Hybrid ARQ Schemes

Amitav Mukherjee

Charter Communications, Englewood, Colorado, USA

Fundamentals of Hybrid ARQ

Communications over wireless channels are subject to errors due to noise, multipath fading, and interference. The errors result in an inability to correctly decode data packets despite the application of error-correction coding. Therefore, data transmissions often need to be retransmitted to compensate for such errors. Any retransmission scheme has two major components:

- An indication from the receiver to the transmitter that a retransmission is required.
- A corresponding retransmission from the transmitter.

A naïve design would involve the transmitter always retransmitting the original data packet whenever the receiver requests a retransmission. However, such a scheme is not always the most efficient method. For example, the receiver may have correctly decoded a subset of bits from the original transmission, and only the remaining bits need to be retransmitted. This principle is the foundation of the hybrid automatic repeat request (HARQ) protocol in cellular networks such as third-generation (3G) Universal Mobile Telecommunications Service (UMTS), fourth-generation (4G) Long-Term Evolution (LTE), and fifth-generation (5G) New Radio (NR). The term hybrid refers to the fact that the selection of retransmitted bits can be varied based on the underlying error correction coding scheme. The automatic repeat request (ARQ) term refers to a conventional retransmission scheme without hybrid capabilities that operates, for example, at higher layers of the protocol stack of LTE and NR.

Theoretically, the following retransmission schemes are supported within a HARQ framework:

1) Retransmission(s) of the original data packet in its entirety. The receiver performs soft combining (i.e. demodulation and decoding using log-likelihood ratios (LLRs)) of the original packet and the subsequently received copies in an operation known as Chase combining. It is evident that this operation is equivalent to time-domain repetitions until the target signal-to-noise ratio necessary for packet decoding is reached.
2) Selective retransmission of the systematic (i.e. input) bits and parity bits of the original data packet. For example, all systematic and parity bits are included in the first

Wiley 5G Ref. Edited by Rahim Tafazolli, Chin-Liang Wang and Periklis Chatzimisios.
Copyright © 2020 John Wiley & Sons Ltd.
DOI: 10.1002/9781119471509.w5GRef015

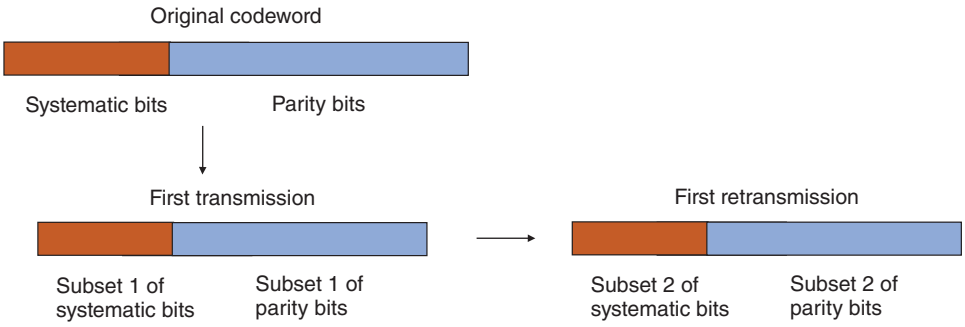

Figure 1 An example of a retransmission scheme that supports incremental redundancy.

transmission, while each subsequent retransmission contains different subsets of systematic and parity bits. This is known as incremental redundancy (IR) that leverages error correction in conjunction with time-domain repetitions.

Figure 1 depicts an example of a basic retransmission scheme. The original codeword is generated by passing the input or systematic bits through a low-rate channel encoder, resulting in a set of systematic bits followed by parity bits. The first transmission takes a specific subset of the systematic and parity bits and transmits them over the wireless channel. The choice of the specific subset is usually based on a carefully designed puncturing scheme. In the event of a retransmission triggered by feedback from the receiver, a second subset of systematic and parity bits is sent to the receiver. If the subsets are the same in the initial transmission and the retransmission, then the receiver can apply Chase combining in order to decode the systematic bits. If the subsets are different and correspond to different redundancy versions (RVs), then the receiver uses the knowledge of which RVs are received in order to decode the packet based on IR. Each RV need not be independently decodable on its own. A maximum number of retransmissions is defined in practical systems and a packet is dropped if this maximum number is exceeded.

HARQ protocols also facilitate high-speed transmissions by enabling pipelining of transmissions/retransmissions of different packets. This is achieved by designating a number of HARQ processes that are active at the same time. Each HARQ process has an identifying number and corresponds to a particular data packet. The receiver stores all the transmitted and retransmitted bits for a given HARQ process until reception is successful, at which point the receive buffer is flushed and that HARQ process is assigned to a new data packet. Therefore, the base station or device can continue to transmit other HARQ processes while waiting for HARQ feedback for the remaining HARQ processes. This is opposed to a classic single-process stop-and-wait protocol where the transmitter does not transmit any new data until a pending transmission is successfully completed. Thus, HARQ with N processes is equivalent to an N-process stop-and-wait protocol.

The receiver alerts the transmitter regarding the successful or failed reception of a HARQ process by feedback of a HARQ acknowledgment (ACK) or HARQ negative acknowledgment (NACK). The status of one HARQ process comprising one codeword can be indicated with one bit of HARQ feedback. For example, the receiver can send a HARQ ACK corresponding to value "1" of one-bit feedback if a particular HARQ process is decoded successfully. Similarly, the receiver can send a HARQ NACK

corresponding to value "0" of one-bit feedback if a particular HARQ process is not decoded correctly. A checksum or cyclic redundancy check (CRC) of the systematic bits can be appended to a data packet so that the receiver can determine if the packet was decoded correctly – if the decoded bits match the checksum then correct decoding is assumed and a HARQ ACK is generated, else a retransmission can be requested from the transmitter.

The benefits of HARQ come with some associated costs. First, the overhead of HARQ feedback detracts from the resources available for data and control transmissions. Second, a delay is incurred between transmissions and retransmissions due to receiver decoding latency, generation of HARQ feedback, transmission of HARQ feedback, and processing latency of the HARQ feedback by the transmitter. This delay is referred to as the HARQ round trip time (RTT). More specifically, RTT can be broken down into the following components (Goktepe et al. 2018):

- Propagation delay between transmitter and receiver;
- Transmission time interval (TTI) duration for the data packet itself;
- T_{RX} – processing time at the receiver;
- $T_{\mathrm{A/N}}$ – transmission time for ACK/NACK feedback;
- T_{TX} – processing time of the feedback at the transmitter.

The propagation delay is a natural artifact arising from the physical length of the radio propagation path and is the only component that cannot be mitigated. The transmission and processing time of the feedback, $T_{\mathrm{A/N}}$ and T_{TX}, are dependent on the radio access network hardware capabilities. The TTI is a parameter of the air interface, and LTE and NR both support multiple TTI durations. In general, TTI can be shortened by reducing the Orthogonal Frequency Division Multiplexing (OFDM) symbol duration and increasing subcarrier spacing. However, the extent to which the TTI can be reduced is limited by the coherence bandwidth of the wireless channel. The processing time at the receiver is the dominant component of HARQ RTT. The processing time T_{RX} comprises of processing the received signal to LLRs T_{LLR} as well as the time for feedback generation T_{FB} involving full decoding, which is represented as $T_{\mathrm{RX}} = T_{\mathrm{LLR}} + T_{\mathrm{FB}}$. TLLR depends on the hardware implementation, while T_{FB} depends upon the particular error correction coding scheme and its implementation at the receiver (Goktepe et al. 2018).

Therefore, fast HARQ protocols need to be designed in order to meet the multi-gigabit throughput requirements of 5G NR systems.

Overview of HARQ in LTE

It is instructive to review the HARQ mechanisms in LTE before moving on to 5G NR.

The LTE Medium Access Control (MAC) protocol layer provides error correction through HARQ and receives HARQ feedback from the physical layer (Layer 1). The HARQ functionality, therefore, ensures delivery between peer entities at the physical layer. There is one HARQ entity at the MAC entity for each serving cell that maintains a number of parallel HARQ processes. A UE may have multiple serving cells in the case of carrier aggregation (CA). Each HARQ process is associated with a HARQ process identifier. The HARQ entity directs HARQ information and associated transport blocks (TBs) received on the downlink (DL) data channel to the corresponding HARQ

Table 1 Maximum number of DL HARQ processes for TDD.

TDD UL/DL configuration	DL subframes per frame	Maximum number of HARQ processes
0	2	4
1	4	7
2	6	10
3	6	9
4	7	12
5	8	15
6	3	6

processes (3GPP 2019a). A TB represents the MAC protocol data unit that is passed to the physical layer for transmission. The physical layer then performs channel encoding, modulation, and scrambling operations to generate a codeword before transmitting over the air. These operations are reversed at the receiver to reconstruct the original TB. The TTI in LTE is one millisecond per DL or UL subframe. For each TTI where a transmission takes place for the HARQ process, one or two (in case of DL spatial multiplexing) TBs and the associated HARQ information are received from the HARQ entity (3GPP 2019a).

Release 8 LTE defines up to eight HARQ processes each for the DL and uplink (UL) in frequency division duplexing (FDD) mode where DL and UL transmissions occupy separate frequencies. The maximum number of DL FDD HARQ processes was increased to as many as 16 in later releases of LTE. The maximum number of DL HARQ processes varies between 4 and 15 for time division duplexing (TDD) mode in Release 8 and depends upon the DL-to-UL resource ratio, as shown in Table 1. On the UL, the maximum number of TDD HARQ processes varies between 1 and 7. The variable number of TDD HARQ processes is required since an UL transmission opportunity is not guaranteed to be available in every subframe of a TDD system, unlike the case of FDD. Therefore, four bits are required for a TDD HARQ process identifier. The aforementioned number of HARQ processes are in addition to the broadcast HARQ process used for system information transmissions without corresponding HARQ feedback.

The DL HARQ mechanism is asynchronous with adaptive retransmissions. In other words, the time difference between a first transmission and a corresponding retransmission is variable, and the time-frequency resource allocation for the retransmitted HARQ process may be different from the resource allocation of the first transmission. The first DL transmission and any retransmissions are scheduled via downlink control information (DCI) sent on the physical downlink control channel (PDCCH). The UE keeps monitoring the PDCCH in order to detect the DCI.

In Release 8, the UL HARQ mechanism is synchronous with either adaptive or nonadaptive retransmissions. In other words, the time difference between a first transmission and a corresponding retransmission is predetermined, and the time-frequency resource allocation for the retransmitted HARQ process is the same as the resource allocation of the first transmission for the nonadaptive case. Table 2 summarizes the UE behavior for operation of synchronous UL HARQ. The UL nonadaptive HARQ

Table 2 LTE UL synchronous HARQ Operation.

HARQ feedback seen by the UE	PDCCH seen by the UE	UE behavior
ACK or NACK	New transmission	New transmission according to PDCCH
ACK or NACK	Retransmission	Retransmission according to PDCCH (adaptive retransmission)
ACK	None	No (re)transmission, keep data in HARQ buffer and a PDDCH is required to resume retransmissions
NACK	None	Nonadaptive retransmission

Source: 3GPP (2019b). Reproduced with permission of 3GPP.

operation requires a predefined RV sequence 0, 2, 3, 1, 0, 2, 3, 1, and so on for successive transmissions of a packet since this information is not explicitly signaled to the UE (Sesia et al. 2011).

Asynchronous UL HARQ was introduced in later releases of LTE to support new use cases such as operation in unlicensed spectrum. HARQ information for DL or UL data transmissions consists of New Data Indicator (NDI) and TB size in the DCI. Toggling the NDI bit indicates whether a new transmission or retransmission is taking place for a particular HARQ process. For DL transmissions, asynchronous UL HARQ, and for autonomous UL HARQ in unlicensed spectrum where UL transmissions take place without an explicit UL grant, the HARQ information also includes the HARQ process ID.

HARQ Feedback in LTE

DL HARQ feedback, i.e. HARQ ACK and HARQ NACK information for DL data transmissions, is classified as a form of uplink control information (UCI) together with channel state information (CSI) and scheduling requests (SR) in LTE. UCI is sent from UEs to the evolved Node-B (eNB). HARQ UCI can be transmitted on either UL control channels known as Physical Uplink Control Channels (PUCCH), or on the UL data channel (with or without UL data multiplexing) known as the Physical Shared Uplink Channel (PUSCH). The advantage of PUCCH is that UCI from multiple UEs can be multiplexed on the same time-frequency resources with users distinguished by means of orthogonal spreading codes, whereas multiuser multiplexing is not allowed for PUSCH.

The eNB delivers UL HARQ ACK/NACK for UL PUSCH transmissions via a special channel known as the Physical Hybrid Automatic Repeat Request Indicator Channel (PHICH). This is more efficient than sending UL HARQ ACK/NACK to individual UEs via the PDCCH since PHICH supports grouping of HARQ feedback to multiple UEs on the same time-frequency resources. The PHICH occupies up to the first three OFDM symbols of a DL subframe and the time duration of the PHICH is configurable.

HARQ feedback timing in LTE depends upon whether the system is operating in FDD or TDD mode. For FDD, the UE shall upon detection of a PDSCH transmission in subframe $n-4$ intended for the UE and for which a HARQ-ACK must be provided, transmit the HARQ-ACK response in subframe n. This yields a HARQ RTT of roughly 8 ms

Table 3 Downlink association set K: $\{k_0, k_1, \ldots k_{M-1}\}$ for TDD.

UL/DL	Subframe n									
configuration	0	1	2	3	4	5	6	7	8	9
0	—	—	6	—	4	—	—	6	—	4
1	—	—	7, 6	4	—	—	—	7, 6	4	—
2	—	—	8, 7, 4, 6	—	—	—	—	8, 7, 4, 6	—	—
3	—	—	7, 6, 11	6, 5	5, 4	—	—	—	—	—
4	—	—	12, 8, 7, 11	6, 5, 4, 7	—	—	—	—	—	—
5	—	—	13, 12, 9, 8, 7, 5, 4, 11, 6	—	—	—	—	—	—	—
6	—	—	7	7	5	—	—	7	7	—

Source: 3GPP (2018a). Reproduced with permission of 3GPP.

between a transmission and a corresponding retransmission. The RTT spans a transmission in subframe n, its corresponding HARQ feedback in subframe $n+4$, and a retransmission if needed in subframe $n+8$. The RTT accounts for UE and eNB processing times.

For TDD, the UE shall upon detection of a PDSCH transmission within subframe(s) $n-k$, where $k \varepsilon K$ and K are defined in Table 3, intended for the UE and for which HARQ-ACK response must be provided, transmit the HARQ-ACK response in UL subframe n. Certain TDD configurations are biased towards DL resources compared to UL resources. For this reason, HARQ feedback compression schemes such as bundling and multiplexing are supported in TDD.

TDD HARQ-ACK bundling is performed per codeword across M multiple subframes in the DL or special subframes associated with a single UL subframe n, where M is the number of elements in the set K defined in Table 3. The bundling involves a logical AND operation of all the individual PDSCH transmission HARQ-ACKs. Therefore, an M-fold reduction in HARQ feedback overhead is possible with bundling.

For TDD HARQ-ACK multiplexing and a subframe n with $M > 1$, M being the number of elements in the set K defined in Table 3, spatial HARQ-ACK bundling across multiple codewords within a DL or special subframe is performed by a logical AND operation of all the corresponding individual HARQ-ACKs. For TDD HARQ-ACK multiplexing and a subframe n with $M = 1$, spatial HARQ-ACK bundling across multiple codewords within a DL or special subframe is not performed. A separate HARQ feedback indicator is still generated per DL subframe, unlike the bundling scheme.

A disadvantage of these HARQ feedback compression schemes is that the eNB does not know exactly, which TB(s) could not be decoded correctly at the UE. In the event of a HARQ NACK, all TBs in the same group must be resent, increasing retransmission overheads and reducing link throughput. Another by-product is that the average HARQ RTT can increase since some TBs cannot be acknowledged until the remainder of the group has been received.

A further complication arises because the detection of PDCCH control signaling is not infallible and it is probable that the UE will miss some DL resource assignments. This would introduce the possibility of HARQ protocol errors, including the erroneous

transmission of ACK in the case when one or more DL assignments were missed in a group of subframes. In order to mitigate this problem, a Downlink Assignment Index (DAI) field is included in the PDCCH to indicate to the UE the number of subframes in a group that actually contains a DL transmission. In the case of ACK/NACK bundling, this helps the UE to detect missed DL assignments and avoid transmitting ACK if one or more DL assignments were missed, while in the case of ACK/NACK multiplexing the DAI helps the UE to determine how many bits of ACK/NACK information should be returned (3GPP 2019b).

LTE MAC HARQ Aspects

As stated previously, the LTE MAC layer coordinates the HARQ procedure. A related aspect is the energy efficiency of HARQ since it is undesirable for the UE to keep waiting indefinitely for a retransmission. Therefore, the UE energy-saving mechanism known as discontinuous reception (DRX) needs to work seamlessly together with HARQ. When the UE enters DRX, it transitions to a low-activity state with intermittent monitoring of PDCCH. Several timers are defined at the MAC layer for efficient HARQ management. These include the following:

- *HARQ RTT Timer*: This parameter specifies the minimum amount of subframe(s) before a DL assignment for HARQ retransmission is expected by the MAC entity. The timer is started upon detection of a PDCCH indicating a DL transmission in a given subframe.
- *UL HARQ RTT Timer*: This parameter specifies the minimum amount of subframe(s) before an UL assignment for HARQ retransmission is expected by the MAC entity in the case of asynchronous UL HARQ. The timer is started upon detection of a PDCCH indicating an UL transmission for an asynchronous HARQ process or if an UL grant has been configured for an asynchronous HARQ process for a given subframe.
- *drx-RetransmissionTimer*: Specifies the maximum number of consecutive PDCCH-subframe(s) until a DL retransmission is received. This timer is started when a HARQ RTT Timer expires in a subframe and the data of the corresponding HARQ process was not successfully decoded. The timer is stopped if a DL grant is received for the corresponding HARQ process.
- *drx-ULRetransmissionTimer*: Specifies the maximum number of consecutive PDCCH-subframe(s) until a grant for UL retransmission or the HARQ feedback is received. This timer is started when an UL HARQ RTT Timer expires in a subframe. The timer is stopped if the PDCCH indicates an UL transmission for an asynchronous HARQ process or if an UL grant has been configured for an asynchronous HARQ process for this subframe, or if the PDCCH indicates an UL transmission for an autonomous HARQ process.

For each serving cell, the HARQ RTT Timer is set to 8 subframes in case of FDD configuration and unlicensed spectrum operation on the serving cell which carries the HARQ feedback for this cell (Goktepe et al. 2018). For each serving cell, in case of TDD configuration, the HARQ RTT Timer is set to $k + 4$ subframes, where k is the interval between the DL transmission and the transmission of associated HARQ feedback as defined in the section titled "HARQ Feedback in LTE." UL HARQ RTT Timer length is set to 4 subframes for FDD and unlicensed spectrum operation and set to $k_{\text{ULHARQRTT}}$

subframes for TDD, where $k_{\text{ULHARQRTT}}$ ranges between 4 and 7 depending upon the TDD configuration.

Release 15 NR HARQ Design

The NR HARQ design in Release 15 bears some similarities to LTE HARQ but is designed to accommodate the increased flexibility of the NR air interface since it does not have separate designs for FDD and TDD modes of operation. DL and UL transmissions are organized into frames with 10 ms duration, each consisting of 10 subframes of 1 ms duration. The slot duration with a SCS of 15 kHz is 14 symbols with normal CP and 12 symbols with Extended CP, and scales in time as a function of the used subcarrier spacing so that there is always an integer number of slots in a subframe. NR OFDM symbols in a slot can be classified as DL, flexible, or UL. The UE assumes that DL transmissions only occur in DL or flexible symbols, while the UE only transmits in UL or flexible symbols.

Similar to LTE, the NR MAC entity includes a HARQ entity for each serving cell. The HARQ entity maintains a number of parallel HARQ processes. Each HARQ process is associated with a HARQ process identifier. The HARQ entity directs HARQ information and associated TBs received on the DL or UL data channel to the corresponding HARQ processes. Note that initial 5G NR deployments will be in a nonstandalone mode with dual connectivity between a LTE master node and a NR secondary node. However, each node will have independent HARQ operations on the DL and UL.

Up to 16 HARQ processes can be configured on the DL and UL for a UE. The number of HARQ processes is increased in NR compared to LTE for a couple of reasons. First, the TTI in NR can be as small as two OFDM symbols, which allows a greater granularity of TB transmissions. Second, the UE processing time is reduced compared to LTE (depending upon UE capability) due to enhancements such as frequency-first and time-second mapping of codewords to resource elements and hardware advances.

The HARQ process supports one TB when the physical layer is not configured for DL spatial multiplexing. The HARQ process supports one or two TBs when the physical layer is configured for DL spatial multiplexing. On the UL, each HARQ process supports only one TB.

Each HARQ process is associated with a HARQ buffer at the receiver. New transmissions are performed on the resource and with the modulation and coding scheme (MCS) indicated by the network. Retransmissions are performed on the resource and, if provided, with the MCS indicated on PDCCH, or on the same resource and with the same MCS as was used for the last made transmission attempt within a bundle.

Asynchronous IR Hybrid ARQ is supported on the DL. The next-generation Node-B (gNB) provides the UE with the HARQ-ACK feedback timing either dynamically in the DCI or semi-statically in an RRC configuration. Asynchronous IR Hybrid ARQ is also supported on the UL. The gNB schedules each UL transmission and retransmission using the UL grant on DCI. Therefore, synchronous HARQ is not supported in Release 15 NR.

When the MAC entity is configured with PUSCH repetitions on the UL, the parameter *repK* provides the number of transmissions of a TB within a bundle of the configured UL grant. After the initial transmission, HARQ retransmissions follow within a bundle.

For both dynamic grant and configured UL grant, bundling operation relies on the HARQ entity for invoking the same HARQ process for each transmission that is part of the same bundle. Within a bundle, HARQ retransmissions are triggered without waiting for feedback from the previous transmission. Each transmission within a bundle is a separate UL grant after the initial UL grant within a bundle is delivered to the HARQ entity.

Code block group (CBG)-based transmission and retransmission is a new feature compared to LTE. The UE may be configured to receive CBG-based transmissions where retransmissions may be scheduled to carry a sub-set of all the code blocks (CBs) of a TB. On the UL, the UE may be configured to transmit CBG-based transmissions where retransmissions may be scheduled to carry a sub-set of all the CBs of a TB. More details are provided in a section titled "Code Block Group-Based HARQ Feedback."

Overview of HARQ Feedback in NR

HARQ feedback is classified as a form of UCI together with CSI and SR in NR. As in LTE, UCI can be transmitted on either the PUCCH or PUSCH. Five formats of PUCCH exist, depending on the duration of PUCCH and the UCI payload size (3GPP 2019b):

- Format #0: Short PUCCH of 1 or 2 OFDM symbols with small UCI payloads of up to two bits with UE multiplexing capacity of up to 6 UEs with 1-bit payload in the same PRB;
- Format #1: Long PUCCH of 4-14 symbols with small UCI payloads of up to two bits with UE multiplexing capacity of up to 84 UEs without frequency hopping and 36 UEs with frequency hopping in the same PRB;
- Format #2: Short PUCCH of 1 or 2 symbols with large UCI payloads of more than two bits with no UE multiplexing capability in the same PRBs;
- Format #3: Long PUCCH of 4-14 symbols with large UCI payloads with no UE multiplexing capability in the same PRBs;
- Format #4: Long PUCCH of 4-14 symbols with moderate UCI payloads with multiplexing capacity of up to 4 UEs in the same PRBs.

The UE selects a particular PUCCH Format based on the size of the UCI to be transmitted. NR PUCCH is flexible in its time and frequency allocation, unlike LTE PUCCH that is located at the edges of the carrier bandwidth and has a fixed duration and timing.

UCI multiplexing in PUSCH is supported when UCI and PUSCH transmissions coincide in time, either due to transmission of an UL TB or due to triggering of aperiodic CSI transmission without UL TB:

- UCI carrying HARQ-ACK feedback with 1 or 2 bits is multiplexed by puncturing PUSCH;
- In all other cases, UCI is multiplexed by rate matching PUSCH. This avoids a degradation in PUSCH decoding performance due to excessive puncturing.

If a UE receives a PDSCH without receiving a corresponding PDCCH, or if the UE receives a PDCCH indicating a SPS PDSCH release, the UE generates one corresponding HARQ-ACK information bit.

If a UE is not provided PDSCH-CodeBlockGroupTransmission, the UE generates one HARQ-ACK information bit per TB.

For a HARQ-ACK information bit, a UE generates an ACK if the UE detects a DCI format 1_0 that provides a SPS PDSCH release or correctly decodes a TB, and generates a NACK if the UE does not correctly decode the TB.

A UE does not expect to be indicated to transmit HARQ-ACK information for more than one SPS PDSCH receptions in a same PUCCH.

The timing of HARQ-ACK feedback is dynamic in NR and can be controlled by DCI. For example, PDSCH-to-HARQ_feedback timing indicator is a 3-bit field in DCI Format 1_0. For DCI format 1_0, the PDSCH-to-HARQ-timing-indicator field values map to {1, 2, 3, 4, 5, 6, 7, 8}. For DCI format 1_1, if present, the PDSCH-to-HARQ-timing-indicator field values map to values for a set of number of slots provided by the parameter *dl-DataToUL-ACK* as defined in Table 4. Therefore, the network has considerable flexibility in controlling when HARQ-ACK feedback is triggered.

With reference to slots for PUCCH transmissions, if the UE detects a DCI format 1_0 or a DCI format 1_1 scheduling a PDSCH reception ending in slot n or if the UE detects a DCI format 1_0 indicating a SPS PDSCH release through a PDCCH reception ending in slot n, the UE provides corresponding HARQ-ACK information in a PUCCH transmission within slot $n+k$, where k is a number of slots and is indicated by the PDSCH-to-HARQ-timing-indicator field in the DCI format, if present, or provided by *dl-DataToUL-ACK*.

Next, consider how HARQ feedback bits are generated in NR. Two types of HARQ-ACK codebooks are defined in Release 15: semi-static (Type-1) and dynamic (Type-2). These codebooks are applicable to both CA and single-carrier scenarios.

In the case of a Type-1 HARQ-ACK codebook for transmission on PUCCH, the first step is for the UE to determine a set of $M_{A,c}$ occasions for candidate PDSCH receptions on serving cell c for which the UE can transmit corresponding HARQ-ACK information in a PUCCH in slot n_{U}. The set $M_{A,c}$ is determined based on a number of semi-static parameters such as the range of HARQ-ACK slot timing values indicated by the PDSCH scheduling grants, the time-domain resource allocation of the candidate PDSCH receptions, and the availability of UL slots for PUCCH transmission for these candidates. The range

Table 4 Slot offset indication in DCI for NR DL HARQ feedback.

PDSCH-to-HARQ feedback timing indicator			
1 bit	**2 bits**	**3 bits**	**Number of slots *k***
'0'	'00'	'000'	1st value from *dl-DataToUL-ACK*
'1'	'01'	'001'	2nd value from *dl-DataToUL-ACK*
	'10'	'010'	3rd value from *dl-DataToUL-ACK*
	'11'	'011'	4th value from *dl-DataToUL-ACK*
		'100'	5th value from *dl-DataToUL-ACK*
		'101'	6th value from *dl-DataToUL-ACK*
		'110'	7th value from *dl-DataToUL-ACK*
		'111'	8th value from *dl-DataToUL-ACK*

Source: 3GPP (2019d). Adapted from 3GPP.

of HARQ-ACK slot timing values is either predefined (between 1 and 8 slots) or semi-statically configured, depending upon the DCI format used to scheduling PDSCH.

Once $M_{A,c}$ is derived, a UE determines $\tilde{o}_0^{ACK}, \tilde{o}_1^{ACK}, \ldots, \tilde{o}_{O^{ACK}-1}^{ACK}$ HARQ-ACK information bits corresponding to a total number of O_{ACK} HARQ-ACK information bits for all of its configured serving cells as follows. The cardinality of the set $M_{A,c}$ defines a total number M_c of occasions for PDSCH reception or SPS PDSCH release for serving cell c corresponding to the HARQ-ACK information bits. For each of these M_c occasions, the number of HARQ-ACK bits are computed depending upon whether one or two DL TBs (i.e. multiple-input multiple-output (MIMO)) or CBG-based feedback (see section titled "Code Block Group-Based HARQ Feedback") has been configured. If the UE does not receive a TB or a CBG, due to the UE not detecting a corresponding DCI format 1_0 or DCI format 1_1, the UE generates a NACK value for the TB or the CBG.

Therefore, the semi-static HARQ-ACK codebook has a fixed codebook size since M_c is determined based on semi-static parameters and the number of serving cells is also semi-static. This codebook size may incur a large overhead for the feedback report since not all of the configured carriers may actually have scheduled PDSCH reception for the UE.

The Type-2 dynamic HARQ-ACK codebook is designed to alleviate the above issues with the Type-1 semi-static codebook. The dynamic codebook only reports HARQ-ACK information for the DL carriers with actual PDSCH scheduling grants, which implies a dynamically varying codebook size. Therefore, in principle, the feedback overhead can be reduced compared to a semi-static codebook. However, this dynamicity comes at a cost. The Type-2 codebook assumes the gNB and the UE have the same understanding regarding which scheduled PDSCH(s) need to have HARQ-ACK feedback reported, but this cannot always be guaranteed since the UE may fail to detect PDCCHs that schedule one or more of the PDSCHs transmitted on the DL. If PDCCH detection is unsuccessful then the UE will be unaware that a PDSCH was even transmitted, while the gNB expects HARQ-ACK feedback for the same.

The solution in the case of Type-2 codebooks is to define a couple of counters, the counter downlink assignment indicator (DAI), and the total DAI. The principle is similar to DAI usage in TDD LTE.

The first step of HARQ-ACK generation is for the UE to determine a set of PDCCH monitoring occasions for PDSCH receptions. The cardinality of the set of PDCCH monitoring occasions defines a total number M of PDCCH monitoring occasions.

Then, a value of the counter DAI field in DCI denotes the accumulative number of {serving cell, PDCCH monitoring occasion}-pair(s) in which PDSCH reception(s) associated with the DCI format is present, up to the current serving cell and current PDCCH monitoring occasion, first in ascending order of serving cell index and then in ascending order of PDCCH monitoring occasion index m, where $0 \leq m \leq M$. In other words, the counter DAI conveys the number of scheduled DL transmissions up to the point the DCI was received.

The value of the total DAI (when present) in DCI denotes the total number of {serving cell, PDCCH monitoring occasion}-pair(s) in which PDSCH reception(s) associated with the DCI format is present, up to the current PDCCH monitoring occasion m and is updated from PDCCH monitoring occasion to PDCCH monitoring occasion.

DAI in the DCI can be 0, 2, or 4 bits. The combination of counter DAI and total DAI helps the UE keep track of how many PDSCH(s) it is expected to provide HARQ feedback for.

Code Block Group-Based HARQ Feedback

In NR, DL and UL TBs are input to a CB segmentation procedure after a TB-level CRC attachment. If the length of the TB with CRC is larger than a specific maximum CB size, then segmentation of the input bit sequence is performed and an additional CB-level CRC sequence of 24 bits is attached to each CB. The maximum number of CBGs for a PDSCH and PUSCH in Release 15 is 8.

If a UE is configured with CBG reception for a serving cell, the UE receives a PDSCH scheduled by DCI format 1_1, which includes CBGs of a TB. The UE is also provided with a maximum number of CBGs for generating respective HARQ-ACK information bits for a TB reception for the serving cell (3GPP 2019c).

If a UE is configured to receive CBG based transmissions by receiving the higher layer parameter CBG transmission for PDSCH, the code block group transmission information (CBGTI) field of DCI format 1_1 is of length $N_{\mathrm{TB}} \times N$ bits, where N_{TB} is the value of the maximum number of codewords scheduled by DCI. The CBGTI can be of length 0, 2, 4, 6, or 8 bits. If $N_{\mathrm{TB}} = 2$ the CBGTI field bits are mapped such that the first set of N bits starting from the most significant bit (MSB) corresponds to the first TB, while the second set of N bits corresponds to a second TB if scheduled. The first M bits of each set of N bits in the CBGTI field have an in-order one-to-one mapping with the M CBGs of the TB, with the MSB mapped to CBG#0.

For the initial transmission of a TB as indicated by the NDI field of the scheduling DCI, the UE may assume that all the CBGs of the TB are present.

For a retransmission of a TB as indicated by the NDI field of the scheduling DCI, the UE may assume that

The CBGTI field of the scheduling DCI indicates which CBGs of the TB are present in the transmission. A bit value of '0' in the CBGTI field indicates that the corresponding CBG is not transmitted and '1' indicates that it is transmitted.

If the code block group flushing out information (CBGFI) field of the scheduling DCI is present, CBGFI set to '0' indicates that the earlier received instances of the same CBGs being transmitted may be corrupted, and CBGFI set to '1' indicates that the CBGs being retransmitted are combinable with the earlier received instances of the same CBGs.

A CBG contains the same CBs as in the initial transmission of the TB.

For C number of CBs in a TB, the UE determines a number of CBGs M and determines a number of HARQ-ACK bits for the TB as $N_{\mathrm{HARQ-ACK}}^{\mathrm{CBG/TB}} = M$.

The UE generates an ACK for the HARQ-ACK information bit of a CBG if the UE correctly received all CBs of the CBG and generates a NACK for the HARQ-ACK information bit of a CBG if the UE incorrectly received at least one CB of the CBG. Therefore, up to 8 bits of HARQ feedback can be generated per TB when CBG-level feedback is enabled. If the UE receives two TBs, the UE concatenates the HARQ-ACK information bits for CBGs of the second TB after the HARQ-ACK information bits for CBGs of the first TB (3GPP 2019c).

The HARQ-ACK codebook includes the $N_{\mathrm{HARQ-ACK}}^{\mathrm{CBG/TB,max}}$ HARQ-ACK information bits and, if $N_{\mathrm{HARQ-ACK}}^{\mathrm{CBG/TB}} < N_{\mathrm{HARQ-ACK}}^{\mathrm{CBG/TB,max}}$ for a TB, the UE generates a NACK value for the last

$N_{\text{HARQ-ACK}}^{\text{CBG/TB,max}} - N_{\text{HARQ-ACK}}^{\text{CBG/TB}}$ HARQ-ACK information bits for the TB in the HARQ-ACK codebook.

If the UE generates a HARQ-ACK codebook in response to a retransmission of a TB, corresponding to a same HARQ process as a previous transmission of the TB, the UE generates an ACK for each CBG that the UE correctly decoded in a previous transmission of the TB. If a UE correctly detects each of the CBGs and does not correctly detect the TB for the CBGs, the UE generates a NACK value for each of the $N_{\text{HARQ-ACK}}^{\text{CBG/TB}}$CBGs (3GPP 2019c).

MAC Aspects of NR HARQ

Since NR UEs can also be configured with DRX, the interplay between HARQ and DRX needs to be handled in NR as well. Several timers are defined at the MAC layer for this purpose, such as (3GPP 2019e):

- *drx-RetransmissionTimerDL* (per DL HARQ process except for the broadcast process): the maximum duration until a DL retransmission is received;
- *drx-RetransmissionTimerUL* (per UL HARQ process): the maximum duration until a grant for UL retransmission is received;
- *drx-HARQ-RTT-TimerDL* (per DL HARQ process except for the broadcast process): the minimum duration before a DL assignment for HARQ retransmission is expected by the MAC entity;
- *drx-HARQ-RTT-TimerUL* (per UL HARQ process): the minimum duration before an UL HARQ retransmission grant is expected by the MAC entity.

Upon detection of a DL grant, the *drx-HARQ-RTT-TimerDL* is started for the corresponding HARQ process in the first symbol after the end of the corresponding transmission carrying the DL HARQ feedback and the *drx-RetransmissionTimerDL* is stopped for the corresponding HARQ process. If a *drx-HARQ-RTT-TimerDL* expires and the data of the corresponding HARQ process was not successfully decoded, the drx-RetransmissionTimerDL is started for the corresponding HARQ process in the first symbol after the expiry of drx-HARQ-RTT-TimerDL.

Upon detection of an UL grant, the *drx-HARQ-RTT-TimerUL* is started for the corresponding HARQ process in the first symbol after the end of the first repetition of the corresponding PUSCH transmission and the *drx-RetransmissionTimerUL* is stopped for the corresponding HARQ process. If a *drx-HARQ-RTT-TimerUL* expires, the drx-RetransmissionTimerUL for the corresponding HARQ process is started in the first symbol after the expiry of *drx-HARQ-RTT-TimerUL*.

Release 16 NR HARQ Enhancements

HARQ mechanisms in NR are continually evolving in order to support the wide diversity of use cases in Release 16 and beyond, as opposed to a one-size-fits-all approach. This section provides an overview of several NR HARQ enhancements for different feature items within Release 16, such as operation in unlicensed spectrum, enhanced ultra-reliable low-latency communications (URLLC), and nonterrestrial networks (NTNs) such as satellite-based 5G.

HARQ in NR-Unlicensed

Cellular networks traditionally operate in licensed spectrum where interference conditions are controlled and quality of service guarantees can be made. This paradigm changed when licensed assisted access for LTE was defined in Release 13 in order to exploit the unlicensed spectrum bands available at 5 GHz for opportunistic transmissions (3GPP 2015; Mukherjee et al. 2016).

The Rel-16 Study Item on New Radio-Unlicensed (NR-U) sought to emulate the foray into unlicensed spectrum by licensed-assisted access (LAA) LTE. The main objectives were to develop a single global solution for NR-based access to unlicensed spectrum in the 5–7.125 GHz range that deviates as little as possible from baseline Rel-15 NR, and to study coexistence methods with other unlicensed spectrum technologies in accordance with regulatory requirements (3GPP 2018b).

With regard to HARQ-ACK feedback, NR-U has the additional requirement of potential channel access procedures (e.g. listen-before-talk with exponential backoff) that need to be performed prior to UL transmission. Transmission of HARQ ACK/NACK for the corresponding DL data in the same shared channel occupancy (CO) is, therefore, beneficial in order to exploit CO sharing and avoid UL LBT for the HARQ feedback. A self-contained burst where all HARQ ACK/NACK bits are transmitted for the corresponding data in the same shared CO is ideal for this reason.

However, in some cases, the HARQ ACK/NACK has to be transmitted in a separate CO from the one in which the corresponding data was transmitted; for example, the maximum CO time imposed by regulation does not leave room for PUCCH. The left-over HARQ ACK/NACK bits can be deferred to a later CO by indicating a nonnumerical value in the PDSCH-to-HARQ-timing-indicator in the DCI scheduling the PDSCH. The gNB can in addition request HARQ feedback for PDSCHs from earlier CO(s), where the exact number of HARQ processes that should be reported, the HARQ feedback timing, and feedback resource is provided to the UE in another DCI (in the same or in another CO). This can be considered as a form of HARQ ACK/NACK polling with an enhanced dynamic HARQ feedback codebook (i.e. feedback for only the indicated HARQ processes needs to be transmitted). Examples of both HARQ feedback enhancements are shown in Figure 2.

HARQ in Enhanced URLLC

Release 15 NR includes a number of enhancements to support ultra-reliable low-latency communication (URLLC) use cases such as remote operation of vehicles and augmented

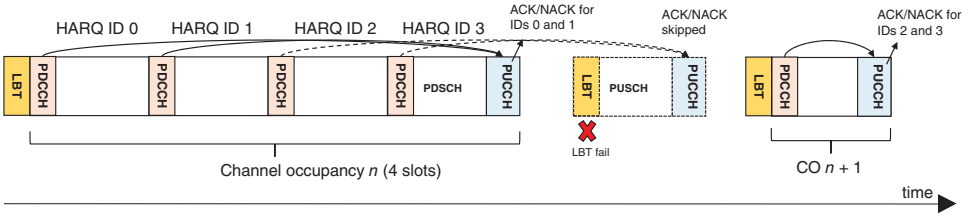

Figure 2 Release 16 NR-Unlicensed HARQ feedback. Indication of delayed HARQ feedback outside channel occupancy n and HARQ feedback polling by gNB in channel occupancy $n + 1$.

Figure 3 Out-of-sequence HARQ feedback for high-priority DL transmissions.

reality. Release 15 (see Low-Latency Transmission) performance requirements are characterized by an over-the-air latency of user plane data of at most 0.5 ms on average for a one-way transmission, and a target reliability of 99.999%. In Release 16, additional use cases such as factory automation with tighter requirements (e.g. target reliability of 99.9999% or a block error rate of 10^{-6}) were identified as important for NR evolution, in addition to the need for enhancing the Release 15 enabled use cases (3GPP 2019f).

One such enhancement in Rel-16 is out-of-order DL HARQ ACK/NACK feedback. Specifically, for a Rel-16 URLLC UE and dynamic DL scheduling, on the active BWP of a given serving cell, the HARQ-ACK associated with the second PDSCH with HARQ process ID x received after the first PDSCH with HARQ process ID y ($x \neq y$) can be sent before the HARQ-ACK of the first PDSCH. This allows prioritization of URLLC HARQ processes over eMBB processes, as shown in Figure 3. The UE always processes the second PDSCH, while the UE may or may not drop the processing of the first channel.

Finally, UCI feedback is enhanced as follows. In NR Rel-15, only one PUCCH for HARQ-ACK transmission is supported within a slot. Enabling more than one PUCCH for HARQ-ACK transmission within a slot is beneficial as it may enable fast HARQ-ACK feedback to reduce the latency and facilitate separate HARQ-ACK feedback for URLLC and eMBB. Therefore, Rel-16 supports transmission of multiple PUCCHs within a slot from a given UE. To enable this, PDSCH(s) are grouped into subslots and each subslot is associated with a group of HARQ ACK/NACK bits. The determination of HARQ-ACK codebook and corresponding PUCCH resource is as in Rel-15, with the difference that subslot boundaries are used instead of slot boundaries. Furthermore, two HARQ-ACK codebooks can be constructed simultaneously for a given UE in order to support both URLLC and eMBB services. The use of separate codebooks allows prioritization of URLLC HARQ ACK/NACK feedback over eMBB HARQ feedback. For example, if there is a collision between eMBB HARQ ACK and URLLC HARQ ACK in the same slot, the UE can drop the eMBB HARQ ACK if certain ACK multiplexing conditions are not met.

HARQ for Nonterrestrial Networks

The flexible air interface of NR makes it a prime candidate for radically new deployment cases such as NTNs. From a 5G perspective, NTN refer to networks, or segments of networks, that use an airborne or spaceborne vehicle (e.g. low earth orbit satellites) for transmission.

The propagation delays in NTN can be orders of magnitude larger than terrestrial deployments. This requires modifications to several NR procedures such as HARQ. It has been shown thus far that the HARQ protocol is a very time-critical mechanism in order to maintain high throughput. NR has 16 HARQ processes and each DL HARQ process waits for an ACK or NACK from the UE to determine whether to transmit a new TB or to retransmit a prior TB. For NTN, the HARQ round-trip time between

initial data transmission and retransmission can be hundreds of milliseconds. Existing MAC timers such as drx-HARQ-RTT-TimerDL (the minimum duration before a DL assignment for HARQ retransmission is expected by the MAC entity) may also need modification. One approach can be to extend the number of HARQ processes to accommodate NTN delays, for example, introducing 50 HARQ processes for satellite-based scenarios. However, the UE complexity would be greatly increased due to the need to handle such a large number of parallel HARQ processes. Therefore, the solution that has been deemed beneficial in Release 16 is to semi-statically disable HARQ ACK/NACK feedback and rely on the ARQ layer for retransmissions (3GPP 2019g).

Conclusions

Hybrid ARQ is a key mechanism that leverages error correction coding and repetitions for error recovery and reliable packet delivery in mobile broadband networks. In particular, the operation of multiple HARQ processes in parallel is critical for maintaining high throughputs at the physical layer and this has been demonstrated in 3G and 4G radio access technologies. Release 15 5G NR introduces a number of HARQ enhancements compared to 4G LTE, such as an increased number of HARQ processes, natively asynchronous DL and UL HARQ with flexible scheduling of DL HARQ-ACK feedback, and retransmissions with CBG granularity as opposed to a coarser TB granularity.

HARQ enhancements continue to be added to NR as it evolves in Release 16 and beyond to encompass new use cases. Examples such as operation in unlicensed spectrum and enhanced URLLC have been addressed previously in the article. The adoption of HARQ-based protocols also is a candidate enhancement for the next generation of IEEE 802.11 WLAN that is currently under development in the form of 802.11be (Lopez-Perez et al. 2019).

Related Article

Low-Latency Transmission

References

3GPP (2015). TR 36.889 V13.0.0, Feasibility study on licensed-assisted access to unlicensed spectrum (Release 13), June 2015.

3GPP (2018a). TS 36.213 V15.2.0 (2018-06), E-UTRA; Physical layer procedures (Release 15).

3GPP (2018b). TR 38.889 V16.0.0, Study on NR-based access to unlicensed spectrum, (Release 16), June 2018.

3GPP (2019a). TS 36.321 V15.3.0 (2019-06), E-UTRA; Medium Access Control (MAC) protocol specification (Release 15).

3GPP (2019b). TS 36.300 V15.7.0 (2019-09), (E-UTRAN); Overall description; Stage 2 (Release 15).

3GPP (2019c). TS 38.212 V15.6.0 (2019-06), NR; Multiplexing and channel coding (Release 15).
3GPP (2019d). TS 38.213 V15.6.0 (2019-06), NR; Physical layer procedures for control (Release 15).
3GPP (2019e). TS 38.321 V15.5.0 (2019-03), NR; Medium Access Control (MAC) protocol specification (Release 15).
3GPP (2019f). TR 38.824 v2.0.0, Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC), Release 16, March 2019.
3GPP (2019g). RP-182880, Study on solutions for NR to support non-terrestrial networks (NTN), 2019.
Goktepe, B., Fahse, S., Thiele, L., et al. (2018). Subcode-based early HARQ for 5G. 2018 IEEE International Conference on Communications Workshops (ICC Workshops), Kansas City, MO, 2018.
Lopez-Perez, D., Garcia-Rodriguez, A., Galati-Giordano, L., et al. (2019). IEEE 802.11be - Extremely High Throughput: The next generation of Wi-Fi technology beyond 802.11ax, 2019. [Online] arXiv:1902.04320v1
Mukherjee, A., Cheng, J.-F., Falahati, S. et al. (2016). Licensed-assisted access LTE: coexistence with IEEE 802.11 and the evolution towards 5G. *IEEE Communications Magazine* 54 (6): 50–57.
Sesia, S., Toufik, I., and Baker, M. (2011). *LTE – The UMTS Long Term Evolution: From Theory to Practice*, 2e. Wiley-Blackwell.

Further Reading

3GPP (2019). TS 38.300 V15.7.0 (2019-09), NR; NR and NG-RAN Overall Description; Stage 2 (Release 15).
3GPP (2019). TS 38.214 V15.5.0 (2019-03), NR; Physical layer procedures for data (Release 15).
Ahmadi, S. (2019). *5G NR: Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards*. Academic Press.
Chandramouli, D., Liebhart, R., and Pirskanen, J. (2019). *5G for the Connected World*. Wiley-Blackwell.
Dahlman, E., Parkvall, S., and Skold, J. (2018). *5G NR: The Next Generation Wireless Access Technology*. Academic Press.
Dahlman, E., Parkvall, S., and Skold, J. (2016). *4G LTE-Advanced Pro and the Road to 5G*. Elsevier.
Mukherjee, A. (2019). *5G New Radio: Beyond Mobile Broadband*. Artech House.
Parkvall, S., Dahlman, E., Furuskar, A., and Frenne, M. (2017). NR: the new 5G radio access technology. *IEEE Communications Standards Magazine* 1 (4): 24–30.