the encoding process, particularly when soft iterative decoding is applied [19]–[21].
4) Delay: The delay results mostly from the iterative sequential decoding process [19].

As can be noted from the main characteristics of channel codes, the coding gain, complexity and delay are affected by the iterative decoding process. Therefore, extensive research effort was devoted to overcome the adverse effects of the iterative decoding process. The main approach followed was the use of early stopping criterion [23]–[41]. Other approaches were also proposed to reduce the complexity and delay of the decoding process. For example, Al-Dweik *et al.* proposed replacing some of the soft decoding iterations with the less complex hard decoding iterations [20]. In [21], the authors proposed a new approach to perform soft decision decoding without algebraic hard decision decoder, which can reduce the processing time and complexity substantially, but at the expense of some coding gain degradation. In [13], the authors utilized the mutual information to stop the decoding process when no further corrections can be made by performing additional iterations, or to skip the decoding process due to the lack of sufficient information for successful decoding, which is particularly suitable for HARQ. In [42], the authors demonstrated that using TPCs with decoding can provide equivalent throughput to the soft-input soft-output (SISO) decoding for several operating scenarios. However, hard-input hard-output (HIHO) decoding has much lower delay and complexity. The advantage of using HIHO can be also improved when using advanced schemes as described in [19], [22]. Examples for research that considered particular FEC and HARQ are:

1) Turbo codes: [43]–[50]
2) TPCs: [42], [51]–[56]
3) LDPC codes: [57]–[70]
4) Polar codes: [71]–[78]

Other codes such as Reed-Solomon, convolutional and Bose–Chaudhuri–Hocquenghem (BCH) codes have been considered for HARQ, however, such codes have been replaced by the aforementioned more powerful codes.

B. Error Detection

After channel decoding, it is generally difficult to accurately know if the decoder output matches the transmitted packet or not, unless some additional verification is performed. Therefore, an additional error detection process is needed to verify the channel decoder output. cyclic redundancy check (CRC) is the most common technique used for error detection, and it has been adopted in several wireless applications such as 4G [7], [8], 5G [9], and WiFi [10]. The CRC operation can be realized using various software and hardware implementations. However, using dedicated hardware is a must for high speed requirements [79], [80]. The generic CRC hardware implementation is based on low complexity linear feedback shift registers (LFSRs) that perform the polynomial division process of the serial input data. In the presence of wide data buses, parallel processing can be applied to enable processing large number of bits simultaneously, at the expense of additional hardware complexity [81]–[83].

In addition to complexity, CRC requires adding redundant bits for error detection. The number of CRC bits depends on the length of the transmitted packet, and it typically ranges from 8 to 64 bits. Therefore, CRC bits can deteriorate the system throughput, particularly for small packet lengths. To avoid throughput reduction, several techniques have proposed CRC-free error detection schemes. For example, a simple error detection scheme is proposed in [55] for TPCs by performing an additional half decoding iteration, and computing the Hamming distance between the input and output of the decoder. If the Hamming distance is zero, the decoding process is considered successful, and failure otherwise. Coulton *et al.* [84] proposed a simple error detection scheme by comparing the ratio of the standard deviation (STD) to the mean of the soft information energy at the output of the decoder. However, the results indicate that a reliable performance requires a large sample size to estimate the mean and STD accurately. Moreover, the system requires a substantial number of operations to compute the energy, and then the mean and STD of the soft information. Buckley and Wicker [85] used neural networks for CRC-free error detection. The main drawback for using neural networks is that network training requires large number of samples, particularly in time-varying channels. In [40], CRC-free error detection is achieved by continuously monitoring the log-likelihood ratio (LLR) of the soft information at the decoder output. However, the results presented in [40] show that the system performance may vary significantly based on the signal to noise ratio (SNR), frame size and code rate. Furthermore, accurate knowledge of the channel statistics is necessary to compute the LLR. Error detection using the word and bit error probabilities is proposed in [86]. Nevertheless, this approach suffers from low detection rates, which causes throughput inflation. Checksum is generally similar to CRC in the sense that it also requires

redundant bits to be added at the transmitter, and then used for data verification at the receiver [87, pp. 106].

C. Retransmission

After channel decoding, the decoder output is applied to an error detection process to verify the data integrity, which is typically performed by computing the CRC bits of decoded packet, and comparing them to the received CRC bits. If the packet passes the CRC check, an acknowledgment (ACK) is sent to the transmitter for confirmation, otherwise a negative acknowledgment (NACK) is sent to request the transmitter to resend the erroneous packet. The retransmission process is terminated when the transmitter receives an ACK for the packet, or when the maximum number of transmissions allowed C is reached. Moreover, the transmitter expects to receive the feedback message (F_M), ACK/NACK, within a limited time after transmitting the packet, if no feedback is received due to the loss of the packet or F_M, the transmitter assumes that the packet is lost and automatically resends it. The specified time limit, denoted as timeout (t_{out}), is set relative to the estimated round trip time (RTT), which is the time elapsed between sending a packet and receiving the F_M after the last successful retransmission of the same message. The RTT can be estimated using a moving average of previously measured RTTs. Generally speaking, there are three main types of ARQ protocols [88], namely, stop-and-wait (SW), go-back-N (GBN), and selective repeat (SR).

1) SW-ARQ: This protocol is given in **Algorithm 1**, where L indicates the total number of packets to be transmitted. SW-ARQ is simple and requires the transmitter to buffer only one packet. However, it is inefficient because the channel remains idle until an F_M is received. Moreover, sending F_M for each packet increases the feedback overhead, which degrades the system throughput. Another drawback is that when an F_M sent by the receiver is damaged or lost, the transmitter will resend the same packet after waiting for t_{out}, which causes packet duplication at the receiver. If an F_M is delayed, the sender assumes that the data packet is damaged, and thus resends that packet. Consequently, the transmitter will consider that the delayed FM acknowledges the last packet, while it is not the case. To avoid packet duplication and false acknowledgments [88], a simple numbering scheme can be used where packets are numbered sequentially as $\{0, 1, 0, 1, ...\}$, and acknowledgments are numbered as $\{1, 0, 1, 0, ...\}$. The acknowledgment number actually indicates the frame number that should be transmitted. For example ACK-0 implies that Packet-0 should be transmitted. Therefore, if the receiver sends two F_Ms with the same number, the last F_M is considered as a NACK. In full-duplex and half-duplex communications, both parties send data and feedback messages because both parties implement ARQ. In such scenarios, piggybacking can be used to improve the system spectral efficiency by sending outstanding F_Ms as part of the packets' header. Consequently, an F_M will consist only of the of packet number to be transmitted, which is one bit for the SW-ARQ.

2) GBN-ARQ: This protocol is widely used in practice because it overcomes the inefficiency limitation of SW protocol as the transmitter continues sending enough packets to keep the channel busy while waiting for an F_M. Unlike the SW-ARQ, the data packets and F_Ms should be numbered regularly, which enables acknowledging multiple packets using a single F_M. The functionality of the GBN-ARQ can be described as follows [88]:

Algorithm 1: SW-ARQ

Input: C, t_{out}, L

1. Initialization: $n = 1$, $c = 1, i = 0$
2. Send packet i
3. Timer Reset, $t = 0$
4. Read: t, Feedback message (F_M)
5. if $\{(F_M =$ ACK-0 or $[F_M = \varnothing, t > t_{out}])$ and $n \leq L\}$
6. if $c \leq C$
7. $c = c + 1$
8. Go to Step 2
9. else
10. $n = n + 1$, $c = 1$, $i = 1$
11. end
12. elseif $F_M =$ ACK-1 and $n \leq L$
13. $n = n + 1$, $c = 1$, $i = 1$
14. Go to step 2
15. end

1) A window of size W is assigned to buffer outstanding packets.
2) An m-bit sequence numbers are used for both the frames and ACKs, $W = 2^m - 1$.
3) The transmitter sends multiple packets numbered sequentially. The number of unacknowledged frames outstanding is determined by window size.
4) Assuming that packets $1, 2, ..., v$ are received with no errors, the receiver sends ACK-$v + 1$ to acknowledge the v packets simultaneously. Transmitter also shifts the window v positions
5) If packet-i, $1 \leq i \leq v$, is not received correctly, packet-i and all subsequent packets are discarded, and ACK-i is sent to the transmitter.
6) The transmitter retransmits packets $i, i + 1, ..., v$.

GBN protocol uses only one timer for the first outstanding packet, because it is the one that expires first. If timeout is reached without receiving F_M, the transmitter resends all outstanding packets. Although GBN is more spectrally efficient than SW, its performance deteriorates significantly in severe communications channels, which due to the fact that multiple packets have to be retransmitted whenever a packet is damaged or lost, or when a timer timeout is reached. Moreover, excessive retransmissions causes large delays. Although the received packets are numbered, the GBN does not support packet reordering because the receiver can buffer only one packet.

3) SR-ARQ: Also known as Selective Reject, mitigates the limitations of the GBN by adding two new features. First, the receiver window is increased to store more than one packet to enable accepting out-of-order but error-free packets. Second, the SR protocol allows retransmission of individual erroneous packet.

III. Classification of HARQ Based on Retransmitted Data

Although SR is more efficient than SW and GBN, it suffers low efficiency due to the retransmission of the entire erroneous packet, and discarding the packets that fail the CRC. Therefore, several techniques were developed to reduce the amount of information retransmitted in the case a transmission was not successful. HARQ can be classified based on the amount of data retransmitted into three main types [89]:

A. HARQ Type-I: Chase Combining

HARQ systems with fixed FEC are denoted as type-I HARQ (HARQ-I). In HARQ-I, the same data packet $\mathbf{s}$ is transmitted in all retransmissions. Then, the receiver can replace the erroneous packet with the new one, and repeat the process until the packet is received correctly, or the maximum number of transmissions is reached. Alternatively, the receiver may exploit the channel variation over consecutive retransmissions, and combine all transmissions that correspond to a particular packet to achieve a significant diversity gain. This type of HARQ is also referred to as chase combining (CC).

By noting that each modulated sequence $\mathbf{s}$ can be detected independently from the other sequences, thus it can be treated as complete packet. Consequently, at the receiver, the received packet during tth transmission session can be expressed as

$$\mathbf{r}^{(t)} = \mathbf{H}^{(t)}\mathbf{s} + \mathbf{w}^{(t)},\ t = 1, 2, \ldots, C \tag{2}$$

where $\mathbf{H} \in \mathbb{C}^{Q\times Q}$ is the channel frequency response coefficients, $\mathbf{H} = \text{diag}[H_1, H_2, \ldots, H_Q]$, $\mathbf{s} \in \mathbb{C}^{Q\times 1}$ is the modulated data sequence and $\mathbf{w} \in \mathbb{C}^{Q\times 1}$ is the additive white Gaussian noise (AWGN). The data symbols in $\mathbf{s}$ are selected uniformly from a particular constellation scheme such as M-ary phase shift keying (MPSK) or quadrature amplitude modulation (QAM). The channel fading coefficients are independent and identically distributed (i.i.d.) zero-mean complex Gaussian random variables with variance σ_H^2, and the AWGN is zero-mean complex Gaussian random variable with variance σ_w^2.

The received sequence $\mathbf{r}$ will be demodulated, deinterleaved, decoded using the FEC decoder, and checked for errors. If the packet is error free, an ACK is sent to the transmitter to proceed with the transmission of the next packet. Otherwise, $\mathbf{r}$ is stored in a buffer and a NACK is sent to instruct the transmitter to retransmit the packet. Once a packet is retransmitted, the newly received sequence will be combined with the stored erroneous versions using CC to enhance the quality of the received signal. The CC is similar to maximum ratio combining (MRC), which is optimal for systems with receiver diversity [5, pp. 994]. However, as reported in [90], MRC is not optimal for HARQ because the knowledge that the packet failed the CRC check can be exploited to optimize the combining process to produce results that are better than MRC. Given that CC is used, the Chase combiner output after ℓ successive transmissions is equal to

$$\mathbf{r}_\Sigma = \left[\sum_{t=1}^{\ell}\left[\mathbf{H}^{(t)}\right]^H \mathbf{H}^{(t)}\right]^{-1} \sum_{t=1}^{\ell}\left[\mathbf{H}^{(t)}\right]^H \mathbf{r}^{(t)} \tag{3}$$

where $[\cdot]^H$ is the Hermitian transpose. The CC of ℓ sequences will be denoted as $\mathcal{C}_\mathcal{H}\left[\mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \ldots, \mathbf{r}^{(\ell)}\right]$.

After combining, there are two possible operations that can be performed, which depend on the channel decoder implementation. First, if channel decoder is configured to perform hard decision decoding (HDD) [19], [20], [22], the combiner output $\mathbf{r}_\Sigma$ will be demodulated to produce the binary bits, deinterleaved, and then applied to the decoder. If the decoder is configured to perform soft decision decoding (SDD), the combiner output will be applied to the demodulator to compute the reliability factor for each bit, known as soft decisions, which are deinterleaved and then applied to the decoder to perform SDD [6], [21]. The remaining processes are identical to the first transmission session. The retransmission process is repeated until the packet becomes error free, or the maximum number of transmissions C is reached.

1) Combining with different modulations schemes or orders: Adaptive modulation is a key tool to improve the performance of wireless communications systems, where the modulation type or order can be changed based on the channel conditions. In ARQ, the modulation used to transmit a particular packet may become a limiting factor for using adaptive modulation in the subsequent transmissions of that packet. Such limitation is due to the challenge of combining information symbols of different modulation schemes or modulation orders. To overcome this problem, researchers proposed using bit-level combining [91]–[93] for cooperative communications systems with relays. In this approach, the LLR for each bit from each transmission is computed, and then combining is performed at the bit level. It is worth noting that bit-level LLR is typically used when higher order modulations are used with binary SDD where the reliability of each bit can be computed even though the modulation is non-binary [94].

2) Optimal combining for HARQ: As described in Sec. III-A, the receiver should buffer all transmissions of a packet that fails the error detection process, and combine such transmissions to improve the SNR of the signal. The combining using CC is applied because it is equivalent to MRC, which is a widely used approach. Nevertheless, when combining multiple retransmissions, the last transmission is combined with all previous transmissions before it is being decoded and checked for errors. Therefore, we are combining packets that are confirmed to be erroneous with the last packet whose status is still unknown. In such scenarios, the useful information about the erroneous packets is ignored. Therefore, Long *et al.* [90] proposed a new combining scheme that outperforms CC by assigning an additional weighing factor for the combining process. The authors of [90] have considered an ARQ with a maximum of two transmissions, and thus, the combining scheme of two symbols received from two transmissions is given by

$$r_\Sigma = \frac{\omega H_1^* r_1 + H_2^* r_2}{|\omega H_1|^2 + |H_2|^2} \tag{4}$$

where ω is the combining factor. Based on the results in [90], the optimum value of the combining factor $\omega < 1$, and the impact of the optimal combining is more significant for small packet lengths.

3) Optimal packet length: The efficiency of the HARQ-I highly depends on the packet length. If the packet length is high, the retransmission process will be very costly in terms of bandwidth and energy consumption. On the other hand, transmitting short packets increases the packet overhead with respect to the data part of the packet. For example, the ratio $k/(k + l_{\mathrm{ed}})$ indicates the impact of the error detection bits on the system throughput, which can be substantial for short packets. Therefore, extensive research efforts have focused on optimizing the packet length of ARQ applications [95]–[99].

B. HARQ Type-II: Incremental Redundancy

At low SNRs, the throughput of HARQ-I systems is remarkably higher than basic HARQ systems as the error correction performed by the FEC and CC can reduce the number of retransmissions significantly. However, at high SNRs, the parity bits introduced for error correction are not often utilized, which reduces the throughput as compared to conventional HARQ systems. Generally speaking, the throughput of HARQ-I is upper bounded by the code rate $\eta \leq k/n$. To overcome this problem, HARQ type-II (HARQ-II) transmits the parity bits only whenever they are needed, which is the reason for the name incremental redundancy. Therefore, in HARQ-II, the data transmitted over the first transmission and the consequent transmissions are generally different. One possible realization of HARQ-II is the following [12]:

1) The transmitter sends the uncoded data sequence $\mathbf{c}$ in (1) in the first transmission. Therefore $\mathbf{r}^{(1)} = \mathbf{H}^{(1)}\mathbf{s}|\mathbf{c}$.
2) The received sequence $\mathbf{r}^{(1)}$ is used to estimate the transmitted sequence $\hat{\mathbf{d}}$. If $\hat{\mathbf{d}} = \mathbf{d}$, the receiver sends an ACK, and the process for that packet is terminated. Otherwise, a NACK is sent.
3) When the NACK is received, the transmitter applies $\mathbf{c}$ to the channel encoder and computes the parity bits $\mathbf{p}$, which are modulated and transmitted without the data bits.
4) The receiver appends $\mathbf{r}^{(1)}$ and $\mathbf{r}^{(2)}$ to form one sequence of n samples that correspond to the received samples of a complete codeword. Consequently, the receiver applies the sequence $\left[\mathbf{r}^{(1)}\ \mathbf{r}^{(2)}\right]$ to the channel decoder, and extracts $\hat{\mathbf{d}}$. If $\hat{\mathbf{d}} = \mathbf{d}$, the receiver sends an ACK, and the session is terminated. Otherwise, A NACK is sent.
5) Once the second NACK is received, the transmitter sends $\mathbf{c}$ again, which is received as $\mathbf{r}^{(3)}$. Then, $\mathbf{r}^{(1)}$ and $\mathbf{r}^{(3)}$ are combined using CC to enhance the sequence SNR. Then $\mathbf{r}^{(2)}$ is appended to the combined sequence to form the sequence $\left[\mathcal{C}_{\mathcal{H}}\left[\mathbf{r}^{(1)}, \mathbf{r}^{(3)}\right]\ \ \mathbf{r}^{(2)}\right]$, which is applied to the channel decoder to compute $\hat{\mathbf{d}}$. If $\hat{\mathbf{d}} \neq \mathbf{d}$, the transmitter resends $\mathbf{p}$, that is received as $\mathbf{r}^{(4)}$, which is then combined with $\mathbf{r}^{(2)}$ to form $\left[\mathcal{C}_{\mathcal{H}}\left[\mathbf{r}^{(1)}, \mathbf{r}^{(3)}\right]\ \ \mathcal{C}_{\mathcal{H}}\left[\mathbf{r}^{(2)}, \mathbf{r}^{(4)}\right]\right]$.
6) If the data is still incorrect, the transmitter keeps sending $\mathbf{c}$ and $\mathbf{p}$ in an alternating manner, until the packet is detected correctly, or the maximum retransmission limit is reached.

HARQ-II based on punctured turbo codes: The incremental redundancy can be also realized by changing the code rate over multiple transmission. One of the common techniques to achieve this goal is puncturing [49], [78], [100]–[102]. The process in such systems can be summarized as follows:

1) The initial transmission may consist only of the information bits, if the packet passes the error check process, an ACK is transmitted and the process is stopped.
2) If the packet fails the error check, then an NACK is transmitted, and the transmitter encodes the information bits and use a puncturing pattern that produces a high code rate. The transmitter sends only the parity bits.
3) The parity bits will be augmented to the information bits of the initial transmission, and will be applied to the channel decoder. The output of the channel decoder is then check for errors, if it passes then an ACK is sent and the process stops. If the packet fails, a NACK is sent.
4) Once the NACK is received, some of the punctured bits will be sent, and the receiver appends them to the information bits to form a high code rate codeword. If the packet fails again, more punctured bits will be sent, which reduces the code rate and improves the error correction capability of the code.
5) If all punctured bits are sent, and a NACK is sent, the process is repeated, and in this case CC can be used to improve the SNR of the bits with multiple transmissions.

This approach can improve the system throughput, but the delay may increase and the number of maximum transmissions should be increased as well to allow correct reception at the end of the process.

C. HARQ Type-III

In general, HARQ Type III, HARQ-III, is similar to HARQ-II except that user data and parity bits are included in every retransmission [103]. Therefore, the information bits can be extracted in each transmission independently of other transmissions. This feature is called self-decodability. A common approach to realize HARQ-III systems is through complementary punctured convolutional codes [104]. A set of punctured convolutional codes derived from the same original low-rate code are said to be complementary if they have equivalent distance properties and their combination yields at least the original low-rate code. Other FEC have been considered as well [105].

IV. Key Performance Measures

The HARQ process is evaluated using several performance metrics, including but not limited to throughput, spectral efficiency, reliability, outage probability, average retransmissions, latency, and delay. Although these metrics are interdependent from each other, each metric will generally give different insight about the system performance. For example, reliability in terms of PER or BER needs to be minimized, where generally, the error rate decreases as the SNR increases. Moreover, there is a strong relation between SNR and throughput which is proportional to SNR. The throughput and reliability have supporting relationship, i.e. the throughput can be maximized and at the same time, the reliability can be minimized. Further, in a HARQ employed system, the reliability needs to be assured with maximum allowable retransmissions. But when the

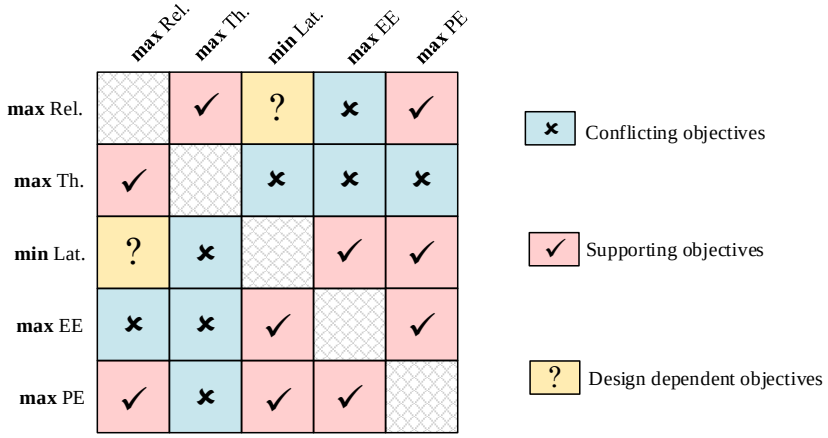

Fig. 2: Relationship among HARQ performance metrics. Rel.: Reliability, Th.: Throughput, Lat.: Latency, EE: Energy efficiency, PE: Power efficiency.

allowable retransmissions are truncated, the outage probability may increase, which shows a conflicting relationship between these two metrics. The relationship between various other HARQ performance metrics is depicted in Fig. 2. It is shown that some of the metrics are in supportive relationship with each other, whereas others are in conflicting relationship. The relationship of few metrics is design dependent.

Some commonly adopted performance metrics are summarized in Tab. III and will be discussed in detail in this section.

A. Throughput and Spectral Efficiency

Although ARQ and HARQ are effective techniques for reliable data transmission over noisy channels, they also increase the delay to the transmitted data because of re-transmission process. Such delays lead to low throughput and spectrally inefficient communication. To reduce the delays caused by re-transmissions, several efficient modifications have been proposed in the literature. These modifications have been validated through different performance metrics. Generally, the throughput of HARQ based systems is evaluated in terms of bits per second (bps), packets per second (pps), frames per second (fps), and bps per hertz, which is the spectral efficiency. The throughput is also evaluated in terms of the number of maximum re-transmissions required for reliable communication. In general, the throughput of HARQ systems is defined as,

$$\eta = \frac{\mathbb{E}[\bar{Z}]}{Z^{max}} \tag{18}$$

where $\mathbb{E}[\bar{Z}]$ is the expected number of correctly received packets, and Z^{max} is the total number of packets generated at the source. The spectral efficiency is the ratio of number of correctly received packets to the total number of transmitted packets in unit time. The metric defined in (18) is generally used to compute the throughput of various communication systems. The commonly used metrics for throughput are summarized in Table III.

In [106], a performance metric (5) to compute the throughput of a cooperative HARQ system is proposed. The network comprises single source, a single relay and a single destination. The throughput of the system is defined as the ratio between the expected number of correctly received bits $\mathbb{E}[\bar{B}]$ to the expected number of channel uses $\mathbb{E}[\bar{N}_s]$ given a particular maximum number of retransmissions. A channel use is defined as the transmission resources required to transmit one information symbol. It is shown that the outage probability is inversely proportional to the overall system throughput, i.e., when the outage probability increases, the system throughput decreases. In [107], [108], a throughout metric (6) is investigated for SW-HARQ for a CR systems. The throughput for a CR system is defined as the total number of packets successfully decoded by the CR receiver in one time slot T. In CR system the cognitive user senses the channel of the primary user and transmits only if the primary user channel is vacant. The throughput in (6) is the ratio between total number of slots required to transmit the desired packets scaled with the propagation time of one time slot, $T = T_d + T_s$, where T_d is the transmission time interval (TTI) and T_s is the channel sensing time. In this case, if the cognitive user obtains the primary user channel in consecutive time slots, the throughput will increase, and vice versa. Similarly, throughput metric (7) for a CR system with GBN-HARQ is proposed in [109]–[111]. In GBN-HARQ, the CR transmitter transmits multiple packets in a time slot without waiting for the acknowledgment from the CR receiver. Once again, if the CR transmitter finds a vacant primary user channel for consecutive time slots, the throughput will increase. If the number of transmissions in one time-slot is increased, the total number of time slots required for successful packet transmission is reduced.

A normalized sum-rate for a cooperative multi-source, multi-relays, single destination cooperative network is investigated in [112], as shown in (8). The overall throughput of such systems depends on the individual throughput of the two phases, where the first phase is when the sources broadcast the packets and the relays and destination receive the packets. This is termed as the initial throughput of each source, the initial throughput increases if, a) the number of transmitted packets increases through less number of channel utilization, or b) the number of transmitted packets increases while the channel

TABLE III: Commonly used performance metrics.

Metrics	Ref.	Expression	Remarks
Throughput	[106]	$\eta = \frac{\mathbb{E}[\bar{B}]}{\mathbb{E}[\bar{N}_s]} = \frac{K(1-P_{out})}{\mathbb{E}[\bar{N}_s]}$ (5)	Throughput of a cooperative HARQ with single source, single relay and single destination
	[107], [108]	$\eta = \frac{Z^{max}}{N_t} \times \frac{T_p}{T}$ (6)	Throughput of cognitive SW-HARQ
	[109]–[111]	$\eta = \frac{Z^{max}}{N_t Z^{tx}}$ (7)	Throughput of cognitive GBN-HARQ
	[112]	$\eta = \frac{1}{S + \alpha \mathbb{E}\left(T_{max}\right)} \sum_{s \in \mathcal{S}} (1 - P_{out}^{s})$ (8)	Sum-rate of cooperative HARQ with multi-source, multi-relay
	[113]	$\eta = R \frac{1 - PER}{\bar{C}}$ (9)	Throughput for HARQ assisted demodulate and forward relaying network
	[114]	$\eta = \frac{1 - P_{out}^{u}}{C+1} R^{u}, \ u \in \{p, s\}$ (10)	Throughput of cooperative CR network, exploiting HARQ to improve secondary user performance
	[115]	$\eta = \frac{R}{C}\left(1 - P_{out}\right)$ (11)	Throughput with optimum energy allocation in ARQ assisted wireless system
	[116]	$\eta = \frac{\mathcal{B}}{\Upsilon}\left(1 - PER\right)$ (12)	Throughput for multi-casting WSNs with HARQ
	[117]	$\eta = \frac{\mathcal{B}/\bar{N}_s}{\sum_{t=1}^{C-1} P_{out}^{t}}$ (13)	Long-term average throughput
Avg. retransmissions	[118]	$\bar{C} = 1 + \sum_{\ell=1}^{\infty} \mathbb{E}\Big\{ \prod_{t=1}^{\ell} \mathbb{P}_f(t) \Big\}$ (14)	The average number of retransmissions in HARQ-incremental redundancy (IR) system
	[113]	$\bar{C} = \frac{1}{Z^{max}} \sum_{t=1}^{C} t \bar{z}_t^{max}$ (15)	The average number of retransmissions in HARQ assisted demodulate and forward relaying network
	[119]	$\bar{C} = 1 + \sum_{t=1}^{C-1} \mathbb{P}_f(t)$ (16)	The average number of HARQ retransmissions for URLLC services
Resources	[120]	$\Upsilon = (\mathcal{P} \cdot n_s) \cdot n_t$ (17)	Resource time and frequency allocation scheme

utilization remains fixed, or c) the number of transmitted packets is not increased, but the channel utilization is reduced. The throughput of each source decreases if the number of retransmissions increases in the second round. The throughput metrics (9) and (12) with PER are given in [114], [115]. The metric (9) is evaluated using the input code rate, whereas (12) uses the achieved rate through the number of successful received packets and the total number of transmitted packets. In both cases, a larger PER value imposes adverse affects on the achieved system throughput.

The throughput for cooperative CR networks (10) and general wireless systems with optimum energy allocation (11), is given in [113], [116], respectively. The throughput increases with a larger code rate and less number of transmissions. Moreover, a larger value of outage probability sinks the overall throughput. Similarly, another important parameter that affects the overall throughput is the PER, or equivalently, the frame error rate (FER) and BER. In (13) the long-term average throughput metric is presented, where $\mathbb{B}$ is packet length and $\bar{N}_s$ is the channel uses for transmission of the packet.

B. Reliability

Reliability is generally defined as the capability of a system to perform consistently well. In the wireless networks, reliability is defined as the probability of successful delivery of a

packet within allowable number of transmissions [121]. At the network level, the reliability is evaluated as a packet delivery ratio [122]. At the application level, reliability is described as the rate of successfully delivered bits, frames or packets.

1) Error rate: The error rate is considered as one of the decisive metrics to evaluate a system's performance in terms of reliability. Error rate can be defined at different levels, e.g. FER, PER, bLock error rate (BLER), BER, and symbol error rate (SER). These various error rate metrics are used for wide range of applications. For example, the FER is mostly adopted to measure the reliability of video communication systems. Similarly, BLER, PER and BER are generic error rate metrics and are used in diverse applications, and have almost similar meanings. The difference between BLER/PER and BER comes from the expected values they acquire. A block or packet is composed of a number of bits, and therefore, for the same SNR value the acquired BLER/PER value is generously larger than BER values. The terms BLER and PER are sometimes used interchangeably by the researchers. The reliability metrics introduced in this section can be defined as, the ratio of the number of bits/blocks/packets received with errors to the total number of bits/blocks/packets transmitted over a communication channel. In this context, the error rate can be seen as the reciprocal of the attained throughput or spectral efficiency. The error metrics influence the average number of re-transmissions for the successful delivery of the data. Moreover, it leads to affect the required energy for the successful transmission [128].

In Table IV, the relation between the error rate metric with the target throughput is provided in (20), which is derived from (12). The error rate metrics introduced in this section are typically used as a threshold for a HARQ system's performance. For example, in [63] the BER is used as a threshold for throughput optimization of a selective HARQ system, whereas tight thresholds for BER are opted in [129] for selective CC methods. A tight BER upper bound is derived in [130] for Nakagami-m fading environment for any constellations pair. It is claimed to be a BER upper bound for coded HARQ systems. Another important closed-form BER expression for different binary modulation schemes with dual branch selection combining is derived in [131]. The closed-form BER expression is not specific to HARQ systems, but to any general communication system. A conditional BER for the HARQ system with various combining schemes is analyzed in [90]. From the theoretical approximated BER expressions, it is concluded that conventional MRC is not optimum for HARQ. The performance of MRC becomes even worse for short packet communications, or if the channel for the re-transmitted packets is not as good as it was during the initial transmission.

Typically, a target BLER, PER or FER is pre-decided for a particular application, mostly to achieve wide range of objectives including throughput, spectral efficiency and energy efficiency. Therefore, the target error rates are pre-computed before the HARQ operation. For example, the influence of considering the FER while computing the effective rates is studied in [132], where it is shown that the performance of the HARQ system significantly improves if the effective rate computation considers the threshold FER. A conditional FER of HARQ system is determined by applying the condition on the previous erroneous transmissions is proposed in [133]. The conditional FER is evaluated while taking into account the effective SNR and rate values during each transmission. The quality of frames received in the past highly affects the proposed conditional FER. In [134], the FER is approximated using a threshold-based method. The proposed FER approximation simplifies the physical layer operations with few parameters. It is further testified that the proposed approximation accurately predicts the FER performance for a wide range of receiving systems. The significance of FER for HARQ performance evaluation is further highlighted in [135], where the FER is evaluated for a HARQ system with systematic polar codes, turbo codes, LDPC and convolutional codes. Further, the performance of a novel polar coded HARQ system is validated in terms of FER in [74], where it is demonstrated that FER has a substantial impact on the system throughput and average number of transmissions.

2) Outage probability: It is one of the important metrics to evaluate the performance of HARQ systems. A transmission is said to be in an outage if the packet is not successfully decoded after the maximum number of allowed transmissions is reached. Reliable data communications require significantly small outage probability. For example, an outage probability of around 10^{-9} is expected to address the low latency communications [136]. Mathematically, outage probability after ℓ^{th} transmissions can be written as,

$$P_{out}^{\ell} = \mathbb{P}(R_c^{\ell} < R) \tag{29}$$

where R is the code-rate, whereas R_c^{ℓ} is the rate achieved after ℓ^{th} round, also termed as effective rate. Outage probability is also measured through various other parameters such as signal to interference plus noise ratio (SINR) [123], [124], [127] and accumulated mutual information (AMI) [125], [126]. Outage probability is also related to the achieved transmission rate. Therefore, various parameters such as the the transmission power, and bandwidth can be adapted to obtain a better rate. High rates that are realized with an increase in transmission power or code-word length eventually decreases the outage probability. Moreover, inclusion of diversity techniques at various levels mitigate the channel effects on the data transmission. In HARQ systems, the combining schemes have shown promising improvements in terms of outage probability. The combining process amplifies the mutual information after each packet retransmission.

C. Energy and Power Efficiency

Energy and power efficiency are widely used performance metrics in communications systems. For a normalized symbol period of unity, both metrics become equivalent. Nevertheless, the two metrics are effectively different may lead to different problem formulation and solution. For the sake of consistency, the two metrics are treated separately based on the system model adopted by the authors.

TABLE IV: Commonly used performance metrics.

Metrics	Ref.	Expression	Remarks
Reliability	[113]	$PER = 1 - \frac{\bar{Z}}{Z^{max}}$ (19)	Packet error rate for a given throughput for HARQ assisted demodulate and forward relaying network
	[116]	$PER = 1 - \frac{\eta \Upsilon}{\mathcal{B}}$ (20)	Packet error rate for a given throughput in multi-casting WSNs with HARQ
Outage Probability	[123]	$P_{out}^{1,\ell} = \mathbb{P}(\bar{\gamma}_{1,\ell}^{x_{1,\ell}} < \gamma_1)$ (21) $P_{out}^{2,\ell} = 1 - \mathbb{P}(\bar{\gamma}_{2,i}^{x_{1,i}} \geq \gamma_1)$, $\forall i \in \{1, \cdots, \ell\}, \sum_{i=1}^{\ell} \bar{\gamma}_{2,i}^{x_{2,i}} \geq \gamma_2$ (22)	Outage probability of a NOMA system after t HARQ rounds
	[124]	$P_{out}^{cc}(\ell) = \mathbb{P}\left(\bar{\gamma}_\ell \leq \frac{2^R - 1}{\ell}\right)$ (23) $P_{out}^{IR}(\ell) = \mathbb{P}\left(\bar{\gamma}_\ell \leq 2^{\frac{R}{\ell}} - 1\right)$ (24)	Outage probability in satellite-terrestrial relay network with HARQ-CC and HARQ-IR
	[125], [126]	$P_{out}^{1,\ell} = \mathbb{P}\left(I_{1\to1,\ell} < R_1 \cup I_{1\to2,\ell} < R_2\right)$ (25) $P_{out}^{2,\ell} = \mathbb{P}\left(I_{2\to2,\ell} < R_2\right)$ (26)	Outage probability in HARQ-CC aided NOMA system
	[127]	$P_{out} = \Pr\left(\log\left(1 + p \cdot h\right) < R\right) = 1 + e^{-\frac{e^R - 1}{p}}$ (27)	Outage probability in HARQ aided URLLC system through optimum power allocation
	[115]	$P_{out} = 1 - \frac{\eta C}{R}$ (28)	Outage probability with a given throughput with optimum energy allocation in ARQ assisted wireless system

1) Energy Efficiency: The total energy required to transmit Z^{max} packets in a HARQ system is given in [137],

$$\begin{aligned} E_{tot} &= E_{tx} + E_{rx} + \tilde{E}_{tx} + \tilde{E}_{rx} \\ &= Z^{max} P_t T_p + \mathbb{E}[\bar{Z}](jE_{tx}) + Z^{max} P_t T_d + j\tilde{E}_{tx} \end{aligned} \quad (30)$$

where E_{tx} and E_{rx} correspond to the energy consumed during the transmission and reception of the data packets, $\tilde{E}_{tx}$ and $\tilde{E}_{rx}$ is the energy consumption during the transmission and reception of the feedback messages. Moreover, $\mathbb{E}[\bar{Z}]$ is the expected number of correctly received packets, T_p is the packet transmission time, T_d is the ACK/NACK transmission time. The energy consumed at the receiver side is less than that at transmitter side by a factor $j \in (0, 1)$. The system in [137] is proposed for underwater sensor networks where it is shown that increasing the hop distance and the number of hops may improve the energy efficiency.

In [138], the energy consumption for a relaying system employing HARQ was studied. The energy consumption is investigated in terms of the total number of data transmissions along with outage probability. A scheme called transmission number relaying (TNR) is proposed, which aims at reducing the total number of transmissions, and hence, reducing the total consumed energy. It is concluded that the energy consumption largely depends on the number of transmissions and the outage probability. Hence, a metric denoted as transmissions per message (TM) is derived as,

$$TM = \sum_{i=1}^{2C-1} \frac{i\mathbb{P}[C_r = i]}{1 - P_{out}} \quad (31)$$

where C is maximum allowable transmissions and C_r is the total number of transmissions of a selected relay. Therefore, (31) can be seen as a ratio of total transmissions to the successfully delivered messages. It is depicted from the simulation results that TM decreases by increasing SNR. Another important energy efficiency metric is proposed in [139], where the multicasting network energy efficiency is characterized in terms of network efficiency and SNR outage probability. Further, an outage-constrained energy minimization and energy-constrained outage minimization problem is formulated as function of ferrying distance. The metric is give as,

$$\zeta = \frac{R}{BE_{tot}}(1 - P_{out}) \quad (32)$$

where E_{tot} is the overall end-to-end energy consumption and B is the bandwidth. In this model, the energy efficiency is

derived as a function of the distance from the user to the base station (BS). Further, the overall energy efficiency is presented as,

$$\zeta = \sum_{s \in S} \frac{R(R_x - \sum_{i=1}^{R_x} s_i)}{(2^{R_x} - 1)BE_{tot}} P_{out} \tag{33}$$

where R_x are total number of receivers, and $s_i = 1$ for a failed transmission and equal to 0 if the transmission is successful for i^{th} receiver. The term $\frac{R(R_x - \sum_{i=1}^{R_x} s_i)}{B}$ represents the total number of successfully received bits. S represents the set containing all possible combinations of successful or erroneous transmissions for each receiver, i.e., 2^{R_x} combinations. For example, in case of $R_x = 3$, the set S contains 8 entries, i.e., $S \in \{000, 001, \cdots, 111\}$, where in each combination the first bit represents the reception status of the first receiver, the second bit indicates the reception of the second receiver, whereas the last bit depicts the status of the third receiver. For example, a combination 010 indicates that the packet is successfully received by the first and third receiver, whereas an erroneous packet is received by the second receiver.

2) Power Efficiency: Transmission power plays a vital role in achieving a desired target error rate, where increasing the power reduces the error rate, and consequently the latency caused by the retransmission process. In the case of high target error rates, the system can transmit limited number of packets due to the large queuing latency. Therefore, these conflicting constraints and limitations further complicate the overall system [140]. Moreover, a significant performance improvement can be realized with an optimal allocation of transmission power, especially in low and medium SNRs [115]. It is therefore necessary to find optimal power allocations for a system to get rid of maximum possible hinges. Optimal average power allocation for URLLC is derived in [127] and given as,

$$p_{avg} = \sum_{\ell=1}^{C} p_\ell P_{out,\ell-1} \tag{34}$$

where p_ℓ is the transmission power in ℓ^{th} HARQ round, whereas $P_{out,\ell-1}$ is outage probability till $(\ell-1)^{\text{th}}$ HARQ rounds. An optimization problem is modeled to minimize the average power given the outage probability does not exceed a given threshold. It is observed that the achieved average power decreases if the outage probability threshold increases. A transmit power policy for multiple nodes is presented in [141] as,

$$p_n = \frac{E^n_{tx,\ell}}{T} \tag{35}$$

where p_n represents the power of n^{th} node, $E^n_{tx,\ell}$ amount of energy consumed by nth node during ℓ^{th} transmission attempt, whereas T is time-slot duration. This power policy can easily be implemented if the nodes attain the information of their own battery state, which reduces any further overhead to implement the power policy. The simulation results validate the performance of the proposed power policy over equal power policy.

D. Delay

The packet transmission delay adversely affects the overall system throughput. Typically, the overall packet transmission delay constitutes the first transmission time coupled with the time required for retransmissions. Keeping in view the stringent delay requirements in the 5G networks, extensive research has been carried out to address the delay sensitivity of HARQ systems.

In [108], [110], [111], [142], an average packet delay for CR systems in terms of the throughput, packets per time slot, is given by,

$$T_D = \frac{T_s + Z^{tx}}{\eta} \tag{36}$$

where η is the system throughput, T_s is the time for channel sensing during one time slot, and Z^{tx} is the number of packets transmitted during one time slot. Another metric for average packet delay is provided in [107], [109] as,

$$T_D = \frac{N_t(T_s + T_d)}{Z^{max}} \tag{37}$$

where N_t is the total number of time slots to transmit total Z^{max} packets. T_s is the channel sensing time and T_d is the round-trip time in one time slot. Another key parameter to estimate the delay is termed as delay outage is studied in [143], [144]. A system is in outage if the average transmission delay exceeds some pre-defined threshold. The outage probability is given as,

$$P_{out} = \mathbb{P}(T_D > T_{thr}) \tag{38}$$

where T_D is the average packet delay or round trip time and T_{thr} is the predefined delay threshold. The average packet delay can be obtained as,

$$T_D = \sum_{i=1}^{\ell} i(1 - \mathbb{P}_f(\ell)) \tag{39}$$

where $\mathbb{P}_f(\ell)$ is the probability that the packet is not decoded after the ℓ^{th} round. In the transmission system, data coding part stimulates major contributions in terms of delay. The transmission delay owing to encoding and decoding is studied in [145]. In case of consecutive transmissions without additional delay, the overall transmission delay is expressed as $C \times T_p$, where C is the maximum number of transmissions and T_p is the duration to transmit one transport block. Therefore, to enable the transmission of new packet soon after the end of current transport block, the encoding delay e_D and decoding delay d_D are constrained as,

$$e_D < T_p \tag{40a}$$

$$d_D < T_p. \tag{40b}$$

A metric for effective delay in a multi-user HARQ-IR employing system is proposed in [146], given as,

$$T_D = \frac{\bar{C}}{R} \tag{41}$$

where $\bar{C}$ is the average number of transmissions and R is the code rate.

E. Resources

Wireless communication infrastructure is mostly confined under limited resources. For example, a wireless sensor network (WSN) or IoT node can be placed at a remote location with limited power resource. Other examples of the resources in a wireless communication system are available spectrum, time slots, and power. Therefore, the main intent of efficient resource allocation is to satisfy the end-users' requirements. For example, the power allocation to admit a maximum number of users, or to maximize the overall rate. Moreover, in URLLC the QoS requirements are defined as the specified number of bits successfully delivered within a pre-defined end-to-end delay and error probability. In [147], the impact of bandwidth resource on the low latency reliable transmission following retransmission is investigated.

In HARQ assisted wireless systems, the importance of intelligently allocating the available resources among the stakeholders becomes even more crucial. An optimal resource time and frequency allocation scheme is proposed in [120]. The resources are allocated such that the number of utilized channel complies with the number of transmitted symbols in a TTI. Therefore, for transmission of n_t symbols in a TTI and n_s subcarriers in a physical resource block (PRB), the least number of required resources are calculated in (17).

V. Advanced HARQ Techniques

The basic HARQ protocols have been considered widely by the researchers where several variations and modifications were adopted for reliable and low latency communications. The foundation of the improved retransmission methods are the basic HARQ protocols described in Sections II and III. In this section, widely employed HARQ variants are reviewed in detail.

A. Adaptive HARQ

The time-varying nature of radio channels provides a strong motivation to design adaptive wireless communications systems that have various components that can be adapted almost in real-time. The main system components that are typically considered for adaptation are the modulation scheme/order, transmission power, maximum number of retransmissions, number of allocated channels, bandwidth, and code rate. The adaptation process usually aims at improving the system performance by increasing the throughput and energy efficiency (EE) while reducing the delay and system complexity. In the literature, it has been shown that the overall performance of HARQ systems can be substantially improved through optimal tuning of the system parameters. A HARQ system with the dynamic parameters is termed as an adaptive HARQ. A detailed survey of the state-of-the-art adaptive HARQ methods is shown in Tab. V.

1) Coding (rate control): In code adaptive HARQ systems, the transmission rates are adjusted for every new transmission, to upgrade the system's performance in terms of throughput and spectral efficiency. In such schemes, the receiver provides the transmitter with a detailed feedback, often termed as "intelligent NACK" [149]. The receiver attempts to decode

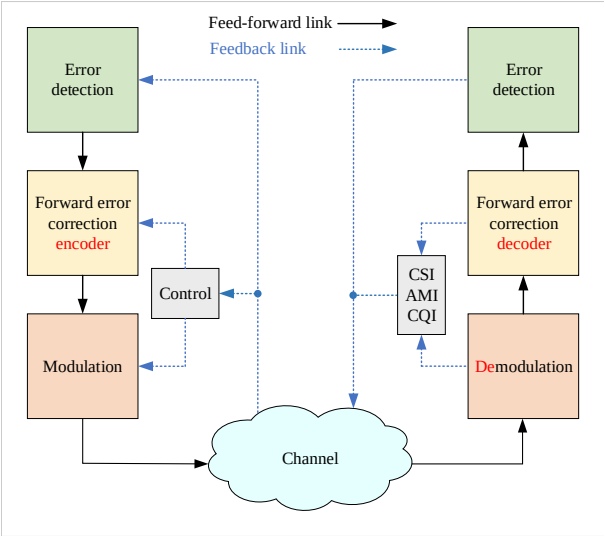

Fig. 3: Basic diagram of adaptive HARQ.

the received packet after the demodulation step, and extracts diverse information related to the channel condition and the transmitted data. This information includes, but is not limited to, the CSI, CQI, and AMI. If the packet decoding is not successful, the channel and data-related information is concatenated with the NACK and transmitted back to the transmitter. Upon receiving the "intelligent NACK", the rate adaptation mechanism is exploited by the transmitter using the feedback information to determine the optimum rate for the following transmission [150]. The overall code adaptive HARQ system is shown in Fig. 3. Following are the steps involved in a code adaptive HARQ systems:

- The transmitter appends the error detection (ED) and FEC bits with the information block, as discussed in Section II, where a packet $\mathbf{z}^{(i)}$ contains N modulated sequences or symbols. The initial code rate is given as $R_o = (k + l_{\text{ed}})/n$.
- A subset of codewords is generated using the initial code rate R_o.
- The receiver demodulates the received sub-codeword and then initiates the decoding process aided by the FEC decoder. Finally, the ED process is applied to detect errors in the received packet. A packet is assumed to be decoded successfully if the accumulated achievable rate is greater than or equal to the mother code rate [152].
- The receiver uses the received information, even with decoding failure, to evaluate various channel estimation and information parameters, mainly AMI and CSI. For example, the AMI is expressed as [149], [150],

$$\mathbf{r}_\ell = \sum_{t=1}^{\ell} \mathbf{z}_t R_t \tag{42}$$

where $\mathbf{r}_\ell$ is the received packet during ℓ^{th} round, whereas $\mathbf{z}_t$ and R_t are the transmitted packet and the adopted code rate during the t^{th} transmission round. Likewise, the CSI is estimated using the SNR of the received information.

TABLE V: Literature review of adaptive HARQ: Power (Pow.), packet length (Len.), modulation scheme (Mod.), number of retransmissions (RTx).

Ref.	Objective	Adaptation					Remarks
		Pow.	Len.	Rate	Mod.	RTx	
[148]	EE	✓	✓				Adaptation using outdated channel state information (CSI) and validated through persistent and truncated HARQ-IR
[149]	Throughput			✓	✓	✓	Employed channel code with puncturing and repetition coding. Validated through fixed-length retransmissions
[150]	Throughput			✓	✓		Based on average channel statistics and employed with HARQ-IR
[151]	Throughput			✓			Outdated CSI is utilized for adapting HARQ-IR based retransmissions
[152]	Throughput		✓	✓			Adaptive HARQ-CC retransmission using AMI
[153]	Throughput			✓			Instantaneous CSI is not available to the transmitter. Only outdated CSI is used to adapt HARQ-IR
[154]	Delay			✓	✓		Channel quality indicator (CQI) reports are utilized for HARQ-IR adaptation
[155]	Spectral efficiency		✓	✓			Outdated CSI feedback information is used to adapt un-truncated HARQ-IR
[156]	Spectral efficiency		✓	✓			Dynamic resource allocation based on outdated CSI feedback information from the receiver to adapt un-truncated HARQ-IR
[157], [158]	Power	✓					Adapted type-I and -II HARQ
[159]	Throughput	✓					Transmit power is adapted in HARQ-CC with CSI feedback information, while the throughput remains almost unchanged for wide range of SNR
[160]	Delay					✓	No. of retransmitted frames are adapted according to the video content and playback deadline. CSI and frame size is used to estimate frames transmission
[161]	Throughput, delay		✓				The AMI is used to estimate the number of bits necessary to decode the message at next transmission in a HARQ-IR system
[162]	Reliability					✓	No. of retransmissions are adapted by prioritizing frames in a group of pictures (GoP)
[163]	Throughput		✓				Estimates the contributions of each candidate packet to the marginal recovery on the erroneous packets and then transmits the packet which generates the most marginal recovery
[164]	Throughput			✓			Space-time block codes are used to mitigate deep fading, but in case of a NACK , a pre-coded matrix is selected based on minimum mean squared error (MSE) after combining
[165]	Reliability		✓				inter-carrier interference (ICI)-cancelled retransmissions using the repetition coding scheme are opted after initial retransmission in HARQ-IR
[166], [167]	Throughput, packet length			✓			An additional convergence NACK is introduced to specify the actual positions of the sub-blocks on which the additional parity bits will be concentrated
[168]	Reliability			✓	✓		Same puncturing pattern is used in HARQ-CC, whereas different puncturing patterns are applied to the successive (re)transmissions
[169]	Throughput	✓					HARQ gains with power control using partial CSI
[170]	Delay			✓			Analysis of SR-HARQ protocol in a multi-rate wireless network in terms of delay statistics

The CQI reports are also exploited for dynamic resource allocation [154], which is considered as a part of CSI.

- The estimated information about the channel and received data is then feedback along with the NACK. Based on the received intelligent information, the transmitter decides how to tailor the code rate for the following transmission.
- The transmitter applies the puncturing technique to increase or decrease the code rate for the following transmission, based on the received channel and data information. In HARQ-CC typically same puncturing pattern is used, whereas in case of HARQ-IR for all successive (re)transmissions different puncturing patterns are applied [168].
- The process is repeated unless all the codewords are successfully delivered or the maximum number of transmissions is reached.

In general, the transmitter is unaware of the channel condition during the initial transmission. Moreover, it perceives the channel condition exploiting the CSI which is sent by the receiver via feedback message after the first transmission. Because of the transmission delay, very often the CSI information becomes partially or fully outdated [149]–[151], [153], [155], [156], [163], which restricts the error-free transmissions.

Several code adaptive HARQ techniques were proposed in the literature. A HARQ proposed in [149] considers puncturing and repetition coding methods to adapt the transmission rates, whereas the same techniques are opted in [150] with non-binary LDPC codes using the average CSI. In [151], [153], [163], the transmission rate is adapted based on the number of codewords transmitted during a particular round. Moreover, the number of codewords to be transmitted is estimated from the residual error sent by the receiver. Therefore, during each round packets are encoded with a variable rate to generate a different number of codewords. AMI is considered as the feedback information in [152]. The transmitter adapts the transmission rate to let the AMI exceeds the code rate.

The puncturing technique is widely preferred in rate-adaptive HARQ. To accelerate the retransmission system to boost the throughput of URLLC system, different pre-encoded redundancy versions, using different systematic and parity bits, are cached in the transmission buffers. Furthermore, the CQI report is exploited to select the best-suited redundancy version by the transmitter [154]. A concept of residual rate is proposed in [155], [156], which is defined as the minimum mutual information required for successful decoding. The residual rate after ℓ^{th} transmission round is given as,

$$R'_\ell = R_{\ell-1} - I_\ell \Upsilon_\ell \tag{43}$$

where $R_{\ell-1}$ and I_ℓ are the code rate and the mutual information during ℓ^{th} transmission round, while Υ_ℓ are the allocated resources which are determined based on the residual rate information. The receiver returns the residual rate information, and in response the transmitter predicts rate adaptation, exploiting the received feedback information, to minimize the number of retransmissions.

2) Packet length: In length adaptive HARQ, the transmitter may vary the packet length during each round. The packet length is often referred to as the bandwidth. Typically in this adaptation scheme, the rate and transmission power are kept constant. If the packet is not successfully decoded by the receiver, it returns a NACK along with the channel information. The length adaptation mechanism is initiated by the transmitter by shrinking the packet length for the next transmission. It is highly likely that the packet is recovered within two transmissions because the redundancy versions can achieve extremely low PER for almost the same SNR [171]. Therefore, to get maximum out of total available bandwidth, it is better to reduce the packet length in the retransmissions according to the received CSI, AMI, or SNR information. The adaptive length HARQ proposed in [148] yields remarkable gain over the conventional HARQ methods. It is shown that in HARQ-IR, length adaptation plays a vital role in increasing the AMI and is given as,

$$I_\ell = I_{\ell-1} + n_\ell \log_2(1 + p_\ell \gamma_\ell) \tag{44}$$

where n_ℓ is the codeword length, whereas p_ℓ, and γ_ℓ are, respectively, the transmission power and the instantaneous nominal SNR experienced by the receiver with unit-power transmission during ℓ^{th} round. Therefore, (44) triggers an option to adapt the length of the codeword to increase the AMI, that along with a NACK is feedback by the receiver. The potential complication in the length adaptive HARQ lies in the signaling overhead. The length of the codeword generated by the relay is adapted in [152] to complete the deficient information at the destination for successful decoding after the retransmission. Moreover, partitioning of the parity bits are determined in a truncated type-II HARQ in [172]. Further, in [173] it is demonstrated that the HARQ-IR performance can be sufficiently improved with the optimal tuning of the block-length. Variable-length transmission for multiple users is not practically efficient, since the subsequent sub-codewords of shorter lengths might lead to bandwidth loss or the subsequent sub-codewords with larger lengths may result in collisions [163].

3) Power control: An energy-efficient communication can be realized by transmission power adaptation throughout the HARQ rounds. Generally, the instantaneous channel conditions are not known to the transmitters. In power controlled adaptive HARQ, the packet is encoded and then transmitted with an initial pre-defined power p_o and a fixed code rate R. The packet can be decoded at the receiver end provided the instantaneous channel gain is larger than the code rate R. Mathematically,

$$\begin{aligned} &\log(1 + h_\ell p_\ell) < R \\ &\implies h_\ell < \frac{e^R - 1}{p_\ell} \end{aligned} \tag{45}$$

where p_ℓ is the transmit power and h_ℓ is the instantaneous channel gain during ℓ^{th} HARQ round. If (45) is true, a NACK, together with the channel information h_ℓ is returned to the transmitter. In response to the feedback, the transmitter increases the transmit power to an amount that the instantaneous channel capacity exceeds the transmission rate.

Besides other modern technologies, the low-power IoT devices are expected to be an integral building block in the future wireless communications systems. Energy efficiency and optimal power control are desired for these low-power devices. Moreover, these devices and systems are likely to incorporate efficient retransmission mechanisms to ensure reliability. Adaptive power-controlled HARQ systems are foreseen to deal with these issues. Different power-controlled adaptive HARQ systems were proposed in the literature. Ref. [148] proposed a rate, length, and power controlled energy-efficient adaptive HARQ and the proposed adaptive solution outperforms the conventional HARQ with fixed codeword length and transmission power. In [157], the transmission and the decoding power is optimized in a HARQ employed communication system. In each HARQ round, the codeword is tailored under the power adaptive assumption to get distinct transmission powers. Moreover, it is shown that the power adaptation leads to a significant reduction in total average power, including the decoding and transmission power. Most of the research work on power adaptive HARQ systems relied on the feedback (CSI) message. In [159], power adaptation is proposed without CSI feedback information. Instead, the transmitter reduces the transmission power with a fixed rate taking into account the received SNR as a feedback message. The receiver estimates the probability of packet error, which is the function of SNR. In this way, a significant reduction in power consumption is observed. A power-controlled HARQ with partial CSI is proposed in [169]. The proposed protocol realizes the HARQ retransmission gains and the power control simultaneously. Notable improvements in throughput are observed through the proposed modifications in the power adaptive HARQ system.

4) Modulation schemes: The transmitter generates coded symbols, which are then mapped to some modulation symbols before the transmission, as depicted in Fig. 3. On the contrary, the receiver de-maps the received modulated channel symbols. Different modulation schemes are employed in the wireless communication systems, e.g., binary phase shift keying (BPSK), quadrature phase shift keying (QPSK),

and QAM. Usually, large constellations like 16- and 64-QAM are employed with channel conditions exhibiting high SNR. Whereas for channels with low SNR, robust modulation like QPSK is preferred [174]. The communication system with adaptive modulation HARQ typically switches between distinct modulation schemes. When the transmitter receives a CQI, AMI, or SNR feedback from the receiver, the transmitter adapts the packet length. The retransmission adaptation is typically based on the estimation of additional bits required for a successful reception. This process is repeated until the packets are successfully decoded by the receiver [149], [150]. In [154], the probability of decoding failure is exploited to adapt the rate and modulation of retransmissions, and is given as

$$\mathbb{P}_f(R) = \mathbb{P}(\text{AMI}(\text{SNR}) \leq R_t) \tag{46}$$

where, AMI is the accumulated mutual information of SNR values and R_t is the target rate given as

$$R_t = MR_c \tag{47}$$

where $M \in \{2, 4, 6\}$ is modulation index for QPSK, 16-QAM, and 64-QAM, respectively and R_c is the effective code rate. Therefore, to minimize the probability of decoding failure (46) a possible solution is to decrease the target rate R_t, which can be adjusted by decreasing the modulation index M in (47). After a certain number of retransmissions the AMI increases which eventually reduces the $\mathbb{P}_f$. In [168], [175], it is shown that the BS is responsible to adjust the modulation index based on the CQI feedback. Moreover, the effective rate of cognitive user employing adaptive M-QAM modulation is analyzed in [176]. Further, in [177] the channel SNRs are divided into non-overlapping intervals. The interval for feedback channel information, i.e., the achieved SNR, is determined and pre-defined modulation scheme for that particular interval is employed for the next retransmission. The same method of SNR partitioning into intervals is realized in [178], [179] and [180] to maintain the targeted BER and PER, respectively. In [119] the spectral efficiency maximization problem for low latency applications is formulated through the adaptive modulation and coding schemes.

5) Number of retransmissions: Another vital parameter that affects the performance of HARQ process is the maximum number of per packet retransmissions. In the HARQ systems, mostly the truncated retransmission model is followed, in which finite number of retransmissions are permitted for a packet. In conventional HARQ these finite number of retransmissions are pre-determined and are not adapted during the process. But in some of the adaptive HARQ systems, the number of transmission are also varied during the communication process. Usually in non-adaptive retransmission systems the delay is reduced but at the same time a lower throughput is achieved because of the additional bits to be transmitted. In [160] the adaptive number of retransmission technique is employed, where the transmitter estimates the possibility of retransmission considering a number of factors derived from feedback information. This information comprises the receiver's buffer starvation information, and the packet playback time information. Moreover, variable retransmissions concept is adopted from a range of discrete number of retransmissions set in [162], [181]. The packets are prioritized and are mapped to any of the values available in the retransmissions set. Typically, the most important packets are allotted more number of retransmissions. For example, in video communications the I-frames are given maximum priority and hence maximum number of retransmissions are assigned to I-frames. The eNodeB grants the number of retransmissions to each video packet according to its CQI information. Moreover, [182] proposed a retransmission policy in which only initial retransmission of P frames are permitted subject to successful reception of preceding I frame. A concept of time-stamped packets is given in [183], where each packet is time-stamped and the retransmission occurs only if the time-stamp is greater than the propagation delay. Similarly, no retransmission policy is adopted for too old and out of sequence packets. The average number of retransmissions for each packet is estimated in [119], [184] using the outage probability and block error rates.

B. Dynamic Resource Allocation

Typically, in conventional HARQ, a fixed number of resources is allocated for each transmission, i.e., termed as static resource allocation. This procedure aids a user to realize better throughput since it gets the predetermined maximum possible resources during each transmission. On the contrary, the constant resource allocation for each transmission might not be an efficient solution from the network perspective. Specifically, in the case of HARQ-IR retransmissions, the required mutual information can be aggregated with fewer resources. The process to determine an optimal number of resources, based on the feedback information, during each transmission is called dynamic resource allocation. The dynamic allocation offers to accommodate an increased number of users in the same network. Usually, a dynamic resource allocation policy is adopted to achieve a range of objectives, such as spectral efficiency, energy/power efficiency maximization, and transmission delay and latency minimization.

A rate-optimized HARQ with dynamic resource allocation is proposed in [156]. During each transmission, the L number of maximum symbols/codewords are transmitted, where the symbols are treated as the resources. Moreover, Υ_ℓ is the number of resources allocated during ℓ^{th} transmission. The AMI after soft combining up to ℓ transmissions is calculated as,

$$I_\ell = \frac{1}{L}\sum_{t=1}^{\ell} \Upsilon_t R_t = \sum_{t=1}^{\ell} \phi_t R_t \tag{48}$$

where R_t is the rate realized during ℓ^{th} transmission and ϕ_t is fraction of allocated resources. The packet is successfully decoded after ℓ^{th} transmission if $I_\ell > R_o$. Moreover, according to the dynamic resource allocation policy, it is highly recommended to allocate more resources to the transmission that is likely to fail, which contributes to reducing the average delay. On the contrary, fewer resources are recommended for the transmission with a high probability of successful decoding for a lower average mean AMI margin. The receiver feedbacks a NACK along with the residual rate information $\bar{R}_\ell$ till

ℓ^{th} round to the transmitter. Using (48), the residual rate is calculated as,

$$R'_\ell = R_o - \sum_{t=1}^{\ell} \phi_t R_t. \tag{49}$$

Based on the value of R'_ℓ, the transmitter dynamically selects the optimal number of resources to minimize the residual rate in the further transmissions. In [163], [185], a dynamic resource allocation policy is adopted taking the AMI into account. The HARQ scheme with dynamic resource allocation exhibits an improved performance compared to the conventional HARQ in terms of a trade-off between spectral efficiency and delay.

Another HARQ scheme employing dynamic resource allocation aiming to maximize overall throughput is proposed in [186]. The technique comprises two major steps, namely, time domain packet scheduling (TDPS) and block time-frequency domain packet scheduling (BTFDPS). At the beginning of transmission, the TDPS selects a subset of users together with the data to be transmitted. The data size is selected given the channel conditions and the buffer size at the transmitter and receiver end. Moreover, based on the channel information and on-going transmission experience, the time-frequency resources are dynamically allocated among the selected users during the BTFDPS phase. This dynamic allocation is aimed at minimizing the number of scheduling decisions and maximizing the throughput. During each transmission, the allocated power is minimized, leading to a significant reduction in transmission energy. A dynamic physical resource allocation scheme for HARQ-IR is proposed in [187] in time-varying channel conditions under the presence of interference. The proposed allocation policy is equally applicable for downlink (DL) and uplink (UL). The bandwidth is dynamically adapted, learning from the earlier transmissions of the same codeword.

C. Low-Complexity HARQ

Error correction and error detection are integral components of a conventional HARQ system, where typically the error correction part anticipates multiple iterations. This iterative decoding consumes significant time, and gives rise to substantial delays. The situation becomes even more challenging if the packet could not be successfully decoded after the maximum decoding iterations and a retransmission is requested. Therefore, the inherently iterative nature of FEC decelerates the overall HARQ system. To address these native concerns of the iterative decoding process, low-complexity HARQ is proposed in [13], [44]. Two novel concepts, namely early stop (ES) and deferred iterations (DI) are proposed where ES and DI are aimed at locating the suspension point and to delay the iterative process, respectively. Typically, the ES criterion is based on the prior target convergence threshold. Therefore, the ES stops the decoding process once the convergence threshold is reached. Another proposed method to reduce FEC complexity is to postpone the decoding iterations until, based on an estimate, the receiver is convinced that it has received plentiful information for successful decoding, otherwise retransmission is requested immediately. In this way, the unproductive iterations are excluded, and hence the complexity of the HARQ process is curtailed.

In [6], [42], [56], [188], the complexity of an HARQ system is reduced with an improvement in the error detection part. A novel concept of parity error checking (PEC) is applied as a replacement of conventional CRC method by exploiting the word-error detection capability of the TPC. Owing to this modification, TPC is jointly employed for error detection and error correction. At the transmitter end, the k information bits, concatenated with l_{ed} detection bits, are supplied to the TPC encoder to form a TPC codeword, attaching the l_{ec} parity bits to the string. On the receiver side, an ACK is sent back if the received packet is error-free, conversely, retransmission is requested. The retransmitted packet is then combined with the soft version of the stored erroneous packet. The combiner output matrix is fed into the TPC which performs a series of soft or hard decoding iterations. The TPC decoder iteratively decodes all the rows and columns of the combiner output matrix. The error detection process launches promptly after the decoding of the last column. During the detection process the syndromes of the first k, where k is the number of information bits, rows are checked and the packet is declared as error-free if all the syndromes are equal to zero. The syndromes of the last $n-k$ rows, corresponding to the parity bits, are essentially not required to be checked. Alternatively, instead of the syndrome check, the parity extended bits are exploited and the error detection is performed with parity checking for the first k rows. Significant complexity reduction is observed with the proposed detection technique as compared to the conventional CRC detection method.

Another low-complexity HARQ with LDPC is proposed in [189] using partial retransmission and diversity combining techniques. The retransmissions are split into equal size sub-packets. On each retransmission, $1/3$ of the original bits are sent with a magnitude factor of $\sqrt{3}$, which leads to a constant energy per transmission. Consequently, the packet length and the magnitude factor in each transmission remains the same. Moreover, the received sub-packets are combined after every retransmission. The same process is repeated during each retransmission.

D. Autonomous/Blind HARQ

In conventional HARQ, the feedback ACK/NACK message triggers an extra overhead to the routine communication. The situation deteriorates if the feedback channel is deeply faded and does not ensure reliable communication. This feedback overhead elevates further in case of multicast communication where multiple devices have to acknowledge simultaneously. On top of it, the situation becomes even more severe under weak channel conditions as it is highly unlikely to deliver the packet successfully during initial transmission and a chance for retransmission is anticipated.

Under such circumstances, it is a smart decision to retransmit promptly rather than waiting for an expected NACK. Therefore, to address the mentioned scenarios, a concept of autonomous HARQ, also referred to as blind HARQ, is proposed by the researchers. The autonomous characteristic

is a supplementary role to the conventional HARQ. In the autonomous HARQ a packet is transmitted a defined number of times, irrespective of the successful or failed reception. After the exhaustion of maximum autonomous transmissions, the system restores itself to the conventional HARQ.

In [116], an autonomous HARQ is proposed for a multicast communications system comprised of two receivers. An optimal number of maximum autonomous HARQ transmissions is determined to improve spectral efficiency. During the first ℓ transmissions, the packets are sent continually without waiting for an ACK. Moreover, an ACK is expected after the completion of autonomous transmissions, thereupon the system switches to the normal HARQ operations if no ACK is received even after the maximum autonomous iterations. The efficiency of the proposed autonomous HARQ is validated through extensive simulations. Further, it is concluded that the proposed solution outperforms the conventional HARQ in terms of spectral efficiency with a tolerable trade-off between HARQ gain and the feedback overhead. An integrated autonomous HARQ mechanism for a cooperative system is adopted in [190], where the packets are transmitted several times without waiting for an acknowledgment. The proposed mechanism determines the optimal number of autonomous retransmissions to realize maximum spectral efficiency.

In [191], an autonomous HARQ mechanism is opted for coverage improvement of universal mobile telecommunications system (UMTS) enhanced uplink transmission. With limited power resources, the user equipment (UE) performs several consecutive transmissions without waiting for an acknowledgment from the BS. Likewise, the BS combines the received packets during these autonomous transmissions. After the autonomous transmission phase, the base station sends an ACK/NACK based on the mutual information from the combined decoded packets. The consecutive transmissions of one packet are regarded as a single transmission. The proposed autonomous retransmissions result in reduced BER, coupled with a considerable coverage gain. Further, it is concluded in [192] that blind retransmissions are convincing if the feedback channel does not ensure sufficient reliability, which leads to additional retransmissions because of a drastic increase in the false NACKs.

E. Multi-layer HARQ

The multi-layer HARQ is devised with a motivation of prominent throughput gain and lower latency over conventional single layer HARQ. The multi-layer HARQ is interchangeably termed as a multi-packet or multi-channel HARQ in the literature. Multi-layer HARQ is a non-orthogonal protocol where multiple packets can be transmitted in the same time slot [193]. In a typical multi-layer HARQ system the transmitter is composed of multiple transmission channels, where packets are produced in parallel by multiple channels and are transmitted in parallel [194]. Usually, the same code rate is adopted in all the channels, but in some cases, each channel employs different code rates [195], [196]. These multiple packets are then combined using linear superposition techniques. On the other side, the receivers employ successive interference cancellation (SIC) to extract their information from the superimposed received packet [197]. The packet is extracted from the received signal if it is successfully decoded by a receiver. The same procedure is followed by all the other receivers. The entire decoding procedure is terminated if any of the recipients could not decode its packet successfully. In this case, a NACK is sent back to the transmitter with an index to the last successfully decoded packet. If the transmitter receives an ACK, it transmits new packets in the next time slot. On the contrary, if the transmitter receives a NACK, it transmits new packets, along with the retransmission of old packets [193] or additional parity bits only for the uncoded layers [195], [196].

A modified version of multi-layer HARQ is proposed in [198]–[200] in which two channels are realized for parallel transmission of two packets in a single time slot. The packets are selected based on received ACK/NACK feedback and are (re)transmitted in parallel. The proposed system comprises of two layers, where one layer is employed for transmission of a new packet in each time slot, whereas the second layer transmits an additional redundant packet. The packet in layer two shares the overall allotted energy with the layer one packet in the same time slot. Moreover, in [201], [202] a multi-layer HARQ system is proposed which supports three modes of operations, i.e., conventional HARQ, HARQ with time-sharing, and with superposition coding. The transmitter comprises two layers and the encoder is capable to jointly encode both packets simultaneously. Another HARQ with three parallel layers is proposed in [73], [203] using Polar codes for FEC. The message bits of the previous packets are jointly encoded with the new packet during the retransmission round.

F. Content-Aware HARQ

Some applications are exceptionally resource-demanding and concurrently are bound by strict latency constraints. These applications include, but are not limited to, video streaming and communication. Moreover, it is an absolute demand to realize the best possible QoS for the end-user in terms of video delivery under the limited available resources. As discussed earlier, HARQ grants a reliable transmission mechanism, but the retransmission mechanism curtails the strict throughput requirements of real-time video transmissions. A truncated version of HARQ may be a potential solution to deal with the delays. But truncated retransmission may adversely impact the video quality in terms of achieved peak signal to noise ratio (PSNR). Therefore, a content-aware mechanism is proposed in HARQ assisted video streaming systems, where the video frames are sampled and categorized in distinct ranks, as shown in Fig. 4. Before real transmission of the video, the frames are compressed utilizing the state-of-the-art encoding standards, i.e., H.264, high efficiency video coding (HEVC), and moving picture experts group (MPEG). These encoding standards exploit the spatial and temporal redundancies and correlation among the neighboring video frames. Generally, the video frames are divided into groups, termed as GoPs. The first frame in each GoP is thoroughly encoded, whereas the

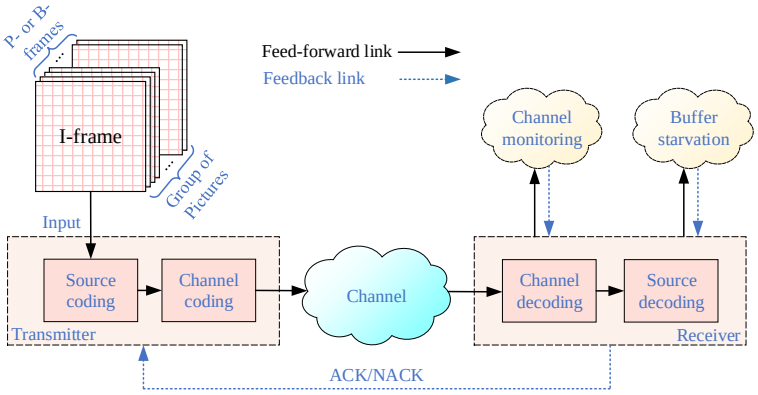

Fig. 4: Content-aware-HARQ system for video streaming.

remaining frames in the same GoP are predicted by exploiting the temporal and spatial redundancies between the consecutive frames in the same GoP. The prediction is based on the difference between the consecutive frames and it determines the motion vectors for each block inside a frame, also termed as motion estimation. The first frame in a GoP is termed as I-frame or independent frame, whereas the other frames are either P- or B-frames, or predicted and bi-predicted frames, respectively. P-frames are encoded through the preceding past frames only, whereas B-frame is a bidirectional frame and is encoded through the past and future frames. The receiver decodes the I-frames independently using the standard image decompression technique, whereas the P- and B-frames are decoded with the help of motion vectors. Hence, the I-frame contains the maximum information and is treated as the most important one because reconstruction of the other frames in current GoP triggers from the I-frame. Considering the I-frame to be the most important one, the HARQ employed video streaming system designates maximum weight to the I-frame. Moreover, the high throughput demands of a video streaming system necessitate the reduced number of retransmissions. Therefore, only the I-frames are preferred for retransmissions and are regarded as high priority frames, which eventually yields to lessen the inherent HARQ delay effects.

In the literature, many solutions are proposed to cope with the implicit delay issues in a HARQ assisted video streaming systems. In [162], [204], an importance-aware HARQ system is proposed where the frames are prioritized within a GoP. The frames are prioritized with a decreasing priority from I-frames to the end of GoP, i.e., the frames close, in order, to the I-frame are given maximum priority and least priority is assigned to the frames closer to the end of GoP. The number of retransmissions is granted according to the priority of frames. Moreover, HARQ mechanism is scheduled according to channel diversity. These diversities are due to the random variations in SNR, distortion, and interference. Therefore, in [183] it is concluded that by selecting a random sub-channel for retransmission, significant throughput enhancement can be realized. Moreover, the frames are categorized into importance levels through time-stamping. In this case, the time-stamp is checked if a NACK is received and the retransmission is aborted if the time-stamp is smaller than the propagation delay. This enhancement shows a notable improvement in throughput. In [181], [205] an unequal error protection mechanism is proposed for video transmission using SR-HARQ. Once again the frames are prioritized and the highest priority is given to I-frames as they contain the maximum information. Retransmissions are proposed for I-frames only, whereas B- and P-frames are discarded if they are not decoded successfully. Moreover, the number of retransmissions is also fixed to avoid extra delays. High throughput is conceived through the proposed enhancements. Usually, satellite communication is highly delay-sensitive with very low path loss. At the same time, the satellites are vulnerable to extreme variable channel conditions. A truncated HARQ-IR is proposed in [206] for LEO satellite transmissions. Rate-less coding is utilized along with truncated HARQ-IR to achieve a fine-grained bit rate and shortened delay.

The real-time video stream imposes stringent latency requirements. Moreover, a HARQ mechanism, employed by a video streaming system, depreciates the throughput in terms of fps due to expected retransmissions. In the real-time video streaming, it is of paramount importance to forecast the playback buffer occupancy at the receiver side. In [160], a link adaptation mechanism is adopted for real-time video streaming. The frames are indexed concerning their importance. I-frames are tagged as the most important frames in a GoP. The HARQ parameters are adapted according to the buffer occupancy and the importance of frames. The delivery budget time for each frame is calculated, where the time varies for different categories of frames. Only important frames are subject to retransmission within their budget time. A concept of false ACK is introduced where the frames are falsely acknowledged if they are well received but after their budget time. In the case of playback buffer starvation, a false ACK is sent back without budget time calculation. A slight degradation in video quality is experienced in terms of PSNR.

Another content-aware HARQ mechanism is proposed in [207], where an unequal protection scheme is proposed for different data partitions. The partitions are encoded through different code rates, according to their importance. Moreover, the retransmissions are controlled through the source coding and channel information. The retransmissions are permitted only for important partitions to realize energy-efficiency and low latency. Furthermore, the receiver sends a NACK only

for the important partitions. A content-aware approach for reliable video transmission is proposed in [182]. Transmission parameters, such as quantization parameter (QP), HARQ, and adaptive coding and modulation (ACM) are adapted at three different layers, i.e., application layer, data link layer, and physical layer, respectively. HARQ mechanism is aware of the frame importance. The joint optimization at three layers improves the quality of the transmitted video. Moreover, to avoid the receiver's buffer starvation each frame is transmitted before the playback deadline.

G. Lessons Learned

As a convincing argument, it is vital to summarize the advance HARQ schemes and to identify the probable pitfalls in them. The general HARQ protocol has demonstrated exceptional success in wireless communications in terms of reliability. Likewise, thorough research is conducted for the advancements in the conventional HARQ to investigate the shortcomings. As discussed in the earlier sections, HARQ scheme offers fruitful results for reliable data transmissions, besides it gives rise to incremental delays, thereby reducing the throughput. In most of the recent wireless applications, throughput and delay are essentially important factors to deal with. For example, the 5G URLLC services force an extremely low latency of 1ms and reliability of around 99.99%. Adaptive HARQ methods are proposed to cope with the shortfalls in the conventional HARQ schemes, exploiting the available resources like power, bandwidth, and the SNR parameters, i.e., the channel conditions, and the combining affects. The code rate, packet length, transmission power, and the number of transmissions are adapted in the succeeding transmissions. These parameters are adapted based on the feedback channel, SNR, AMI, CQI, and the residual rate information. A significant improvement in terms of throughput is realized with the adaptive HARQ methods. Typically, the iterative decoding process induces extra delays in the retransmission process, that increases the complexity of HARQ system. To reduce the computational complexity of the HARQ system different solutions have been proposed. Autonomous HARQ retransmissions are exploited in presence of bad channel conditions, where the retransmissions are highly likely. Content-aware or importance-ware HARQ methods are proposed mainly for video communications. These advancements exploit the importance of each video frame and the retransmissions are carried out only for the frames with comprehensive information, like the I-frames in a GoP. Further advanced methods in the HARQ schemes involve the multi-layer transmissions, where the same data is transmitted in parallel through multiple layers. These multiple transmissions have revealed significant throughput gains. The allocation of available resources dynamically, exploiting the feedback information is also investigated. The dynamic resource allocation realizes an increase in user admission.

So far, advancements in HARQ have projected significant improvement in several aspects. But the global coverage, pretty high throughput and spectral efficiency demands, extreme energy and cost efficiency, ultra-reliability, and security requirements in the future sixth generation (6G) wireless networks seek plenty of analysis.

VI. HARQ in Wireless Standards

A. Wireless Wide Area Networks: UMTS

The UMTS is developed and maintained by the 3rd generation partnership project (3GPP) [208], [209]. It is a third generation (3G) mobile cellular system based on the global system for mobile communications (GSM) standard. UMTS defines three different terrestrial air interfaces based on code division multiple access (CDMA), which are wideband-CDMA (W-CDMA) that uses frequency division duplexing (FDD), time division (TD)-CDMA, and TD-synchronous CDMA (SCDMA) that use time division duplexing (TDD). Since the initial release of UMTS, 3GPP has released updated 3GPP standards that extend and improve the performance of existing 3G mobile telecommunication networks using the W-CDMA protocol. In particular, the high speed downlink packet access (HSDPA) was introduced with 3GPP Release 5, the high speed uplink packet access (HSUPA) was specified and standardized in 3GPP Release 6, and the evolved high speed packet access (HSPA+) was defined in 3GPP Release 7.

UMTS adopts asynchronous-downlink and synchronous-uplink N-process SW-HARQ scheme. UMTS supports both incremental redundancy and Chase combining. Up to a maximum of eight simultaneous processes may exist for a single UE. A packet will be retransmitted if it is not received correctly, signaled with a NACK, or if the ACK/NACK was not received within a specified amount of time in the case of synchronous HARQ. The packet will be dropped if the maximum number of retransmissions, an attribute of the considered HARQ profile, has been exhausted.

B. Wireless Wide Area Networks: CDMA2000

CDMA2000 is a 3G mobile technology family of standards that is the backwards-compatible successor to second-generation CDMAOne [210]. It was the developed by the 3rd Generation Partnership Project 2 (3GPP2), which should not be confused with 3GPP. CDMA2000 was a competing 3G standard to UMTS, and has evolved through a number of evolutionary stages including 1XRTT, 1X Advanced, and EV-DO (EVolution-Data Optimized) with all its revisions.

CDMA2000 uses four-processes HARQ with a maximum of three retransmissions. It also has a synchronous-uplink HARQ and supports both HARQ-IR and HARQ-CC. A HARQ with energy reduction was proposed in Revision D whereby the power associated with the retransmissions of a packet are not equal. The main concept is to give higher power to the early transmission/retransmissions and less power to the later retransmissions. The goal is to achieve higher HARQ gain and less transmission time, however, interference with other users is an issue.

C. Wireless Wide Area Networks: LTE

The air interface of the 3GPP long term evolution (LTE) network is defined by the Evolved UMTS evolved universal

terrestrial radio access (E-UTRA) specifications [7]. Both ARQ and HARQ are used in the Evolved UMTS evolved universal terrestrial radio access network (EUTRAN) protocol stack. These two mechanisms complement each other and their usage is due to the trade-off between fast and reliable feedback. Their combination leads to small round-trip time with a modest feedback overhead. The protocol stack consists of a number of layers and sublayers. The radio link control (RLC) sublayer [211] uses ARQ while HARQ is used by the medium access control (MAC) [212] and Physical [213] layers.

1) RLC layer: ARQ can perform data transfer in one of the following three modes: transparent mode, unacknowledged mode (UM) or acknowledged mode (AM). ARQ is used only in the AM data transfer, and uses a storage window at the sender and receiver. Both windows have the same size, which is defined by the parameter AM_Window_Size. This parameter is set to 512 when a 10 bit sequence number (SN) is used, or 32768 when a 16 bit SN is used. ARQ can send ACK and NACK through the STATUS protocol data unit (PDU) to indicate the reception status of the transferred PDUs. A NACK has a sequence number field that indicates the SN of the acknowledged mode data (AMD) PDU that has been detected as lost at the receiving side. The ACK message also includes a sequence number that indicates all AMD PDUs up to, but not including, the AMD PDU where SN = ACK_SN have been received correctly, excluding those indicated by the NACK sequence number.

An AMD PDU can be retransmitted up to a maximum number indicated by the maxRetxThreshold parameter. This parameter is set by the radio resources control (RRC) sublayer [214] and can get a value in the set $\{1, 2, 3, 4, 6, 8, 16, 32\}$ with a default value of 4.

2) MAC and Physical layers: The MAC and Physical layers collaborate to implement HARQ. While the MAC layer is responsible of the HARQ protocol management and signaling, the physical layer is responsible of the generation of different redundancy versions at the transmitter as well as the soft combining at the receiver. The LTE-HARQ mechanism, implemented by a HARQ entity, is structured as a set of multiple SW processes operating in parallel. SW is adopted for its simplicity while using multiple processes in parallel allows for increased throughput through continuous transmission without wait stages. Typically, there is one HARQ entity per device. The number of HARQ processes depends on whether FDD or TDD is used. In FDD, eight HARQ processes are used, while in TDD the number of HARQ processes depends on several configuration parameters and can range from 1 to 15 processes. The maximum number of HARQ retransmissions can be configured based on the UE capabilities signaled by the RRC sublayer.

LTE-HARQ uses soft combining, where the receiver combines the received signals from multiple transmission attempts. LTE uses IR, a type of soft combining where the retransmissions do not have to be identical. Instead, for each retransmission, multiple sets of coded bits are generated, each representing the same set of information bits. IR is used as it results in coding gain for each retransmission in addition to a gain in accumulated received E_b/N_0.

HARQ protocols can be identified as synchronous versus asynchronous based on their flexibility in the time domain, as well as adaptive versus non-adaptive based on their flexibility in the frequency domain. LTE uses an asynchronous adaptive HARQ for the downlink, and a synchronous HARQ for the uplink. Uplink HARQ retransmissions are typically non-adaptive, but they can be made adaptive if needed. In the adaptive HARQ, the coding and modulation schemes, the redundancy version, and the subcarrier can change for each retransmission. Therefore, any change in the values of these parameters have to be indicated to the receiver by the protocol.

Soft combining is supported through transmission of an explicit *new-data indicator* flag, toggled for each new transport block. This bit allows the receiver to know when to perform soft combining prior to decoding, in the case of a retransmission, and when to clear the soft buffer, which is for the case of an initial transmission. It also allows the transmitter to know whether to retransmit erroneously received data or to transmit new data.

A notable behavior of the uplink HARQ is the usage of the new-data indicator to alter the traditional semantic of ACK/NACK. To support both adaptive and non-adaptive HARQ, the transmission buffer is not flushed when receiving a positive acknowledgment. Instead, the actual control of whether data should be retransmitted or not is performed by the *new-data indicator*. If this indicator is toggled, the device flushes the transmission buffer and transmits a new data packet. However, if this indicator is not toggled, the previous transport block is retransmitted. In this case, the negative HARQ acknowledgment plays the role of a scheduling grant for retransmissions.

Finally, through TTI bundling, LTE allows a UE to transmit the same data multiple times in a row, TTI bundle, to increase the possibility of data reception and decoding. The number of TTIs of a TTI bundle is four. In this case, HARQ works at the level of the bundle and a retransmission of a TTI bundle is also a TTI bundle.

D. Wireless Wide Area Networks: WiMAX

worldwide interoperability for microwave access (WiMAX) is the IEEE 802.16 set of standards that was proposed as a candidate for the 4G wireless networks and a competitor of LTE Advanced standard. The physical layer of the WiMAX standard supports two modes of operation for the HARQ mechanism, namely CC and IR [8]. The standard uses multiple HARQ channels each of which uses SW protocol. This allows one channel to transmit data while others are waiting for acknowledgments, which compensates for the propagation delay of the SW. The maximum number of HARQ channels per mobile station is 16 in the uplink and downlink. The use of the SW protocol allows also for low memory requirements. Each HARQ channel has a HARQ channel identifier. To specify the start of a new transmission on each HARQ channel, a one-bit HARQ Identifier Sequence Number is toggled on each successful transmission.

In the standard, the maximum number of retransmissions of the same data packet is 4. For each HARQ physical data

service unit (PSDU) 4 subpackets are generated. When the HARQ is operating in IR mode, each subpacket contains different parity information generated by different FEC puncturing pattern. In this case, a subpacket identifier (SPID) is used to uniquely identify each subpacket. In CC mode, the SPID is not used since the content of each subpacket associated with a HARQ PSDU is the same. If a subpacket is not received successfully a retransmission is requested and the next subpacket for the same HARQ PSDU is transmitted. The HARQ packet is discarded if the maximum number of retransmissions is reached.

WiMAX uses adaptive asynchronous HARQ in the downlink. Hence, in the case of a retransmission, appropriate signaling is required to indicate the resource allocation and transmission format, along with other HARQ parameters. An asynchronous HARQ scheme is used to provide more flexibility for the downlink scheduler. A failed HARQ burst should be retransmitted within the maximum retransmission delay bound. The delay between two consecutive HARQ transmissions of the same data burst should not exceed the maximum value chosen in the set $\{1, 2, \ldots, 8\}$. Synchronous HARQ is chosen for the uplink transmission. That is the user equipment sends HARQ feedback after a fixed delay. A dedicated ACK channel is provided for HARQ ACK/NACK signaling. In synchronous HARQ, resource allocation for the retransmissions in the uplink can be fixed or adaptive based on the control signaling. The default operation mode of HARQ in the uplink is non-adaptive, however the BS can signal an adaptive uplink HARQ mode if needed.

E. Wireless Wide Area Networks: 5G

The 3GPP specification 38 series [9] provides the technical details of 5G new radio (NR), the radio access technology beyond LTE. The HARQ process in 5G NR is an enhanced and more flexible version of HARQ in LTE. More specifically, the HARQ in 5G is adaptive and asynchronous for both the downlink and uplink. A UE can support up to 16 HARQ processes if needed. In contrary to LTE where the time gap between the reception of data and the reporting of HARQ-ACK feedback is fixed to 4 ms, 5G NR allows for more flexibility in reporting HARQ feedback. For example, for low latency services, the network is able to request very fast HARQ feedback. In other circumstances, the network may delay the HARQ feedback if needed.

In addition, to save downlink transmission resources, the UE can use a codeblock-group (CBG) based HARQ feedback instead of a transport block (TB) based HARQ feedback. In a CBG-HARQ, each TB is divided into several CBGs each of which is transmitted with its own CRC bits. This allows the UE to report on each CBG individually, and allows the network to retransmit only the CBGs that the UE failed to receive. The uplink in 5G NR also supports TB-based and CBG-based HARQ.

F. Wireless Local Area Networks (WLANs): IEEE 802.11 Standards

The IEEE 802.11 is a standard that provides the basis for wireless communication in wireless local area network (WLAN)s by specifying a set of MAC and physical layer (PHY) protocols. Since its first release in 1997, it has gone through many improvements over the years with the goal of improving the throughput, supporting QoS, and allowing more devices in the network. The latest version up to date is [10]. In addition to the main standard defined in [10], the IEEE 802.11 working group also publishes amendments that detail specific protocols using different frequency bands, modulation schemes, and capabilities. Over time, some of these amendments get incorporated into the main standard while new amendments get published.

For frames that need to be acknowledged, 802.11 supports two styles of ARQ: SW and SR.

- SW-ARQ: In this mode, every frame has to be acknowledged before the following frame can be transmitted. An ACK timeout is used to detect lost frames. The frames have sequence numbers so that duplicate retransmissions can be detected and dropped.
- SR-ARQ: In 802.11 protocols that supports QoS and high throughput, such as 802.11ac and 802.11ax, an SR-ARQ is used. In this scheme, the transmitting station maintains a transmission window of frames that have been sent and for which an acknowledgment is not yet received. The frames in this window can be individually acknowledged using the legacy ACK frame or in groups using the block ACK (BAck) frame according to the type of frames. The BAck frame allows informing the transmitting station about which frames did arrived successfully and which did not, using their sequence numbers. Only the frames that did not arrive successfully, will be retransmitted. When the first frame in the transmission window is acknowledged, the window is advanced by one place to include a new frame at the end of the transmission window. To reduce the size of the BAck frame, a compressed version of the BAck is defined where it contains a starting sequence number SSN and then a Bit-map. Every bit in the Bit-map corresponds to a sequence number, starting from SSN and progressing in order. A bit position n of the bitmap is set to 1 if the frame with the sequence number $SSN + n$ is received successfully, and it is set to 0 otherwise. The maximum number of frames in the transmission window is set to 64.

IEEE 802.11 has several parameters that can affect the number of transmissions per frame. For frames that require acknowledgment, 802.11 defines two different retry limits for short and long frames. The limits specify how many times a frame should be retransmitted before it is dropped and the higher layers are informed. The limits are defined by two parameters *dot11ShortRetryLimit* and *dot11LongRetryLimit* for the short and long retry limits, respectively. The parameters are stored in the management information base (MIB) of the device and can be accessed using the simple network management protocol (SNMP). Hence, every wireless station (STA) shall maintain a station short retry count (SSRC) as well as a station long retry count (SLRC), both of which shall take an initial value of 0 at the beginning of the initial transmission and incremented as appropriate, short or long, after

every retransmission. When a counter reaches the associated limit, the frame is dropped and the counter is reset. What constitutes a short or long frame is controlled by the parameter *dot11RTSThreshold* which is also in the MIB. A frame can be also discarded if its lifetime has been reached regardless of counter value. The lifetime of a frame is controlled by a number of parameters that can be set based on the frame type, access category (AC), or traffic category (TC) that the frame belongs to.

G. Wireless Personal Area Networks (WPANs): Bluetooth

Bluetooth is a wireless technology standard that is used for building WPANs [215]. Bluetooth is managed by the Bluetooth Special Interest Group [216]. Bluetooth defines several types of packets, which may or may not require acknowledgment, depending on the application specifications. For the packets that require acknowledgment, Bluetooth uses FEC and ARQ. Typically a packet that is not received correctly will be retransmitted until successfully received or a timeout expired. Bluetooth uses a HARQ-I, and also uses a fast, unnumbered acknowledgment scheme, that is a kind of synchronous HARQ. Therefore, an acknowledgment has to follow immediately in the slot following the reception of a packet.

H. WPANs: Low-Rate (LR) WPANs

The low-rate (LR)-WPANs operations are defined in the IEEE $802.15.4$ standard [217] that is maintained by IEEE 802.15 working group. The technical standard specifies the physical and MAC layers for LR-WPANs, and is the basis for several WPAN standards such as Zigbee and 6LoWPAN that specify the upper layers of the protocol. The IEEE $802.15.4$ standard supports both acknowledged and non-acknowledged frames. An acknowledgment is sent when a frame is received correctly, and a negative acknowledgment is sent when the frame is received incorrectly. A retransmission is initiated only if the transmission was direct. If the transmission was indirect, the frame shall remain in the queue of the coordinator and can only be extracted following the reception of a new Data Request command. An indirect transmission is a type of transfer where the device instigates the transfer of data rather than the coordinator. In this case, either the coordinator needs to indicate in its beacon when messages are pending for a device, or the device itself should poll the coordinator to determine whether it has any pending messages.

In the case of direct transmission, the device retransmits the frame and waits for an acknowledgment, up to a maximum of *macMaxFrameRetries* times. The retransmission shall be attempted only if it can be completed within the same portion of the superframe in which the original transmission was attempted, otherwise it is deferred until the next superframe. If *macMaxFrameRetries* retransmissions have been reached without success, the MAC sublayer assumes that the transmission has failed and notifies the next higher layer of the failure.

The retransmissions are exact copies of the original transmission, hence, the IEEE $802.15.4$ standard uses HARQ-I. The maximum number of retransmissions allowed after a transmission failure, *macMaxFrameRetries*, can range from 0 to 7 with a default value of 3. The retransmissions are also controlled by a number of timing constraints and a non-successfully delivered frame will be dropped if *macTransactionPersistenceTime* is reached. This maximum time, in unit periods, is chosen between 0 and 65535 with a default value of 500.

I. WPANs: High-rate (HR) WPANs

The IEEE Standard for High Data Rate Wireless Multi-Media Networks, IEEE Std $802.15.3$-2016 [218], was originally developed to provide superior QoS over relatively medium range wireless links. The standard defines PHY and MAC specifications for high data rate wireless connectivity, typically over 200 Mbps, with fixed, portable, and moving devices. The standard then focused on the concept of piconet, which is defined as a wireless ad hoc network that allows a number of independent devices to communicate with each other. A piconet is distinguished from other types of data networks in that communications are normally confined to a small area around a person or object that typically covers 10 m in all directions.

The standard defines five acknowledgment types that can be used between communicating devices as needed by the application:

1) No acknowledgment (no-ACK): Used when the receiver does not need to acknowledge the received frame and the sender assumes a successful reception. This ACK policy is used for broadcast and multicast transmissions.
2) Immediate acknowledgment (Imm-ACK): Used when the receiver has to immediately send an ACK after correctly receiving a frame.
3) Delayed acknowledgment (Dly-ACK): Can be used after negotiation between the source and the destination. Two parameters are negotiated, the Max Burst field and the Max Frames field, that control the amount of data the source can send in one burst. The Max Burst field is a value representing the maximum number of *pMaxFrameBodySize* MAC protocol data units (MPDUs) the source may send in one burst. If the frames are smaller than *pMaxFrameBodySize*, then the number of frames should not exceed the Max Frames field. The maximum frame length allowed, *pMaxFrameBodySize*, is 2048 octets.
4) Implied acknowledgment (Imp-ACK): Is a technique that allows a more efficient use of channel time allocation by allowing bidirectional data transfer. With Imp-ACK, the ACK is implied when the target device sends any frame in response to a frame that has an ACK policy of Imp-ACK.
5) BAck: Used with an aggregated frame, which is a frame that is composed of number of subframes. The BAck has a Bitmap field in its MAC subheader to be used by the destination. The destination, upon receiving an aggregated frame, checks each subframe. Based on the status of the subframe, either correctly or incorrectly received, the corresponding ACK bit of the BAck Bitmap field will

be set. The source, after reading the BAck Bitmap field in MAC subheader, will handle the incorrectly received subframes retransmissions.

In all cases, an incorrectly received frame will be retransmitted until a retransmission limit is reached, and in this case the frame will be discarded. The retransmissions are the exact copy of the original transmission, hence, the IEEE $802.15.3$ standard uses HARQ-I.

VII. HARQ in Emerging Wireless Technologies

This section considers the integration of ARQ/HARQ in various emerging wireless technologies to provide an overview about key issues such as reliability, latency and throughput.

A. URLLC

URLLC is one of the latest operating modes in 5G. The URLLC mode requires almost 100% reliable communication with a round-trip latency of around 1 ms, which are significantly challenging requirements to satisfy. In the literature, several HARQ techniques were proposed to satisfy these requirements. The retransmission process associated with HARQ protocols is proven to be efficient for improving the data reliability. Nevertheless, the retransmission process also increases the latency. In addition to reliability and latency issues, URLLC applications also require a high throughput as 5G is expected to provide data rates more than 1 Gbps. Therefore, providing highly reliable communications with low latency has attracted extensive research efforts. Some of the key works that considered HARQ for URLLC are discussed below.

In [127], the authors propose a power allocation algorithm to specify the outage probability for repetition and parallel HARQ retransmissions. The authors implement the algorithms for various HARQ protocols, including HARQ-CC, HARQ-IR and the ARQ protocols. In [219], a NOMA based retransmission is proposed for uplink in URLLC systems. In the proposed NOMA-HARQ, the new and retransmitted information packets can share the transmission resource to reduce the delay and maximize the throughput. The authors in [220] consider throughput maximization by optimizing the number of retransmissions. Moreover, the optimal modulation and coding scheme (MCS) is determined in order to obtain the maximum achievable throughput. The concept of partial retransmission is considered in [221], which provides an opportunity to retransmit the damaged or corrupted part of data due to puncturing. Consequently, high throughput gains can be achieved.

Generally speaking, the decoding process at the receiver consists of several processes such as computing the LLRs, deinterleaving, generating the soft data, searching for the nearest codeword, etc. [19]–[21]. Therefore, the decoding process is time consuming and may induce significant time delay, which contradicts the critical timing requirements of URLLC. Therefore, an early retransmission HARQ technique is proposed in [222] where the retransmission decision is made based on the CSI. The early retransmission before channel decoding eventually reduces the time delay. Nevertheless, relying on the CSI might result in unnecessary early retransmission, and hence, throughput loss. To overcome this limitation, the authors propose a multistage decision process where the early retransmitted bits start from a small value, and increase gradually based on the CSI [223]. The issue of throughput loss with early retransmission is also discussed in [224], where the authors propose a method to multiplex early retransmission data with the initial data, within the same channel, for next transmission using superposition coding. In [120], the HARQ-IR and HARQ-CC are investigated with two different cases, i.e., with BLIND retransmission and NACK-based retransmissions. In BLIND retransmissions, the transmitter resends the packets aggressively until it receives an ACK, whereas in NACK-based schemes, the retransmission takes place only when the NACK is received.

Energy efficiency is another important aspect to be considered for URLLC. For example, some real-time applications implemented using IoT require energy-efficient communications. Also, it is observed that energy and latency performances are conflicting in nature. In [225], a trade-off between energy and latency is studied, where HARQ-IR is employed as a retransmission scheme. It is shown that 25% energy is saved with HARQ-IR retransmission when compared with no-HARQ transmission. A blind retransmission on the shared radio resources along with SIC is considered in [226] for reduced latency. Another novel concept for short packet communications in URLLC is discussed in [173], where short packets are obtained by converting the original long packet into a number of sub-words, and the HARQ-IR is investigated for short packet communications. The short packet communication is compared with one-shot communication to evaluate the trade-off between energy and latency. In wireless communications systems, the resource allocation process also contributes towards the delay. Therefore, an efficient solution is proposed in [227] using the concept of pre-scheduled resource sharing. In the proposed solution, all users are supposed to retransmit at fixed time intervals. The retransmission is named as synchronous HARQ. The proposed scheme shows an imminent improvement in terms of resource efficiency.

The authors in [147] propose an HARQ optimization algorithm to maximize URLLC capacity, or minimize bandwidth utilization subject to packet delivery deadline, available bandwidth, and a decoding failure probability. The optimization parameters are the maximum number of retransmissions and number of allocated resource blocks. It is assumed that simultaneous transmissions of multiclass packets each of which is a SW-ARQ with CC. The queuing model assumed is M/GI/∞ with no queuing delay. In [228], a similar optimization model is considered, however, an M/G/1 queuing model with queuing delay is assumed. A truncated HARQ-CC is also proposed in [229] for URLLC where a new approach is proposed to tune the maximum number of retransmissions to minimize energy consumption in HARQ-CC under latency, reliability, and transmit power constraints.

In [145], a discrete-time finite state Markov model is used to theoretically characterize the throughput performance of HARQ-IR in URLLC systems as a function of transport block

size, modulation scheme and maximum number of transmission rounds. The throughput performance of HARQ-IR with Luby transform codes is also evaluated through computer simulation to validate the proposed theoretical model. The performance of HARQ-IR is also theoretically analyzed in [230]. The theoretical model is then used to derive delay/error-rate bounded QoS metrics to optimize millimeter wave (mmWave) cell-free massive MIMO (mMIMO) for URLLC.

In [231], the authors point out that HARQ with ACK/NACK feedback may not be suitable for many URLLC use cases with strict latency constraints. Instead, they proposed allocation of transmission resource slots to transmission replicas to achieve high reliability through diversity rather than HARQ retransmissions. They considered and compared different multi-user resource allocation approaches, namely, fully dedicated slots, fully shared slots, and hybrid allocation of dedicated and shared slots. The authors in [232] also propose repeated packet transmissions within a transmission instance. However, unlike [231], they allow retransmissions with individual feedback for each repeated packet.

B. Massive machine type communication (mMTC)

With the emerging machine type communication (MTC) protocols, a large number of geographically distributed machines/devices are supposed to communicate with minimal or no human intervention. In 5G, MTC/mMTC is one of the main topics to address the expected inclusion of millions of IoT devices in the network. mMTC involves a handshake process between the UE and evolved node B (eNB). The UE establishes a connection with the eNB and then requests for uplink resources. During the handshaking process, if two or more UEs select the same preamble sequence number, a collision occurs. This collision is referred to as MSG3 collision. Therefore, a probabilistic HARQ is proposed in [233] to tackle the MSG3 collision problem in cellular IoT networks. In IoT networks, secrecy could be an issue as the MTC devices work in an autonomous fashion. Therefore, to ensure the secrecy of IoT devices, a secure HARQ is proposed in [234] for NOMA-based networks. Secure HARQ is employed at the user which demands security to ensure that the SIC is performed successfully. Another challenge for MTC is the frequent interaction of machines with each other, which eventually introduces contentions, and thus, latency issues. To address the latency due to this signaling overhead among the machines, a detection-based ARQ protocol is proposed in [235] where each machine monitors the resource blocks during the transmission, and a retransmission is performed soon after the detection of a collision, without waiting for any kind of acknowledgment. The performance of HARQ in mMTC applications is mathematically modeled in [236], [237]. Moreover, in [238] the performance of HARQ in mMTC is numerically evaluated when joint network and fountain coding is used.

C. MIMO and mMIMO

MIMO is widely considered in wireless communication to mitigate the adversarial effects of multipath propagation. Multipath propagation is handled using multiple transmitting and receiving antennas, which eventually improves the SNR. Alternatively, MIMO can be used for data multiplexing, where the transmitter sends data in multiple parallel streams to increase the data rate. However, these transmissions are needed to be reliable. For this purpose, [239]–[242] investigate the performance of MIMO with different types of HARQ.

Adaptive power allocation is combined with ARQ in [243] and with HARQ-CC and HARQ-IR in [244] to minimize error probabilities in MIMO systems. A hybrid scheme of bit level combining (BLC) and symbol level combining (SLC) is proposed in [245] for MIMO systems with HARQ-CC. The obtained results show that the hybrid BLC-SLC outperforms BLC and provides similar performance to SLC with simplified reception procedures.

Point-to-point mMIMO with HARQ-CC is considered in [246] where precoding and combining are progressively optimized during retransmission to maximize spectral efficiency. The number of data streams is assumed to be less than the number of transmitter and receiver antennas. The authors in [247] consider an uplink scenario with one transmitting antenna and multiple receiving antennas. It is shown that the age of information (AoI) and energy consumption of HARQ are improved by jointly exploiting time diversity and space receiver diversity. Here AoI is defined as, *the time elapsed since the generation of the last successfully received message containing update information about its source system* [248]. AoI is an important metric in real-time applications, where the access point (AP) is always looking for the updated information from the end devices. On the other hand, a multi-user overloaded MIMO system with HARQ-IR is considered in [249] where the number of transmitting antennas is larger than the number of receiving antennas. In the proposed model, retransmissions and new transmissions share the same MIMO channel. The authors show that the proposed shared HARQ outperforms conventional HARQ in terms of BER and throughput in overloaded MIMO systems.

D. Edge Caching and HARQ

Caching is a promising solution to reduce the network load by storing popular content in local caches of network terminals such as small base stations (SBSs), wireless access points (WAPs), or end-user devices. Therefore, various caching aspects have been considered in the literature. For example, the authors in [250] considered a two-tier cooperative cellular network with macro base stations (MBSs) and SBSs. The source content is stored in the MBSs whereas subsets of the content that are most popular are cached in the SBSs. Users request data from the nearest SBS, which transmits the requested data to the user when the data is available. Otherwise, the SBS requests the data from the nearest MBS. Once the requested data is successfully received at the SBS, it will forward the data to the user. The authors assume an ARQ-based transmission from the MBS to the SBS, and from SBS to the user. It is shown that the content caching based on popularity improves the average transmission delay and average outage probability.

A cache-aided HARQ with soft information combining is proposed in [251] for cooperative device-to-device (D2D)

communications where each device performs probabilistic caching of popular content up to its caching capacity. When a user fetches a content, the user's device first checks self-cached content, and if required content is not available locally, it is requested from nearby peers. The authors evaluated the impact of the HARQ combining strategy on the obtained performance gain of caching. In particular, mutual information accumulation soft combining provided better cache-aided successful transmission probability when compared to energy accumulation or no combining schemes. However, the performance gain vanishes under severe channel conditions, or when the users' requests are concentrated on popular data.

The authors of [252] proposed caching the packets that are to be dropped at intermediate routers to improve the performance of ARQ schemes in transmission control protocol (TCP). When a data packet is to be dropped due to overflow of the main buffer of a router, the packet is cached on an auxiliary buffer at the router. When ACK timeout occurs, the source sends a sniffer packet rather than retransmitting the original data packet. The sniffer packet contains the identity of the timed-out packet. It is assumed that the sniffer packet is likely to reach the router at which the packet is cached and subsequently triggers the packet retransmission from the router to the intended destination. The obtained evaluation results show that the average end-to-end transmission delay is reduced when the proposed caching is implemented.

E. Edge Computing and cloud-radio access networks (CRANs)

Edge computing has emerged as an innovative technique to exploit the processing power of computing devices on the network edges. In edge computing, the devices, especially the low power devices, can offload their data to some nearby edge device for computation. The tasks are either partially or fully offloaded to the edge device. The decision depends on the local or offloaded computation time and resources. The interaction between edge computing and ARQ can be used to improve the system efficiency. For example, the authors in [253], [254] proposed a novel importance-aware ARQ protocol for data acquisition in edge learning systems. This protocol efficiently adapts the retransmission considering the data importance. Consequently, the limited transmission budget is efficiently utilized.

CRAN is a candidate 5G cellular technology that is based on decoupling the baseband processing unit (BPU) and radio access unit (RAU), also know as remote radio head (RRH), in every eNB [255]. The RRH performs radio functionalities whereas the BPU performs baseband processing and the rest of the protocol stack. This architecture is proposed to reduce deployment cost where multiple RRHs can share a single BPU for joint baseband processing. However, a main drawback of CRAN is the increased latency due to the added fronthaul network between the RRHs and BPUs. Moreover, the additional data transfer between the RRHs and BPU may introduce additional errors. Therefore, the authors in [256] proposed an uplink HARQ protocol with separation of control and data planes to mitigate the latency drawback. The retransmission control decisions are handled at the edge of the network, namely, the RRH, while data decoding is centrally performed at the BPU.

F. Machine Learning for HARQ

A turbo-coded HARQ scheme is proposed in [85] where artificial neural networks are employed at the receiver to predict decoding errors. It is shown that the proposed error prediction allows for efficient retransmission process with reduced overhead and decoding complexity, while providing similar reliability when compared to CRC-based error detection. A reinforcement learning (RL) algorithm is proposed in [257] to learn, in real-time, the optimal transmission policy which minimizes the average AoI. At the beginning of each transmission slot, the sender decides based on the RL algorithm whether to stay idle, transmit a new packet, or retransmit a previously failed packet.

In [258], one-shot transmission is adopted to minimize the delay in underwater WSNs by dropping the ARQ process. Instead, the temporal dependencies between transmitted sensors' data are exploited to estimate the missing data using a recurrent neural network prediction model at the data collection center. A machine learning approach is proposed in [259] to predict decoding results in early HARQ (E-HARQ) employing LDPC codes. The authors argue that their proposed enhanced E-HARQ provides a potential solution for URLLC since it can achieve lower latency and/or higher reliability by allowing more retransmission rounds when compared to regular HARQ.

G. NOMA and HARQ

NOMA is a major breakthrough in wireless communication that can address the demands of concurrent connections and resource sharing. In power-domain NOMA, multiple users are multiplexed on the same channel using different power levels. The UEs apply SIC to decode their intended signals. With the emergence of NOMA, the spectrum scarcity constraint is noticeably relaxed. However, NOMA suffers from the error propagation problem due to its sequential decoding process. Therefore, the performance of NOMA with different types of HARQ has been investigated [125], [260]–[262].

Adaptive power allocation for HARQ-CC and HARQ-IR are studied in [123], [263], [264] to ensure low errors or delays in NOMA systems. Moreover, dynamic UE pairing in NOMA based systems is proposed in [265] where different UE pairs may be considered in each retransmission with the objective to improve the transmission diversity, delay, and fairness.

H. Cognitive Radio and HARQ

The preoccupied available spectrum has limited the admission of more users in wireless networks. Massive connections and devices need to be incorporated in 5G, but the mostly saturated spectrum is unable to serve such devices. Therefore, researchers have focused on various aspects to find solutions to effectively utilize the spectrum. CR has become a topic of interest for the wireless communication society in recent years to address spectrum scarcity. In CR, the unlicensed/secondary

devices can detect and opportunistically use the spectrum allocated for the licensed/primary devices without degrading the performance of licensed devices. Many researchers have studied HARQ for reliable transmission in CR systems. Some considered HARQ for secondary users [108], [109], [111], [142] and others considered HARQ for primary users [266]–[269]. When HARQ is employed for primary users, secondary users cooperate by relaying primary users' data and get spectrum access as a reward for their cooperation. Alternatively, secondary users jointly transmit their data and primary failed data using adaptive power allocation.

I. Cooperative Communications and HARQ

Cooperative communications has become an integral part of 5G networks. The diversity provided by cooperative communications plays a vital role in terms of spectral efficiency and reliability. Many researchers have proposed some novel techniques to ensure reliability and throughput maximization in cooperative communications using retransmission protocols.

An energy-efficient relay selection for cooperative HARQ is proposed in [138]. HARQ-IR is considered with decode-and-forward (DCF) relaying over Rayleigh fading channels. The performance of cooperative HARQ-IR with DCF relaying over a general time-correlated Nakagami-m fading channel is analyzed in [270]. Other examples of cooperative HARQ in DCF relay networks are considered in [271], [272]. On the other hand, the authors in [273] analyze the performance of cooperative ARQ in amplify-and-forward (AF) relay networks to demonstrate the achievable throughput enhancement when cooperative ARQ is used. A probabilistic HARQ for demodulated-and-forward (DF) relay networks is considered in [113]. The probabilistic HARQ is shown to offer significant improvements in terms of PER, when compared to deterministic HARQ.

In [274], cooperative HARQ-IR is employed with multi-sources and multi-relay scenario. It is shown that by deploying cooperative HARQ-IR, a better throughput along with lower outage probability is achieved. Another multi-source and multi-relay scenario is discussed in [112] where a centralized scheduling strategy is proposed with HARQ-IR. The communication is divided into two phases, where the sources or relays transmit during the first phase using time multiplexing. In the second phase, the destination schedules the relays or sources for retransmission. The retransmissions are scheduled in limited time slots. The relays or sources are selected by the destination with the objective to maximize the throughput.

J. WSNs and HARQ

Retransmission using HARQ protocols is employed in WSNs to improve the reliability of the transmitted data. Sensor nodes are usually battery operated and therefore have limited power. Hence, HARQ schemes for WSNs need to be designed under power constraints. For example, a HARQ scheme is proposed in [275] where only the least reliable bits are retransmitted. The reliability of the sensor network is further evaluated in terms of BER through simulation results. A similar concept is proposed in [276] where partial retransmission is used to achieve reliability with enhanced throughput. In [116], HARQ with autonomous retransmission is proposed for multicast WSNs. The autonomous retransmission transmits the data regardless of decoding results at the receiver to achieve high spectral efficiency.

K. SWIPT and HARQ

Power constrained devices, such as IoT devices and sensor nodes are equipped with small on-board power sources. Such devices are usually put in sleep mode when they have no information to transmit or when their data is not needed, and radio frequency (RF) signals are used to wake them up. These devices can make use of SWIPT where a device may harvest energy from radio signals to charge its battery. In [277] the ratio between accumulated information and harvested RF energy is optimized for HARQ-IR to minimize the number of retransmissions.

L. Multimedia Streaming and Transmission

With the advancement in multimedia-enabled devices, the demand of high data rate with reliable transmission is continuously increasing. In [182], a cross-layer adaptive scheme is proposed for H.264 video transmission. A media-aware HARQ protocol is presented by prioritizing the importance of frames, like I-frame, B-frame and P-frame. Similarly, the GoP and quantization parameters are varied to achieve Rate Distortion Optimization. Current cellular networks have multimedia broadcast systems, like Multimedia Broadcast and Multicast System (MBMS) and enhanced Multimedia Broadcast and Multicast System (eMBMS). These systems are designed to efficiently broadcast multimedia messages to a group of devices in a cell. Multimedia broadcasting with HARQ, is discussed in [278].

M. UAV-assisted Communications

Existing wireless networks are unable to accommodate the exponential surge in new devices. The use of UAVs has become one of the potential solutions to rapidly deploy wireless networks based on demand. However, there are certain limitations and constraints with UAVs as well, particularly the service (flight) time of UAVs is limited.

A UAV assisted data ferrying network is analyzed in terms of energy efficiency, reliability and outage probability in [139]. In this data ferrying network, the UAV carries the data from a base station and moves towards multiple destinations. The reliability of data communication is evaluated using standard ARQ retransmissions and without retransmissions.

N. Limitation of Existing Work

Most existing work of HARQ in emerging wireless technologies mainly focus on time diversity. Proposing solutions with combined time, frequency and space diversity may reduce the reliance on time diversity and can enhance the latency performance of these systems while also improving their

reliability due to the added frequency and space diversity. Another limitation of some of the existing work is assuming perfect knowledge of channel status when designing HARQ optimization algorithms. In practical communication systems, channel status may be inaccurately estimated. Hence, additional studies of the performance of HARQ emerging systems in the presence of channel estimation errors are needed. Binary ACK/NACK feedback is usually assumed in HARQ systems. Additional work on non-binary granular feedback and its impact on HARQ performance would be beneficial. Most work in the literature consider transmission delay in terms of number of retransmissions when analyzing HARQ latency. A more comprehensive analysis considering other latency components such as processing, queuing and routing delays will provide a more accurate evaluation of HARQ performance in emerging systems. Moreover, some of the existing work consider limited enumerations of block sizes, code rates, modulation orders and coding schemes. Conducting exhaustive simulations and real experiments of a wide range of these parameters will be of great value for academia and industry. Studying the implementation complexity of the proposed solutions will also be useful. In addition, real deployment of some the proposed solutions for specific applications in an industrial setting will provide strong validation of these proposed systems.

VIII. Future Directions

As can be noted from the aforementioned discussions in this work, using ARQ or HARQ is essential to provide guaranteed QoS in terms of PER. However, the cost is a reduced throughput, increased delay and higher energy consumption. Therefore, mitigating the adversary effects of the retransmission process has attracted tremendous efforts. However, the need for more efficient retransmission process remains an open problem as the demand for higher throughput, less delay and more energy efficient communications is still increasing. In the following, we discuss some potential solutions that may form the basis for more efficient retransmission process.

A. Proactive ARQ

ARQ is generally reactive by nature. The retransmission process is typically triggered based on the receiver feedback. As such, the process is not truly automatic as the name implies. The interaction between the transmitter and receiver causes large delay and reduces the throughput as the feedback messages to the transmitter consume some bandwidth, although they are generally short packets. Consequently, significant benefits can be gained if the transmitter behaves in a proactive manner instead of being reactive. Proactiveness can be realized if the transmitter can predict the channel conditions, and thus, the packet failure probability can be computed in advance. Therefore, the transmitter may decide to transmit the same packet for a certain number of times without waiting for the receiver feedback, which may reduce the delay and feedback overhead significantly. The receiver may use CC to combine the repeated packets before being sent for subsequent processing. It is worth noting that this approach is already used in some standards such as LTE, which is known as TTI bundling. Nevertheless, the current use of TTI bundling depends on the statistical channel information, which limits its application to users who suffer large scale fading, such as cell-edge users. Moreover, the number of repeated packets is fixed to 4. To fully utilize the potential of this approach, it should be applied proactively to small scale fading, and with variable number of transmissions. Therefore, it is pivotal to use efficient channel prediction schemes that are capable of providing future channel information prediction for a few packets duration. There are some promising approaches in the literature that are based on artificial intelligence (AI) [279] and ray tracing [280], however, the computation time and prediction duration should be enhanced.

B. Intelligent ARQ

Retransmission protocols usually specify the interaction between the transmitter and receiver based on the outcome of the decoding process at the receiver side. A successful decoding should be followed and an ACK, otherwise a NACK will be sent. Therefore, the protocols do not generally consider the content or context where the protocol is applied. The factors that can be taken into consideration while designing an ARQ protocol may include the battery level, power source, buffer occupancy, packet type, user vigilance, and user safety or comfort level. Some examples for exploiting such information are given below.

Battery level and power source: Given that the transmitter and receiver exchange their battery energy levels or power source, then the ARQ process can be designed to prolong the connection level. For example, if the receiver is powered by a battery with low energy level, then the protocol can reduce the power consumption at receiver by asking the transmitter to increase the power for all retransmitted packets, which saves the energy required to send frequent feedback to the transmitter. Moreover, since most receivers employ iterative detection and decoding schemes, having high SNR packets may reduce the number of detection and decoding iteration, and hence reduce the processing power [281]. Similarly, if the transmitter has low energy level, then the receiver might increase the number of iterations to recover the data, and thus, reduce the number of required retransmissions. Another interesting scenario is when the transmitted data have unequal error importance, and error concealment techniques are employed. Video coding is an example for such scenario where the I-frames are generally more important than P-frames, and lost I-frames can be recovered using error concealment techniques. If an P- or B-frame is damaged or lost, the receiver can voluntarily chose to send a false ACK to the transmitter, to save the transmitter energy. The receiver may use error concealment to recover or substitute the damaged frame.

Buffer occupancy: Video traffic constituted about 80% of the Internet traffic in 2019. Moreover, video communications requires high data rate transmission and processing large amounts of data, which may quickly deplete the batteries of the transmitter and receiver. In the case that the playback buffer at the receiver is empty, the video player may pause, causing watching interruption. Therefore, the receiver might chose to

send a false ACK for less important video frame types to prevent buffer starvation. It is worth noting that the receiver in this case may behave autonomously in the sense that it is the one who decides the type of the feedback message regardless of the decoding process outcome. On the other hand, if the buffer is nearly full, the receiver may decide to behave selfishly and send false NACKs to the transmitter. The receiver then combines these packets which will increase the probability of having smooth video playing.

User attentiveness: AI is currently widely used to estimate the human behavior, which is extensively adopted in intelligent transportation systems and smartphones. For example, the smart stay feature in some smart devices uses the front camera to sense when the user is looking at the device, and it keeps the screen on regardless of the screen timeout setting. In a more intelligent design, the device can also control the quality of media being presented to the user based on the user attentiveness level, by controlling the ARQ process. If the device identifies that the user is not paying attention while streaming a video, for instance, the receiver can reduce the video quality by sending false ACKs to save the transmission power and bandwidth by dropping some video frames instead of asking for retransmitting them. Similarly, the device microphone can be used to detect the background noise to adjust the quality of the played audio in the case that ear-speakers are not used.

User safety or comfort level: This scenario is expected in case of disasters where mobile users are in critical need for energy saving. In this case, the base station can increase the power for the transmitted and retransmitted packets to save the energy for the mobile user by avoiding sending frequent feedback messages or performing substantial signal processing.

C. Non-orthogonal multiplexing (NOM)

As indicated by several studies, retransmitted packets require much less power than the initial transmission due to the CC. Therefore, the reduced power of the retransmitted packets enables multiplexing new packets with the retransmitted packets. As a result, significant throughput improvement can be achieved. Inspired by NOMA, the authors of [282] proposed using NOM to perform the multiplexing in such scenarios. Nevertheless, the work mostly presents a proof of concept without any optimization.

IX. Conclusion

This survey considered various retransmission schemes in terms of advantages, limitations, integration with emerging technologies, adoption in industrial standards and future directions. As can be noted from the presented literature, using HARQ is essential to provide reliable communications for most legacy and emerging technologies. However, the large delay, increased power consumption, and reduced throughput are among the main HARQ limitations that most research work tried to address.

Although the work reported in the literature has covered a broad range of aspects, the work to develop more efficient HARQ schemes is expected to continue due to the nature of the new applications. More specifically, new HARQ schemes should be more intelligent by considering all the operating conditions while into the retransmission protocol. Therefore, AI is expected to play a major role in the future to collect information about the user vigilance and preferences, in addition to the application type while configuring the transmission protocol parameters.

Appendix I: Acronyms

F_M	feedback message.
3G	third generation.
3GPP	3rd generation partnership project.
4G	fourth generation.
5G	fifth generation.
6G	sixth generation.
AC	access category.
ACK	acknowledgment.
ACM	adaptive coding and modulation.
AF	amplify-and-forward.
AI	artificial intelligence.
AM	acknowledged mode.
AMD	acknowledged mode data.
AMI	accumulated mutual information.
AoI	age of information.
AP	access point.
ARQ	automatic repeat request.
AWGN	additive white Gaussian noise.
BAck	block ACK.
BCH	Bose–Chaudhuri–Hocquenghem.
BER	bit error rate.
BLC	bit level combining.
BLER	bLock error rate.
bps	bits per second.
BPSK	binary phase shift keying.
BPU	baseband processing unit.
BS	base station.
BTFDPS	block time-frequency domain packet scheduling.
CBG	codeblock-group.
CC	chase combining.
CDMA	code division multiple access.
CQI	channel quality indicator.
CR	cognitive radio.
CRAN	cloud-radio access network.
CRC	cyclic redundancy check.
CSI	channel state information.
D2D	device-to-device.

DCF	decode-and-forward.
DF	demodulated-and-forward.
DI	deferred iterations.
DL	downlink.
E-HARQ	early HARQ.
E-UTRA	evolved universal terrestrial radio access.
ED	error detection.
EE	energy efficiency.
eNB	evolved node B.
ES	early stop.
EUTRAN	evolved universal terrestrial radio access network.
FDD	frequency division duplexing.
FEC	forward error correction.
FER	frame error rate.
fps	frames per second.
GBN	go-back-N.
GoP	group of pictures.
GSM	global system for mobile communications.
HARQ	hybrid-ARQ.
HDD	hard decision decoding.
HEVC	high efficiency video coding.
HIHO	hard-input hard-output.
HR	high-rate.
HSDPA	high speed downlink packet access.
HSPA+	evolved high speed packet access.
HSUPA	high speed uplink packet access.
i.i.d.	independent and identically distributed.
ICI	inter-carrier interference.
IoT	internet of things.
IR	incremental redundancy.
ITU	international telecommunication union.
LDPC	low density parity check.
LFSRs	linear feedback shift registers.
LLR	log-likelihood ratio.
LR	low-rate.
LTE	long term evolution.
MAC	medium access control.
MBS	macro base station.
MCS	modulation and coding scheme.
MIB	management information base.
MIMO	multiple-input multiple-output.
mm-Wave	millimeter wave.
mMIMO	massive MIMO.
mMTC	massive machine type communication.
MPDU	MAC protocol data unit.
MPEG	moving picture experts group.
MPSK	M-ary phase shift keying.
MRC	maximum ratio combining.
MSE	mean squared error.
MTC	machine type communication.
NACK	negative acknowledgment.
NB	narrow band.
NOM	non-orthogonal multiplexing.
NOMA	non-orthogonal multiple access.
NR	new radio.
PDR	packet drop rate.
PDU	protocol data unit.
PEC	parity error checking.
PER	packet error rate.
PHY	physical layer.
pps	packets per second.
PRB	physical resource block.
PSDU	physical data service unit.
PSNR	peak signal to noise ratio.
PTD	packet time delay.
QAM	quadrature amplitude modulation.
QoS	quality of service.
QP	quantization parameter.
QPSK	quadrature phase shift keying.
RAU	radio access unit.
RF	radio frequency.
RL	reinforcement learning.
RLC	radio link control.
RRC	radio resources control.
RRH	remote radio head.
RTT	round trip time.
SBS	small base station.
SCDMA	synchronous CDMA.
SDD	soft decision decoding.
SER	symbol error rate.
SIC	successive interference cancellation.
SINR	signal to interference plus noise ratio.
SISO	soft-input soft-output.
SLC	symbol level combining.
SLRC	station long retry count.
SN	sequence number.

SNMP	simple network management protocol.
SNR	signal to noise ratio.
SPID	subpacket identifier.
SR	selective repeat.
SSRC	station short retry count.
STA	wireless station.
STD	standard deviation.
SW	stop-and-wait.
SWIPT	simultaneous wireless information and power transmission.
TB	transport block.
TC	traffic category.
TCP	transmission control protocol.
TD	time division.
TDD	time division duplexing.
TDPS	time domain packet scheduling.
TM	transmissions per message.
TNR	transmission number relaying.
TPC	turbo product code.
TTI	transmission time interval.
UAV	unmanned aerial vehicle.
UE	user equipment.
UL	uplink.
UM	unacknowledged mode.
UMTS	universal mobile telecommunications system.
URLLC	ultra reliable low latency communication.
VoIP	voice over internet protocol.
W-CDMA	wideband-CDMA.
WAP	wireless access point.
WiFi	wireless fidelity.
WiMAX	worldwide interoperability for microwave access.
WLAN	wireless local area network.
WPAN	wireless personal area network.
WSN	wireless sensor network.

APPENDIX II: LIST OF SYMBOLS

B	Bandwidth
C	Maximum number of transmissions
C_r	Total number of transmissions of a selected relay
E_ℓ^n	Amount of energy consumed by n^{th} node during ℓ^{th} transmission attempt
E_{rx}	Energy consumed during the reception of the data packets
E_{tot}	Total end-to-end consumed energy
E_{tx}	Energy consumed during the transmission
I	Mutual information
K	Total number of bits to be transmitted
L	Total number of codewords to be transmitted
M	Modulation index
N	Number of symbols per packet
N_t	Total number of time slots
PER	Packet error rate
P_{out}	Outage probability
Q	Number of symbols per codeword
R	Code rate
R'	Residual rate
R_c	Effective rate
R_o	Initial code rate
R_t	Target rate
R_x	Total number of receivers
S	Maximum number of sources in a relay system
T	Time slot duration
TM	Transmissions per message
T_D	Time delay
T_d	Acknowledgment time
T_p	Transmission/propagation time
T_s	Sensing time duration
T_{max}	Maximum transmissions during second round in cooperative communication
T_{thr}	Packet delay threshold
W	Transmitter buffer size
Z^{max}	Total number of transmitted packets
Z^{tx}	Maximum allowable packets to be transmitted in a single time slot
Υ	Number of allocated resources
α	Ratio of number of channels uses between source and relays/destinations (first phase) to the number of channels uses between the source/relays and destinations/relays (second phase)
$\bar{B}$	Number of successfully received bits
$\bar{C}$	Average number of transmissions
$\bar{N}_s$	Total number of channel uses
$\bar{T}_d$	Round-trip time
$\bar{Z}$	Successfully received packets
$\bar{Z}_t^{max}$	Total number of packets that are transmitted exactly t times
$\bar{\gamma}_{u,i}^{x_{v,i}}$	Signal to interference plus noise ratio for user u during i^{th} round when user v transmits a symbol $x_{v,i}$ during same round
ℓ	Current transmission number
η	Throughput
γ	SNR
$\mathbb{E}$	Expectation operator
$\mathbb{P}_f$	Probability of decoding failure
$\mathbf{B}$	Interleaver output
$\mathbf{H}$	Channel frequency response
$\mathbf{U}$	A set of buffered L codewords
$\mathbf{c}$	Error detection encoder output
$\mathbf{d}$	Binary information block
$\mathbf{p}$	Parity bits for error correction
$\mathbf{r}$	Received packet
$\mathbf{r}_\Sigma$	Combined signal
$\mathbf{s}$	Modulated codeword
$\mathbf{u}$	Codeword
$\mathbf{w}$	AWGN vector
$\mathbf{z}$	Transmitted packet
$\mathcal{B}$	Total number of bits per packet

$\mathcal{P}$	Number of physical resource blocks
$\mathcal{S}$	Set of sources in relay system
ω	Combining factor
ϕ	Fraction of allocated resources
σ_H^2	Channel variance
σ_w^2	Noise variance
$\tilde{E}_{rx}$	Energy consumed during the reception the feedback messages
$\tilde{E}_{tx}$	Energy consumed during the transmission of the feedback messages
ζ	Energy efficiency
d_D	Decoding delay
e_D	Encoding delay
h	Instantaneous channel gain
k	Number of information bits per codeword
l_{ec}	Number of parity bits for error correction
l_{ed}	Number of bits for error detection
n	Total number of bits per codeword
n_s	Number of subcarriers in a physical resource block
n_t	Number of symbols to be transmitted in one round trip
p	Transmission power
s	Information symbols
t	Transmission round index
t_{out}	Specified time limit (timeout)

REFERENCES

[1] M. A. Al-Jarrah, M. A. Yaseen, A. Al-Dweik, O. A. Dobre, and E. Alsusa, "Decision fusion for IoT-based wireless sensor networks," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1313–1326, Feb. 2020.

[2] X. Xu, Y. Pan, P. P. M. Y. Lwin, and X. Liang, "3D holographic display and its data transmission requirement," in *2011 Int. Conf. inf. Photon. Opt. Commun.*, 2011, pp. 1–4.

[3] T. Janevski, *Quality of service regulation manual*, Intl. Telecommun. Union (ITU), Geneva, Switserland, 2017.

[4] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[5] J. Proakis and M. Salehi, *Digital Communications*, 5th ed. USA: McGraw-Hill Education, 2008.

[6] H. Mukhtar, A. Al-Dweik, and A. Shami, "Turbo product codes: Applications, challenges, and future directions," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 3052–3069, 2016.

[7] "3GPP 36 series of specifications: Evolved Universal Terrestrial Radio Access (E-UTRA)," https://www.3gpp.org/ftp/Specs/html-info/36-series.htm, accessed: 2020-06-14.

[8] "IEEE Standard for Air Interface for Broadband Wireless Access Systems," *IEEE Std 802.16-2017 (Revision of IEEE Std 802.16-2012)*, pp. 1–2726, 2018.

[9] "3GPP 38 series of specifications: 5G New Radio," https://www.3gpp.org/DynaReport/38-series.htm, accessed: 2020-08-18.

[10] "IEEE Standard for Information technology—Telecommunications and information exchange between systems Local and metropolitan area networks—Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pp. 1–3534, 2016.

[11] Y. . E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A primer on 3GPP narrowband internet of things," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 117–123, 2017.

[12] H. A. Ngo and L. Hanzo, "Hybrid automatic-repeat-request systems for cooperative wireless communications," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 25–45, 2014.

[13] H. Chen, R. G. Maunder, and L. Hanzo, "A survey and tutorial on low-complexity turbo coding techniques and a holistic hybrid ARQ design example," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1546–1566, 2013.

[14] S. Jiang, "On reliable data transfer in underwater acoustic networks: A survey from networking perspective," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 1036–1055, 2018.

[15] H. Mukhtar, A. Al-Dweik, M. Al-Mualla, and A. Shami, "Low complexity power optimization algorithm for multimedia transmission over wireless networks," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 113–124, Feb. 2015.

[16] Y. Iraqi and A. Al-Dweik, "Efficient information transmission using smart OFDM for IoT applications," *IEEE Internet Things J., IEEE early access*, pp. 1–1, 2020.

[17] ——, "Adaptive bit loading with reduced computational time and complexity for multicarrier wireless communications," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 3, pp. 2497–2506, June 2020.

[18] F. Kalbat, A. Al-Dweik, Y. Iraqi, H. Mukhtar, B. Sharif, and G. K. Karagiannidis, "Direct bit loading with reduced complexity and overhead for precoded OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 7169–7173, July 2019.

[19] A. Al-Dweik and B. Sharif, "Non-sequential decoding algorithm for hard iterative turbo product codes," *IEEE Trans. Commun.*, vol. 57, no. 6, pp. 1545–1549, 2009.

[20] A. Al-Dweik, S. Le Goff, and B. Sharif, "A hybrid decoder for block turbo codes," *IEEE Trans. Commun.*, vol. 57, no. 5, pp. 1229–1232, 2009.

[21] A. Al-Dweik, H. Mukhtar, E. Alsusa, and J. Dias, "Ultra-light decoder for turbo product codes," *IEEE Commun. Lett.*, vol. 22, no. 3, pp. 446–449, 2018.

[22] A. Al-Dweik and B. Sharif, "Closed-chains error correction technique for turbo product codes," *IEEE Trans. Commun.*, vol. 59, no. 3, pp. 632–638, 2011.

[23] C. Simsek and K. Turk, "Simplified early stopping criterion for belief-propagation polar code decoders," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1515–1518, 2016.

[24] T. Xia, H. Wu, and H. Jiang, "New stopping criterion for fast low-density parity-check decoders," *IEEE Commun. Lett.*, vol. 18, no. 10, pp. 1679–1682, 2014.

[25] G. Han and X. Liu, "A unified early stopping criterion for binary and nonbinary LDPC codes based on check-sum variation patterns," *IEEE Commun. Lett.*, vol. 14, no. 11, pp. 1053–1055, 2010.

[26] B. Liu, G. Dou, W. Tao, and J. Gao, "Efficient stopping criterion for hybrid weighted symbol-flipping decoding of nonbinary LDPC codes," *IEEE Commun. Lett.*, vol. 15, no. 3, pp. 337–339, 2011.

[27] R. Y. Shao, Shu Lin, and M. P. C. Fossorier, "Two simple stopping criteria for turbo decoding," *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1117–1120, 1999.

[28] M. Ferrari, S. Bellini, and A. Tomasoni, "Safe early stopping for layered LDPC decoding," *IEEE Commun. Lett.*, vol. 19, no. 3, pp. 315–318, 2015.

[29] Y. Wu, B. D. Woerner, and W. J. Ebel, "A simple stopping criterion for turbo decoding," *IEEE Commun. Lett.*, vol. 4, no. 8, pp. 258–260, 2000.

[30] W. Zhanji, P. Mugen, and W. Wenbo, "A new parity-check stopping criterion for turbo decoding," *IEEE Commun. Lett.*, vol. 12, no. 4, pp. 304–306, 2008.

[31] L. Huang, Q. T. Zhang, and L. L. Cheng, "Information theoretic criterion for stopping turbo iteration," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 848–853, 2011.

[32] B. Yuan and K. K. Parhi, "Early stopping criteria for energy-efficient low-latency belief-propagation polar code decoders," *IEEE Trans. Signal Process.*, vol. 62, no. 24, pp. 6496–6506, 2014.

[33] G. T. Chen, L. Cao, L. Yu, and C. W. Chen, "An efficient stopping criterion for turbo product codes," *IEEE Commun. Lett.*, vol. 11, no. 6, pp. 525–527, 2007.

[34] C. Lin and C. Wei, "Efficient window-based stopping technique for double-binary turbo decoding," *IEEE Commun. Lett.*, vol. 17, no. 1, pp. 169–172, 2013.

[35] X. Liu, J. Cai, and L. Wu, "Improved decoding algorithm of serial belief propagation with a stop updating criterion for LDPC codes and applications in patterned media storage," *IEEE Trans. Magn.*, vol. 49, no. 2, pp. 829–836, 2013.

[36] Y. Yan, X. Zhang, and B. Wu, "Simplified early stopping criterion for belief-propagation polar code decoder based on frozen bits," *IEEE Access*, vol. 7, pp. 134 691–134 696, 2019.

[37] J. Zhang and M. Wang, "Belief propagation decoder with multiple bit-flipping sets and stopping criteria for polar codes," *IEEE Access*, vol. 8, pp. 83 710–83 717, 2020.

[38] F. Li and A. Wu, "On the new stopping criteria of iterative turbo decoding by using decoding threshold," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5506–5516, 2007.

[39] D. Kim and I. Park, "A fast successive cancellation list decoder for polar codes with an early stopping criterion," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4971–4979, 2018.

[40] Fengqin Zhai and I. J. Fair, "Techniques for early stopping and error detection in turbo decoding," *IEEE Trans. Commun.*, vol. 51, no. 10, pp. 1617–1623, 2003.

[41] Jin Li, Xiao-hu You, and Jing Li, "Early stopping for LDPC decoding: convergence of mean magnitude (CMM)," *IEEE Commun. Lett.*, vol. 10, no. 9, pp. 667–669, 2006.

[42] H. Mukhtar, A. Al-Dweik, M. Al-Mualla, and A. Shami, "Adaptive hybrid ARQ system using turbo product codes with hard/soft decoding," *IEEE Commun. Lett.*, vol. 17, no. 11, pp. 2132–2135, 2013.

[43] B. Zhang, P. Cosman, and L. Milstein, "Energy optimization for hybrid ARQ with turbo coding: Rate adaptation and allocation," *IEEE Trans. Veh. Technol.*, pp. 1–1, 2020.

[44] B. Zhang, H. Chen, M. El-Hajjar, R. Maunder, and L. Hanzo, "Distributed multiple-component turbo codes for cooperative hybrid ARQ," *IEEE Signal Process. Lett.*, vol. 20, no. 6, pp. 599–602, 2013.

[45] R. D. Souza, M. E. Pellenz, and T. Rodrigues, "Hybrid ARQ scheme based on recursive convolutional codes and turbo decoding," *IEEE Trans. Commun.*, vol. 57, no. 2, pp. 315–318, 2009.

[46] T. Shi and L. Cao, "On the performance of turbo codes-based hybrid ARQ with segment selective repeat in WCDMA," *J. Commun. Netw.*, vol. 8, no. 2, pp. 212–219, 2006.

[47] L. Cao and T. Shi, "Turbo codes based hybrid ARQ with segment selective repeat," *Electron. Lett.*, vol. 40, no. 18, pp. 1140–1141, 2004.

[48] Tingfang Ji and W. E. Stark, "Turbo-coded ARQ schemes for DS-CDMA data networks over fading and shadowing channels: throughput, delay, and energy efficiency," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 8, pp. 1355–1364, 2000.

[49] D. N. Rowitch and L. B. Milstein, "On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo (RCPT) codes," *IEEE Trans. Commun.*, vol. 48, no. 6, pp. 948–959, 2000.

[50] K. R. Narayanan and G. L. Stuber, "A novel ARQ technique using the turbo coding principle," *IEEE Commun. Lett.*, vol. 1, no. 2, pp. 49–51, 1997.

[51] H. Mukhtar, A. Al-Dweik, and M. Al-Mualla, "Hybrid ARQ with partial retransmission using turbo product codes," in *2015 Intl. Conf. inf. Commun. Technol. Res. (ICTRC)*, 2015, pp. 28–31.

[52] Jhi-Siang Liao, Y. T. Su, and Shen-Chih Chen, "Type II ARQ schemes based on turbo product codes," in *Global Telecommun. Conf., 2002. GLOBECOM '02. IEEE*, vol. 1, 2002, pp. 846–850 vol.1.

[53] Sunheui Ryoo, Sooyoung Kim, and Sung Pal Lee, "Hybrid ARQ using rate compatible block turbo codes for mobile satellite systems," in *2003 IEEE 58th Veh. Technol. Conf.. VTC 2003-Fall (IEEE Cat. No.03CH37484)*, vol. 4, 2003, pp. 2673–2677 Vol.4.

[54] P. Paul, A. Kumar, and K. C. Roy, "Throughput analysis on a scheme of product codes for ARQ protocol," in *2010 15th IEEE Intl. Workshop Comput. Aided Modeling, Anal. Des. Commun. Links Netw. (CAMAD)*, 2010, pp. 36–40.

[55] H. Mukhtar, A. Al-Dweik, and M. Al-Mualla, "CRC-free hybrid ARQ system using turbo product codes," *IEEE Trans. Commun.*, vol. 62, no. 12, pp. 4220–4229, 2014.

[56] ——, "Low complexity hybrid ARQ using extended turbo product codes self-detection," in *2015 IEEE Global Commun. Conf. (GLOBECOM)*, 2015, pp. 1–6.

[57] C. Kim, S. Kim, and J. No, "New GRP LDPC codes for H-ARQ-IR over the block fading channel," *IEEE Trans. Commun.*, pp. 1–1, 2020.

[58] M. Xie, Q. Wang, J. Gong, and X. Ma, "Age and energy analysis for LDPC coded status update with and without ARQ," *IEEE Internet Things J.*, pp. 1–1, 2020.

[59] H. Wang, S. V. S. Ranganathan, and R. D. Wesel, "Variable-length coding with shared incremental redundancy: Design methods and examples," *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 5981–5995, 2019.

[60] T. Shafique, M. Zia, and H. Han, "Channel constrained multiple selective retransmissions for OFDM system: BER and throughput analysis," *IEEE Access*, vol. 7, pp. 4317–4326, 2019.

[61] T. Shafique, M. Zia, H. Han, and H. Mahmood, "Cross-layer chase combining with selective retransmission, analysis, and throughput optimization for OFDM systems," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2311–2325, 2016.

[62] K. Vakilinia, S. V. S. Ranganathan, D. Divsalar, and R. D. Wesel, "Optimizing transmission lengths for limited feedback with nonbinary LDPC examples," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2245–2257, 2016.

[63] Z. Muhammad, H. Mahmood, A. Ahmed, and N. A. Saqib, "Selective HARQ transceiver design for OFDM system," *IEEE Commun. Lett.*, vol. 17, no. 12, pp. 2229–2232, 2013.

[64] "IEEE standard for air interface for broadband wireless access systems," *IEEE Std 802.16-2012 (Revision of IEEE Std 802.16-2009)*, pp. 1–2542, 2012.

[65] A. G. D. Uchôa, R. D. Souza, and M. E. Pellenz, "Hybrid ARQ with partial retransmissions and LDPC codes and its impact on TCP," *IEEE Latin Amer. Trans.*, vol. 8, no. 4, pp. 417–424, 2010.

[66] M. El-Khamy, J. Hou, and N. Bhushan, "Design of rate-compatible structured LDPC codes for hybrid ARQ applications," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 6, pp. 965–973, 2009.

[67] "IEEE standard for local and metropolitan area networks part 16: Air interface for broadband wireless access systems," *IEEE Std 802.16-2009 (Revision of IEEE Std 802.16-2004)*, pp. 1–2080, 2009.

[68] Y. Cao, J. Gu, L. Qi, and D. Yang, "Degree distribution based HARQ for irregular LDPC," *Electron. Lett.*, vol. 42, no. 6, pp. 363–364, 2006.

[69] S. Sesia, G. Caire, and G. Vivier, "Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes," *IEEE Trans. Commun.*, vol. 52, no. 8, pp. 1311–1321, 2004.

[70] M. R. Yazdani and A. H. Banihashemi, "On construction of rate-compatible low-density parity-check codes," *IEEE Commun. Lett.*, vol. 8, no. 3, pp. 159–161, 2004.

[71] K. Chen, K. Niu, and J. Lin, "A hybrid ARQ scheme based on polar codes," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1996–1999, 2013.

[72] M. S. Mohammadi, I. B. Collings, and Q. Zhang, "Simple hybrid ARQ schemes based on systematic polar codes for IoT applications," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 975–978, 2017.

[73] H. Liang, A. Liu, Y. Zhang, and X. Liang, "Efficient design of multi-packet hybrid ARQ transmission scheme based on polar codes," *IEEE Access*, vol. 6, pp. 31 564–31 570, 2018.

[74] H. Liang, A. Liu, Y. Zhang, and Q. Zhang, "Analysis and adaptive design of polar coded HARQ transmission under SC-list decoding," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 798–801, 2017.

[75] K. Wang and Z. Ding, "Diversity integration in hybrid-ARQ with chase combining under partial CSI," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2647–2659, 2016.

[76] H. Liang, A. Liu, Y. Zhang, and X. Liang, "Corrections to "efficient design of multi-packet hybrid ARQ transmission scheme based on polar code"," *IEEE Access*, vol. 8, pp. 135 296–135 297, 2020.

[77] J. Gao, P. Fan, and L. Li, "Optimized polarizing matrix extension based HARQ scheme for short packet transmission," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 951–955, 2020.

[78] H. Saber and I. Marsland, "An incremental redundancy hybrid ARQ scheme via puncturing and extending of polar codes," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 3964–3973, 2015.

[79] M. Grymel and S. B. Furber, "A novel programmable parallel CRC circuit," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 10, pp. 1898–1902, 2011.

[80] G. D. Nguyen, "Fast CRCs," *IEEE Trans. Comput.*, vol. 58, no. 10, pp. 1321–1331, 2009.

[81] X. Zhang, "A low-power parallel architecture for linear feedback shift registers," *IEEE Trans. Circuits Syst., II, Exp. Briefs*, vol. 66, no. 3, pp. 412–416, 2019.

[82] M. Ayinala and K. K. Parhi, "High-speed parallel architectures for linear feedback shift registers," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4459–4469, 2011.

[83] J. Jung, H. Yoo, Y. Lee, and I. Park, "Efficient parallel architecture for linear feedback shift registers," *IEEE Trans. Circuits Syst., II, Exp. Briefs*, vol. 62, no. 11, pp. 1068–1072, 2015.

[84] P. Coulton, C. Tanriover, B. Wright, and B. Honary, "Simple hybrid type II ARQ technique using soft output information," *Electron. Lett.*, vol. 36, no. 20, pp. 1716–1717, 2000.

[85] M. E. Buckley and S. B. Wicker, "The design and performance of a neural network for predicting turbo decoding error with application to hybrid ARQ protocols," *IEEE Trans. Commun.*, vol. 48, no. 4, pp. 566–576, 2000.

[86] J. C. Fricke and P. A. Hoeher, "Reliability-based retransmission criteria for hybrid ARQ," *IEEE Trans. Commun.*, vol. 57, no. 8, pp. 2181–2184, 2009.

[87] B. B. Agarwal and S. P. Tayal, *Computer Network*. Laxmi Publications, 2009.

[88] B. Forouzan, *Data communications and networking*, 5th ed. New York, NY, USA: McGraw-Hill, 2013.

[89] S. Ahmadi, Ed., *LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies*. Elsevier, 2014.

[90] H. Long, W. Xiang, S. Shen, Y. Zhang, K. Zheng, and W. Wang, "Analysis of conditional error rate and combining schemes in HARQ," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2677–2682, 2012.

[91] A. B. Sediq and H. Yanikomeroglu, "Selection combining of signals with different modulation levels in nakagami-m fading," *IEEE Commun. Lett.*, vol. 16, no. 5, pp. 752–755, 2012.

[92] A. Bin Sediq and H. Yanikomeroglu, "Performance analysis of selection combining of signals with different modulation levels in cooperative communications," *IEEE Trans. Veh. Technol.*, vol. 60, no. 4, pp. 1880–1887, 2011.

[93] A. B. Sediq and H. Yanikomeroglu, "Performance analysis of soft-bit maximal ratio combining in cooperative relay networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 4934–4939, 2009.

[94] R. Yazdani and M. Ardakani, "Efficient LLR calculation for non-binary modulations over fading channels," *IEEE Trans. Commun.*, vol. 59, no. 5, pp. 1236–1241, 2011.

[95] S. Liu, X. Wu, Y. Xi, and J. Wei, "On the throughput and optimal packet length of an uncoded ARQ system over slow rayleigh fading channels," *IEEE Commun. Lett.*, vol. 16, no. 8, pp. 1173–1175, 2012.

[96] S. Hara, A. Ogino, M. Araki, M. Okada, and N. Morinaga, "Throughput performance of SAW-ARQ protocol with adaptive packet length in mobile packet data transmission," *IEEE Trans. Veh. Technol.*, vol. 45, no. 3, pp. 561–569, 1996.

[97] M. S. Mohammadi, Q. Zhang, E. Dutkiewicz, and X. Huang, "Optimal frame length to maximize energy efficiency in IEEE 802.15.6 UWB body area networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, pp. 397–400, 2014.

[98] K. Vakilinia, S. V. S. Ranganathan, D. Divsalar, and R. D. Wesel, "Optimizing transmission lengths for limited feedback with nonbinary LDPC examples," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2245–2257, 2016.

[99] Yiqing Zhou and Jiangzhou Wang, "Optimum subpacket transmission for hybrid ARQ systems," *IEEE Trans. Commun.*, vol. 54, no. 5, pp. 934–942, 2006.

[100] C. . Cho, J. . Won, and H. . Lee, "Performance of hybrid II ARQ schemes using punctured RS code for wireless ATM," *IEEE Proc. - Commun.*, vol. 148, no. 4, pp. 229–233, 2001.

[101] S. B. Wicker and M. J. Bartz, "Type-II hybrid-ARQ protocols using punctured MDS codes," *IEEE Trans. Commun.*, vol. 42, no. 234, pp. 1431–1440, 1994.

[102] S. Kallel and D. Haccoun, "Generalized type II hybrid ARQ scheme using punctured convolutional coding," *IEEE Trans. Commun.*, vol. 38, no. 11, pp. 1938–1946, 1990.

[103] R. Prasad and M. Ruggieri, *Technology Trends in Wireless Communications*, ser. Artech House universal pers. commun. ser. Artech House, 2003. [Online]. Available: https://books.google.ae/books?id=DmA0d_B8dZcC

[104] S. Kallel, "Complementary punctured convolutional (CPC) codes and their applications," *IEEE Trans. Commun.*, vol. 43, no. 6, pp. 2005–2009, 1995.

[105] Jung-Fu Cheng, "Coding performance of hybrid ARQ schemes," *IEEE Trans. Commun.*, vol. 54, no. 6, pp. 1017–1029, 2006.

[106] S. R. Khosravirad, L. Szczecinski, and F. Labeau, "Rate allocation for HARQ in relay-based cooperative transmission," in *IEEE Wireless Commun. Netw. Conf., WCNC*, vol. 3, 2014, pp. 2757–2762.

[107] A. U. Rehman, L. Yang, and L. Hanzo, "Performance of cognitive hybrid automatic repeat request: Stop-and-wait," in *2015 IEEE 81st Veh. Technol. Conf. (VTC Spring)*, 2015, pp. 1–5.

[108] A. U. Rehman, C. Dong, L. L. Yang, and L. Hanzo, "Performance of cognitive stop-and-wait hybrid automatic repeat request in the face of imperfect sensing," *IEEE Access*, vol. 4, pp. 5489–5508, 2016.

[109] A. U. Rehman, C. Dong, V. A. Thomas, L. L. Yang, and L. Hanzo, "Throughput and delay analysis of cognitive Go-Back-N hybrid automatic repeat request using discrete-time markov modelling," *IEEE Access*, vol. 4, pp. 9659–9680, 2016.

[110] A. U. Rehman, L. Yang, and L. Hanzo, "Performance of cognitive hybrid automatic repeat request: Go-Back-N," in *2016 IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, 2016, pp. 1–5.

[111] A. U. Rehman, V. A. Thomas, L. L. Yang, and L. Hanzo, "Performance of cognitive selective-repeat hybrid automatic repeat request," *IEEE Access*, vol. 4, pp. 9828–9846, 2016.

[112] S. Cerovic, R. Visoz, L. Madier, and A. O. Berthet, "Centralized scheduling strategies for cooperative HARQ retransmissions in multi-source multi-relay wireless networks," in *2018 IEEE Intl. Conf. Commun. (ICC)*, 2018, pp. 1–6.

[113] F. Maliqi, F. Bassi, P. Duhamel, and I. Limani, "A probabilistic HARQ protocol for demodulate-and-forward relaying networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1623–1636, 2019.

[114] Q. Li, X. Zhang, A. Pandharipande, and J. Zhang, "Cooperative spectrum sharing on SWIPT-based DF relay: An energy-aware retransmission approach," *IEEE Access*, vol. 7, pp. 120 802–120 816, 2019.

[115] L. Nasraoui, L. N. Atallah, and M. Siala, "Throughput maximization with optimum energy allocation for ARQ retransmission protocol," in *2018 Seventh Intl. Conf. Commun. Netw. (ComNet)*, 2018, pp. 1–5.

[116] Y. H. Jung and J. Choi, "Hybrid ARQ scheme with autonomous retransmission for multicasting in wireless sensor networks," *Sensors (Switzerland)*, vol. 17, no. 3, 2017.

[117] H. El Gamal, G. Caire, and M. O. Damen, "The MIMO ARQ channel: Diversity–multiplexing–delay tradeoff," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3601–3621, 2006.

[118] F. A. de Souza, R. D. Souza, G. Brante, M. E. Pellenz, and F. Rosas, "Code rate, frequency and snr optimization for energy efficient underwater acoustic communications," in *2015 IEEE Intl. Conf. Commun. (ICC)*, 2015, pp. 6351–6356.

[119] M. Deghel, S. E. Elayoubi, A. Galindo-Serrano, and R. Visoz, "Joint optimization of link adaptation and HARQ retransmissions for URLLC services," in *2018 25th Intl. Conf. Telecommun. (ICT)*, 2018, pp. 21–26.

[120] L. Buccheri, S. Mandelli, S. Saur, L. Reggiani, and M. Magarini, "Hybrid retransmission scheme for QoS-defined 5G ultra-reliable low-latency communications," in *2018 IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2018, pp. 1–6.

[121] R. Sattiraju and H. D. Schotten, "Reliability modeling, analysis and prediction of wireless mobile communications," in *2014 IEEE 79th Veh. Technol. Conf. (VTC Spring)*, 2014, pp. 1–6.

[122] L. Zheng, N. Lu, and L. Cai, "Reliable wireless communication networks for demand response control," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 133–140, 2013.

[123] Y. Xu, D. Cai, F. Fang, Z. Ding, C. Shen, and G. Zhu, "Outage constrained power efficient design for downlink NOMA systems with partial HARQ," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5188–5201, 2020.

[124] R. Wang, F. Zhou, J. Bian, K. An, and K. Guo, "Performance evaluation of HARQ-assisted hybrid satellite-terrestrial relay networks," *IEEE Commun. Lett.*, vol. 24, no. 2, pp. 423–427, 2019.

[125] Z. Shi, C. Zhang, Y. Fu, H. Wang, G. Yang, and S. Ma, "Achievable diversity order of HARQ-aided downlink NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 471–487, 2020.

[126] Z. Shi, C. Zhang, Y. Fu, H. Wang, G. Yang, and S. Ma, "Diversity analysis of HARQ-CC-aided NOMA," in *2019 IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6.

[127] E. Dosti, M. Shehab, H. Alves, and M. Latva-aho, "Ultra reliable communication via optimum power allocation for HARQ retransmission schemes," *arXiv preprint arXiv:2002.11524*, 2020.

[128] F. Rosas, R. D. Souza, M. E. Pellenz, C. Oberli, G. Brante, M. Verhelst, and S. Pollin, "Optimizing the code rate of energy-constrained wireless communications with HARQ," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 191–205, 2015.

[129] T. Shafique, M. Zia, and H. D. Han, "Analysis and throughput optimization of selective chase combining for OFDM systems," *arXiv preprint arXiv:1503.05819*, 2015.

[130] M. C. Ilter, H. Yanikomeroglu, and P. A. Dmochowski, "BER upper bound expressions in coded two-transmission schemes with arbitrarily spaced signal constellations," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 248–251, 2015.

[131] I. S. Ansari, S. Al-Ahmadi, F. Yilmaz, M.-S. Alouini, and H. Yanikomeroglu, "A new formula for the BER of binary modulations with dual-branch selection over generalized-K composite fading channels," *IEEE Trans. Commun.*, vol. 59, no. 10, pp. 2654–2658, 2011.

[132] H. Zheng and H. Viswanathan, "Optimizing the ARQ performance in downlink packet data systems with scheduling," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 495–506, 2005.

[133] J. Gu, Y. Zhang, and D. Yang, "Modeling conditional FER for hybrid ARQ," *IEEE Commun. Lett.*, vol. 10, no. 5, pp. 384–386, 2006.

[134] G. Wang, J. Wu, and Y. R. Zheng, "An accurate frame error rate approximation of coded diversity systems with non-identical diversity branches," in *2014 IEEE Intl. Conf. Commun. (ICC)*, 2014, pp. 5312–5317.

[135] Q. Wang, P. Fu, and S. Zhang, "A design of hybrid automatic repeat request scheme based on systematic polar codes," in *2020 IEEE 10th*

Intl. Conf. Electron. Info. Emergency Commun. (ICEIEC), 2020, pp. 85–89.

[136] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 119–125, 2018.

[137] K. Geethu and A. Babu, "A hybrid ARQ scheme combining erasure codes and selective retransmissions for reliable data transfer in underwater acoustic sensor networks," *Eurasip j. Wireless Commun. Netw.*, vol. 2017, no. 1, pp. 1–18, 2017.

[138] J. Kim, K. Kim, and J. Lee, "Energy-efficient relay selection of cooperative HARQ based on the number of transmissions over rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 610–621, 2017.

[139] T. Shafique, H. Tabassum, and E. Hossain, "End-to-end energy-efficiency and reliability of UAV-assisted wireless data ferrying," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1822–1837, 2020.

[140] X. Du, Y. Sun, N. B. Shroff, and A. Sabharwal, "Balancing queueing and retransmission: Latency-optimal massive MIMO design," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2293–2307, 2020.

[141] M. K. Sharma and C. R. Murthy, "Distributed power control for multi-hop energy harvesting links with retransmission," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4064–4078, 2018.

[142] A. U. Rehman, L. Yang, and L. Hanzo, "Delay and throughput analysis of cognitive Go-Back-N HARQ in the face of imperfect sensing," *IEEE Access*, vol. 5, pp. 7454–7473, 2017.

[143] M. Maaz, P. Mary, and M. Hélard, "Delay outage probability in block fading channel and relay-assisted hybrid-ARQ network," *IEEE Wireless Commun. Lett.*, vol. 3, no. 2, pp. 129–132, 2013.

[144] ——, "Energy minimization in HARQ-I relay-assisted networks with delay-limited users," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 6887–6898, 2017.

[145] C. Sahin, L. Liu, E. Perrins, and L. Ma, "Delay-sensitive communications over IR-HARQ: Modulation, coding latency, and reliability," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 749–764, 2019.

[146] J. Choi, J. Ha, and H. Jeon, "On the energy delay tradeoff of HARQ-IR in wireless multiuser systems," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3518–3529, 2013.

[147] A. Anand and G. de Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2411–2421, 2018.

[148] M. Jabi, M. Benjillali, L. Szczecinski, and F. Labeau, "Energy efficiency of adaptive HARQ," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 818–831, 2016.

[149] S. Pfletschinger and M. Navarro, "Adaptive HARQ for imperfect channel knowledge," in *2010 Intl. ITG Conf. Source Channel Coding (SCC)*, 2010, pp. 1–6.

[150] S. Pfletschinger, D. Declercq, and M. Navarro, "Adaptive HARQ with non-binary repetition coding," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4193–4204, 2014.

[151] L. Szczecinski, P. Duhamel, and M. Rahman, "Adaptive incremental redundancy for HARQ transmission with outdated CSI," in *2011 IEEE Global Telecommun. Conf.-GLOBECOM 2011*, 2011, pp. 1–6.

[152] P. Weitkemper and H. Taoka, "Adaptive HARQ with memoryless relays," in *2011 IEEE Veh. Technol. Conf. (VTC Fall)*, 2011, pp. 1–5.

[153] S. R. Khosravirad, L. Szczecinski, and F. Labeau, "Rate-adaptive HARQ in relay-based cooperative transmission," in *2013 IEEE Intl. Conf. Commun. (ICC)*, 2013, pp. 5328–5333.

[154] E. Cabrera, G. Fang, and R. Vesilo, "Adaptive hybrid ARQ (A-HARQ) for ultra-reliable communication in 5G," in *2017 IEEE 85th Veh. Technol. Conf. (VTC Spring)*, 2017, pp. 1–6.

[155] H. Zhuang, "A high spectral efficiency hybrid ARQ protocol with low latency," in *2017 IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2017, pp. 1–6.

[156] ——, "Rate adaptive hybrid ARQ with optimal spectral efficiency and delay tradeoff," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4052–4064, 2017.

[157] H. Fares, B. Vrigneau, and O. Berder, "Green communication via HARQ protocols using message-passing decoder over AWGN channels," in *2015 IEEE 26th Annual Intl. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, 2015, pp. 197–201.

[158] H. Farès, B. Vrigneau, O. Berder, and P. Scalart, "Green communication via cooperative protocols using message-passing decoder over additive white gaussian noise channels," *IET Commun.*, vol. 11, no. 15, pp. 2320–2327, 2017.

[159] H. Mukhtar, A. Al-Dweik, and M. Al-Mualla, "On the performance of adaptive HARQ with no channel state information feedback," in *2015 IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2015, pp. 446–451.

[160] ——, "Content-aware and occupancy-based hybrid ARQ for video transmission," in *2016 IEEE 59th Intl. Midwest Symp. Circuits Syst. (MWSCAS)*, 2016, pp. 1–4.

[161] R. A. Ahmad, J. Lacan, F. Arnal, M. Gineste, and L. Clarac, "Enhancing satellite system throughput using adaptive HARQ for delay tolerant services in mobile communications," in *2015 Wireless Telecommun. Symp. (WTS)*, 2015, pp. 1–7.

[162] A. Rapaport, W. Liu, L. Ma, G. S. Sternberg, A. Ziera, and A. Balasubramanian, "Adaptive HARQ and scheduling for video over LTE," in *2013 Asilomar Conf. Signals, Syst. Comput.*, 2013, pp. 1584–1588.

[163] L. Szczecinski, S. R. Khosravirad, P. Duhamel, and M. Rahman, "Rate allocation and adaptation for incremental redundancy truncated HARQ," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2580–2590, 2013.

[164] J. Chen, G. Shen, and S. Jin, "An adaptive HARQ algorithm in MIMO systems," in *2009 IEEE 20th Intl. Symp. Pers., Indoor Mobile Radio Commun.*, 2009, pp. 2012–2015.

[165] R.-T. Juang, P. Ting, K.-Y. Lin, H.-P. Lin, and D.-B. Lin, "Adaptive HARQ for mobile wimax systems over time-varying channels," in *2009 IEEE Mobile WiMAX Symp.*, 2009, pp. 1–4.

[166] B. Mielczarek and W. A. Krzymien, "Adaptive hybrid ARQ systems with BCJR decoding," Jun. 29 2010, uS Patent 7,747,922.

[167] ——, "Adaptive hybrid ARQ systems with BCJR decoding," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1606–1619, 2008.

[168] J. Ramis and G. Femenias, "Cross-layer design of adaptive multirate wireless networks using truncated HARQ," *IEEE Trans. Veh. Technol.*, vol. 60, no. 3, pp. 944–954, 2011.

[169] D. Tuninetti, "On the benefits of partial channel state information for repetition protocols in block fading channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5036–5053, 2011.

[170] L. B. Le and E. Hossain, "Delay statistics for selective repeat ARQ protocol in multi-rate wireless networks with non-instantaneous feedback," in *GLOBECOM '05. IEEE Global Telecommun. Conf., 2005.*, vol. 5, 2005, pp. 5 pp.–2479.

[171] R. H. Chen, C. Hsiao, R. Chen, and W. Chung, "Adaptive HARQ scheme for reliable multicast communications," in *2012 IEEE 23rd Intl. Symp. Pers., Indoor Mobile Radio Commun. - (PIMRC)*, 2012, pp. 238–242.

[172] E. Uhlemann, L. K. Rasmussen, A. J. Grant, and P.-A. Wiberg, "Optimal incremental-redundancy strategy for type-II hybrid ARQ," in *IEEE Intl. Symp. Info. Theory, Pacifico Yokohama, Yokohama, Japan, June 29-July 4, 2003*, 2003, p. 448.

[173] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2475–2485, 2018.

[174] G. Alnwaimi and H. Boujemaa, "Adaptive packet length and MCS using average or instantaneous SNR," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10 519–10 527, 2018.

[175] J. Ramis and G. Femenias, "Cross-layer QoS-constrained optimization of adaptive multi-rate wireless systems using infrastructure-based cooperative ARQ," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2424–2435, 2013.

[176] Y. Yang, H. Ma, and S. Aissa, "Cross-layer combining of adaptive modulation and truncated ARQ under cognitive radio resource requirements," *IEEE Trans. Veh. Technol.*, vol. 61, no. 9, pp. 4020–4030, 2012.

[177] J. S. Harsini, F. Lahouti, M. Levorato, and M. Zorzi, "Analysis of non-cooperative and cooperative type II hybrid ARQ protocols with AMC over correlated fading channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 3, pp. 877–889, 2011.

[178] R. Zhang and L. Cai, "Joint AMC and packet fragmentation for error control over fading channels," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 3070–3080, 2010.

[179] W. Jiao, H. Ding, H. Wu, and G. Yu, "Spectrum efficiency of jointing adaptive modulation coding and truncated ARQ with QoS constraints," *IEEE Access*, vol. 6, pp. 46 915–46 925, 2018.

[180] X. Wang, I. Li, D. Wang, H. Zhuang, and S. D. Morgera, "Incorporating retransmission in quality-of-service guaranteed multiuser scheduling over wireless links," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4388–4397, 2009.

[181] J. Zhang, W. Liang, J. Wu, and D. Shi, "A novel retransmission scheme for video services in hybrid wireline/wireless networks," in *2010 IEEE 71st Veh. Technol. Conf.*, 2010, pp. 1–5.

[182] C. Taneja, S. Chatterjee, and S. De, "QoE-aware cross-layer adaptation for delay-constrained video transmission over wireless channels," in *2019 IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2019, pp. 1–6.

[183] S. Soundararajan, P. Agrawal, and Y. Li, "An efficient HARQ retransmission algorithm in OFDMA based wireless networks," in *2009 41st Southeastern Symp. Syst. Theory*, 2009, pp. 88–93.

[184] M. Varga, M. A. Badiu, and V. Bota, "Combined hybrid ARQ and link adaptation for coded cooperation in block-fading channels," *Advances Electron. Telecommun.*, vol. 2, no. 4, pp. 47–53, 2011.

[185] Z. Mheich, M. Le Treust, F. Alberge, and P. Duhamel, "Rate adaptation for incremental redundancy secure HARQ," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 765–777, 2016.

[186] D. J. Dechene and A. Shami, "Energy-aware resource allocation strategies for LTE uplink with synchronous HARQ constraints," *IEEE Trans. Mobile Comput.*, vol. 13, no. 2, pp. 422–433, 2014.

[187] T. Villa, R. Merz, and R. Knopp, "Dynamic resource allocation in heterogeneous networks," in *2013 IEEE Global Commun. Conf. (GLOBECOM)*, 2013, pp. 1915–1920.

[188] H. Mukhtar, A. Al-Dweik, and M. Al-Mualla, "CRC-free hybrid ARQ system using turbo product codes," *IEEE Trans. Commun.*, vol. 62, no. 12, pp. 4220–4229, 2014.

[189] A. G. D. Uchôa, R. D. Souza, and M. E. Pellenz, "Type-I HARQ scheme using LDPC codes and partial retransmissions for AWGN and quasi static fading channels," in *2010 7th Intl. Symp. Wireless Commun. Syst.*, 2010, pp. 571–575.

[190] H. A. Sakr and M. A. Mohamed, "Performance evaluation using smart: HARQ versus HARQ mechanisms beyond 5G networks," *Wireless Pers. Commun.*, vol. 109, no. 3, pp. 1503–1528, 2019.

[191] J. Liu, J. Bergman, R. Fan, R. Hu, M. Ericson, S. Craig, and N. Brannstrom, "Coverage improvements for enhanced uplink," in *VTC Spring 2009 - IEEE 69th Veh. Technol. Conf.*, 2009, pp. 1–5.

[192] S. R. Khosravirad and H. Viswanathan, "Backwards composite feedback for configurable ultra-reliability of retransmission protocols," in *2018 IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2018, pp. 1–6.

[193] Z. Mheich, W. Yu, P. Xiao, A. U. Quddus, and A. Maaref, "On the performance of HARQ protocols with blanking in NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7423–7438, 2020.

[194] Q. Li and C.-X. Chen, "A hybrid ARQ protocol for the communication system with multiple channels," in *2018 Intl. Conf. Info. Commun. Technol. Convergence (ICTC)*, 2018, pp. 222–227.

[195] A. Steiner and S. Shamai, "Multi-layer broadcast hybrid-ARQ strategies," in *2008 IEEE Intl. Zurich Seminar Commun.*, 2008, pp. 148–151.

[196] ——, "Multi-layer broadcasting hybrid-ARQ strategies for block fading channels," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2640–2650, 2008.

[197] A.-N. Assimi, C. Poulliat, and I. Fijalkow, "Packet combining for multi-layer hybrid-ARQ over frequency-selective fading channels," in *2009 17th Eur. Signal Process. Conf.*, 2009, pp. 671–675.

[198] A. Khreis, F. Bassi, P. Ciblat, and P. Duhamel, "Multi-layer HARQ with delayed feedback," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 6224–6237, 2020.

[199] A. Khreis, P. Ciblat, F. Bassi, and P. Duhamel, "Multi-packet HARQ with delayed feedback," in *2018 IEEE 29th Annu. Intl. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, 2018, pp. 1–5.

[200] A. Khreis, F. Bassi, P. Ciblat, and P. Duhamel, "Analysis of multi-messages retransmission schemes," *arXiv preprint arXiv:2004.01090*, 2020.

[201] A. El Hamss, L. Szczecinski, and P. Piantanida, "Increasing the throughput of HARQ via multi-packet transmission," in *2014 IEEE Global Commun. Conf.*, 2014, pp. 1485–1491.

[202] M. Jabi, A. El Hamss, L. Szczecinski, and P. Piantanida, "Multipacket hybrid ARQ: Closing gap to the ergodic capacity," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 5191–5205, 2015.

[203] H. Liang, A. Liu, C. Gong, F. Cheng, and X. Liang, "A systematic multi-packet HARQ scheme using polar codes for IoT," in *2019 IEEE 19th Intl. Conf. Commun. Technol. (ICCT)*, 2019, pp. 307–311.

[204] Song Xiao, Chengke Wu, Jianchao Du, and Yadong Yang, "Reliable transmission of H.264 video over wireless network," in *20th Intl. Conf. Advanced Info. Netw. Appl. - Volume 1 (AINA'06)*, vol. 2, 2006, pp. 5 pp.–.

[205] V. Vadori, A. V. Guglielmi, and L. Badia, "Markov analysis of video transmission based on differential encoded HARQ," in *2016 IEEE 17th Intl. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, 2016, pp. 1–9.

[206] H. Bian, R. Liu, X. Shi, and J. Thompson, "A high-throughput fine-grained rate adaptive transmission scheme for a LEO satellite communication system," in *2018 NASA/ESA Conf. Adaptive Hardware Syst. (AHS)*, 2018, pp. 192–197.

[207] X.-f. Qin, D. Sui, and X. Zhang, "Power-distortion optimization and delay constraint scheme for layered video transmission," in *2008 IEEE 19th Intl. Symp. Pers., Indoor Mobile Radio Commun.*, 2008, pp. 1–5.

[208] 3GPP, "User Equipment (UE) radio transmission and reception (FDD)," 3rd Gener. Partnership Project (3GPP), Tech. Specification (TS) 25.101, 08 1999. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/25101.htm

[209] ——, "User Equipment (UE) radio transmission and reception (TDD)," 3rd Gener. Partnership Project (3GPP), Tech. Specification (TS) 25.102, 1999. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/25102.htm

[210] 3GPP2, "Interoperability Specification (IOS) for CDMA2000 Access Network Interfaces — Part 1 Overview," 3rd Gener. Partnership Project2 (3GPP2), Tech. Specification (TS) 3G-IOS v5.1.1, 2009. [Online]. Available: http://www.3gpp2.org/Public_html/Specs/A.S0011-D_v2.0_090825.pdf

[211] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification," 3rd Gener. Partnership Project (3GPP), Tech. Specification (TS) 36.322, 09 2019, version 15.3.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2438

[212] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification," 3rd Gener. Partnership Project (3GPP), Tech. Specification (TS) 36.321, 04 2020, version 16.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2437

[213] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," 3rd Gener. Partnership Project (3GPP), Tech. Specification (TS) 36.213, 04 2020, version 16.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2427

[214] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification," 3rd Gener. Partnership Project (3GPP), Tech. Specification (TS) 36.331, 04 2020, version 16.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2440

[215] Bluetooth SIG, "Core Specification," Bluetooth Special Interest Group, Tech. Specification (TS) 5.2, 2019. [Online]. Available: https://www.bluetooth.org/docman/handlers/downloaddoc.ashx?doc_id=478726

[216] https://www.bluetooth.com, accessed: 2020-12-24.

[217] "IEEE Standard for Low-Rate Wireless Networks," *IEEE Std 802.15.4-2020 (Revision of IEEE Std 802.15.4-2015)*, pp. 1–800, 2020.

[218] "IEEE Standard for High Data Rate Wireless Multi-Media Networks," *IEEE Std 802.15.3-2016 (Revision of IEEE Std 802.15.3-2003*, pp. 1–510, 2016.

[219] R. Kotaba, C. N. Manchon, N. M. K. Pratas, T. Balercia, and P. Popovski, "Improving spectral efficiency in URLLC via NOMA-based retransmissions," in *ICC 2019 - 2019 IEEE Intl. Conf. Commun. (ICC)*, 2019, pp. 1–7.

[220] M. Deghel, S. E. Elayoubi, A. Galindo-Serrano, and R. Visoz, "Joint optimization of link adaptation and HARQ retransmissions for URLLC services," *2018 25th Intl. Conf. Telecommun., ICT 2018*, no. i, pp. 21–26, 2018.

[221] J. Yeo, S. Park, J. Oh, Y. Kim, and J. Lee, "Partial retransmission scheme for HARQ enhancement in 5G wireless communications," in *2017 IEEE Globecom Workshops (GC Wkshps)*, 2017, pp. 1–5.

[222] Y. Imamura, D. Muramatsu, Y. Kishiyama, and K. Higuchi, "Low latency hybrid ARQ method using channel state information before channel decoding," in *2017 23rd Asia-Pacific Conf. Commun. (APCC)*, 2017, pp. 1–6.

[223] K. Taniyama, Y. Kishiyama, and K. Higuchi, "Low latency HARQ method using early retransmission prior to channel decoding with multistage decision," *IEEE Veh. Technol. Conf.*, vol. 2019-Septe, no. 1, pp. 5–6, 2019.

[224] K. Miura, Y. Kishiyama, and K. Higuchi, "Low latency HARQ method using early retransmission before channel decoding based on superposition coding," in *2019 Intl. Symp. Intelligent Signal Processing Commun. Systems (ISPACS)*, 2019, pp. 1–2.

[225] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2475–2485, 2018.

[226] R. Abreu, G. Berardinelli, T. Jacobsen, K. Pedersen, and P. Mogensen, "A blind retransmission scheme for ultra-reliable and low latency communications," *IEEE Veh. Technol. Conf.*, vol. 2018-June, pp. 1–5, 2018.

[227] R. Abreu, P. Mogensen, and K. I. Pedersen, "Pre-scheduled resources for retransmissions in ultra-reliable and low latency communications," in *2017 IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2017, pp. 1–5.

[228] H. Jang, J. Kim, W. Yoo, and J. Chung, "URLLC mode optimal resource allocation to support HARQ in 5G wireless networks," *IEEE Access*, vol. 8, pp. 126 797–126 804, 2020.

[229] J. P. Battistella Nadas, O. Onireti, R. D. Souza, H. Alves, G. Brante, and M. A. Imran, "Performance analysis of hybrid ARQ for ultra-reliable low latency communications," *IEEE Sensors J.*, vol. 19, no. 9, pp. 3521–3531, 2019.

[230] X. Zhang, J. Wang, and H. V. Poor, "Statistical delay/error-rate bounded QoS provisioning over mmWave cell-free M-MIMO and FBC-HARQ based 5G+ mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1661–1677, 2020.

[231] R. Kotaba, C. Navarro Manchón, T. Balercia, and P. Popovski, "Uplink transmissions in URLLC systems with shared diversity resources," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 590–593, 2018.

[232] E. Kim, Y. Lee, and H. Lee, "An applicable repeated transmission for low latency and reliable services," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8468–8482, 2020.

[233] C. A. Astudillo, F. H. S. Pereira, and N. L. S. da Fonseca, "Probabilistic retransmissions for the random access procedure in cellular IoT networks," in *ICC 2019 - 2019 IEEE Intl. Conf. Commun. (ICC)*, 2019, pp. 1–7.

[234] Z. Xiang, W. Yang, G. Pan, Y. Cai, Y. Song, and Y. Zou, "Secure transmission in HARQ-assisted non-orthogonal multiple access networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2171–2182, 2019.

[235] L. Li, H. Long, L. Zhao, H. Yang, and K. Zheng, "Retransmission scheme for contention-based data transmission systems," *IET Commun.*, vol. 12, no. 2, pp. 144–151, 2018.

[236] C. V. Anamuro, N. Varsier, J. Schwoerer, and X. Lagrange, "Modeling of MTC energy consumption for D2D communications with chase combining HARQ scheme," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6.

[237] B. Yu, Y. Cai, D. Wu, and Z. Xiang, "Average age of information in short packet based machine type communication," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10 306–10 319, 2020.

[238] A. Nessa and M. Kadoch, "Joint network channel fountain schemes for machine-type communications over LTE-advanced," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 418–427, 2016.

[239] S. Park and S. Choi, "Performance of symbol-level combining and bit-level combining in MIMO multiple ARQ systems," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1517–1528, 2016.

[240] S. H. Kim, T. V. K. Chaitanya, and T. Le-Ngoc, "Hybrid ARQ in multicell MU-SIMO with large-scale antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5861–5874, 2016.

[241] J. Jiao, Y. Hu, Q. Zhang, and S. Wu, "Performance modeling of LTP-HARQ schemes over OSTBC-MIMO channels for hybrid satellite terrestrial networks," *IEEE Access*, vol. 6, pp. 5256–5268, 2018.

[242] M. G. Khoshkholgh and V. C. Leung, "Delay analysis of spatially coded MIMO-ZFBF with retransmissions in random networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2652–2667, 2019.

[243] C. Jeon, I. Hwang, and J. W. Lee, "MSSTC-based MIMO-ARQ system with two transmit antennas," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2128–2143, 2017.

[244] T. V. K. Chaitanya and T. Le-Ngoc, "Energy-efficient adaptive power allocation for incremental MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2820–2827, 2016.

[245] S. Park, "Aggregation-assisted combining for MIMO multiple ARQ systems," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 348–351, 2018.

[246] R. Mai, T. V. K. Chaitanya, and T. Le-Ngoc, "Progressive hybrid precoding and combining for massive MIMO ARQ systems," *IEEE Access*, vol. 6, pp. 34 503–34 515, 2018.

[247] J. F. Grybosi, J. L. Rebelatto, G. L. Moritz, and Y. Li, "Age-energy tradeoff of truncated ARQ retransmission with receiver diversity," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1961–1964, 2020.

[248] A. Kosta, N. Pappas, and V. Angelakis, "Age of information: A new concept, metric, and tool," *Found. Trends® Netw.*, vol. 12, no. 3, pp. 162–259, 2017. [Online]. Available: http://dx.doi.org/10.1561/1300000060

[249] M. Kashif, M. Iqbal, Z. Ullah, T. Dagiuklas, S. Sarwar, Z. Ul-Qayyum, and M. Safyan, "Shared hybrid ARQ with incremental redundancy (SHARQ IR) in overloaded MIMO systems to support energy-efficient transmissions," *IEEE Access*, vol. 8, pp. 111 653–111 659, 2020.

[250] F. Rostami Ghadi and M. R. Javan, "Outage and delay performance of content caching in two-tier cooperative cellular networks," *IET Commun.*, vol. 13, no. 16, pp. 2492–2499, 2019.

[251] J. Ma, L. Liu, B. Shang, and P. Fan, "Cache-aided cooperative device-to-device (D2D) networks: A stochastic geometry view," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7444–7455, 2019.

[252] A. Bhatia, R. Nehete, P. Kaushik, P. Murkya, and K. Haribabu, "A lightweight ARQ scheme to minimize communication overhead," in *2019 11th Intl. Conf. Commun. Syst. Netw. (COMSNETS)*, 2019, pp. 507–509.

[253] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Wireless data acquisition for edge learning: Data-importance aware retransmission," *arXiv preprint arXiv:1812.02030*, 2018.

[254] D. Liu, G. Zhu, Q. Zeng, J. Zhang, and K. Huang, "Wireless data acquisition for edge learning: Data-importance aware retransmission," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2020.

[255] M. Kalil, A. Al-Dweik, M. F. Abu Sharkh, A. Shami, and A. Refaey, "A framework for joint wireless network virtualization and cloud radio access networks for next generation wireless networks," *IEEE Access*, vol. 5, pp. 20 814–20 827, 2017.

[256] S. Khalili and O. Simeone, "Uplink HARQ for cloud RAN via separation of control and data planes," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4005–4016, 2017.

[257] E. T. Ceran, D. Gündüz, and A. György, "Average age of information with hybrid ARQ under a resource constraint," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1900–1913, 2019.

[258] X. Wei, Y. Liu, S. Gao, X. Wang, and H. Yue, "An RNN-based delay-guaranteed monitoring framework in underwater wireless sensor networks," *IEEE Access*, vol. 7, pp. 25 959–25 971, 2019.

[259] N. Strodthoff, B. Göktepe, T. Schierl, C. Hellge, and W. Samek, "Enhanced machine learning techniques for early HARQ feedback prediction in 5G," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2573–2587, 2019.

[260] D. Cai, Z. Ding, P. Fan, and Z. Yang, "On the performance of NOMA with hybrid ARQ," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 10 033–10 038, 2018.

[261] D. Cai, Y. Xu, F. Fang, Z. Ding, and P. Fan, "On the impact of time-correlated fading for downlink NOMA," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4491–4504, 2019.

[262] Z. Shi, S. Ma, H. ElSawy, G. Yang, and M. Alouini, "Cooperative HARQ-assisted NOMA scheme in large-scale D2D networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4286–4302, 2018.

[263] J. Choi, "On HARQ-IR for downlink NOMA systems," *IEEE Trans. Commun.*, vol. 64, no. 8, pp. 3576–3584, 2016.

[264] R. Chandran and S. R. Pal, "A novel retransmission scheme for HARQ enhancement in NOMA based LTE systems," in *2019 IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2019, pp. 1–5.

[265] B. Makki, K. Chitti, A. Behravan, and M. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 179–189, 2020.

[266] M. Song, Y. Kim, E. Park, and G. Im, "Rate adaptation and power allocation for cognitive radio networks with HARQ-based primary system," *IEEE Trans. Commun.*, vol. 62, no. 4, pp. 1178–1187, 2014.

[267] R. Joda and M. Zorzi, "Access policy design for cognitive secondary users under a primary type-I HARQ process," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4037–4049, 2015.

[268] E. Park, M. Song, W. Choi, and G. Im, "Maximization of long-term average throughput for cooperative secondary system with HARQ-based primary system in cognitive radio network," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 356–359, 2016.

[269] V. Towhidlou and M. Shikh-Bahaei, "Improved cognitive networking through full duplex cooperative ARQ and HARQ," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 218–221, 2018.

[270] Z. Shi, H. Ding, S. Ma, K. Tam, and S. Pan, "Inverse moment matching based analysis of cooperative HARQ-IR over time-correlated nakagami fading channels," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3812–3828, 2017.

[271] H. Chiu and S. Wu, "Cross-layer performance analysis of cooperative ARQ with opportunistic multi-point relaying in mobile networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4191–4205, 2018.

[272] S. S. Hosseini, J. Abouei, B. Champagne, and X. Chang, "A novel cooperative HARQ protocol for free-space optical broadcasting systems," *J. Lightw. Technol.*, vol. 38, no. 7, pp. 1789–1799, 2020.

[273] C. Tseng and S. Wu, "Effective protocols and channel quality control mechanisms for cooperative ARQ with opportunistic AF relaying," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2382–2397, 2018.

[274] A. Mohamad, R. Visoz, and A. O. Berthet, "Cooperative incremental redundancy hybrid automatic repeat request strategies for multi-source multi-relay wireless networks," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1808–1811, 2016.

[275] M. A. Hassanien, P. Loskot, S. M. Al-Shehri, T. Numanoglu, and M. Mert, "Bitwise retransmission schemes for resources constrained uplink sensor networks," *arXiv preprint arXiv:1707.09696*, 2017.

[276] M. A. Hassanien, P. Loskot, S. M. Al-Shehri, T. Numanoğlu, and M. Mert, "Design and performance evaluation of bitwise retransmission schemes in wireless sensor networks," *Physical Commun.*, vol. 37, 2019.

[277] M. S. H. Abad, O. Ercetin, T. Elbatt, and M. Nahe, "Wireless energy and information transfer in networks with hybrid ARQ," in *2018 IEEE Wireless Commun. Netw. Conf., (WCNC)*, 2018, pp. 1–6.

[278] B. Mouhouche and M. Al-Imari, "Optimization of delivery time in broadcast with acknowledgement and partial retransmission," in *2016 IEEE Intl. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2016, pp. 1–5.

[279] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel state information prediction for 5g wireless communications: A deep learning approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 227–236, 2020.

[280] K. Hammad, M. Mirahmadi, S. Primak, and A. Shami, "On a throughput-efficient look-forward channel-aware scheduling," in *2015 IEEE Int. Conf. Commun/ (ICC)*, 2015, pp. 6234–6239.

[281] M. Tariq, A. Al-Dweik, B. Mohammad, H. Saleh, and T. Stouraitis, "Computational power evaluation for energy-constrained wireless communications systems," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 308–319, 2020.

[282] A. Al-Dweik and Y. Iraqi, "ARQ: The gateway from NOMA to NOM," Oct. 2020. [Online]. Available: https://www.techrxiv.org/articles/preprint/ARQ_The_Gateway_from_NOMA_to_NOM/13157714.