CHAPTER 2

New Radio Access Layer 2/3 Aspects and System Operation

2.1 Overview of Layer 2 and Layer 3 Functions

The NR radio interface protocols (alternatively referred to as layer-1, layer-2, and layer-3 protocols) operate between the NG-RAN and the UE and consist of user-plane protocols, for transfer of user data (IP packets) between the network and the UE, and control-plane protocols, for transporting control signaling information between the NG-RAN and the UE. The non-access stratum (NAS) protocols terminate in the UE and the AMF entity of the 5G core network and are used for core network related functions and signaling including registration, authentication, location update and session management. In other words, the protocols over Uu and NG interfaces are categorized into user plane and control plane protocols. User plane protocols implement the actual PDU Session service which carries user data through the access stratum. Control plane protocols control PDU Sessions and the connection between UE and the network from various aspects which include requesting the service, controlling different transmission resources, handover etc. The mechanism for transparent transfer of NAS messages is also included. The layer 2 of the new radio protocol stack is split into four sublayers: medium access control (MAC), radio link control (RLC), packet data convergence protocol (PDCP), and service data adaptation protocol (SDAP), where each sublayer hosts a number of functions and performs certain functionalities that are configurable via radio resource control (RRC) at layer 3. Fig. 2.1 shows the OSI protocol layers and how they map to 3GPP radio protocol architecture. As shown in the figure (*dark-shaded boxes*), the data link layer of OSI network model maps to these four sublayers that constitute layer 2 of the 3GPP new radio protocols. The layer 2 protocols of NR have some similarities with the corresponding LTE protocols; however, the NR has added more configuration flexibility and more functionalities to support the new features such as beam management that have no counterpart in LTE. A significant difference in NR RRC compared to LTE RRC is the introduction of a 3-state UE behavior with the addition of the RRC_INACTIVE state. The RRC_INACTIVE provides a state with battery efficiency similar to RRC_IDLE while storing the UE context within the NG-RAN so that the transitions to/from RRC_CONNECTED are faster and incur less signaling overhead. The other significant improvements relative to LTE RRC are the support of on-demand system information

5G NR. DOI: https://doi.org/10.1016/B978-0-08-102267-2.00002-6

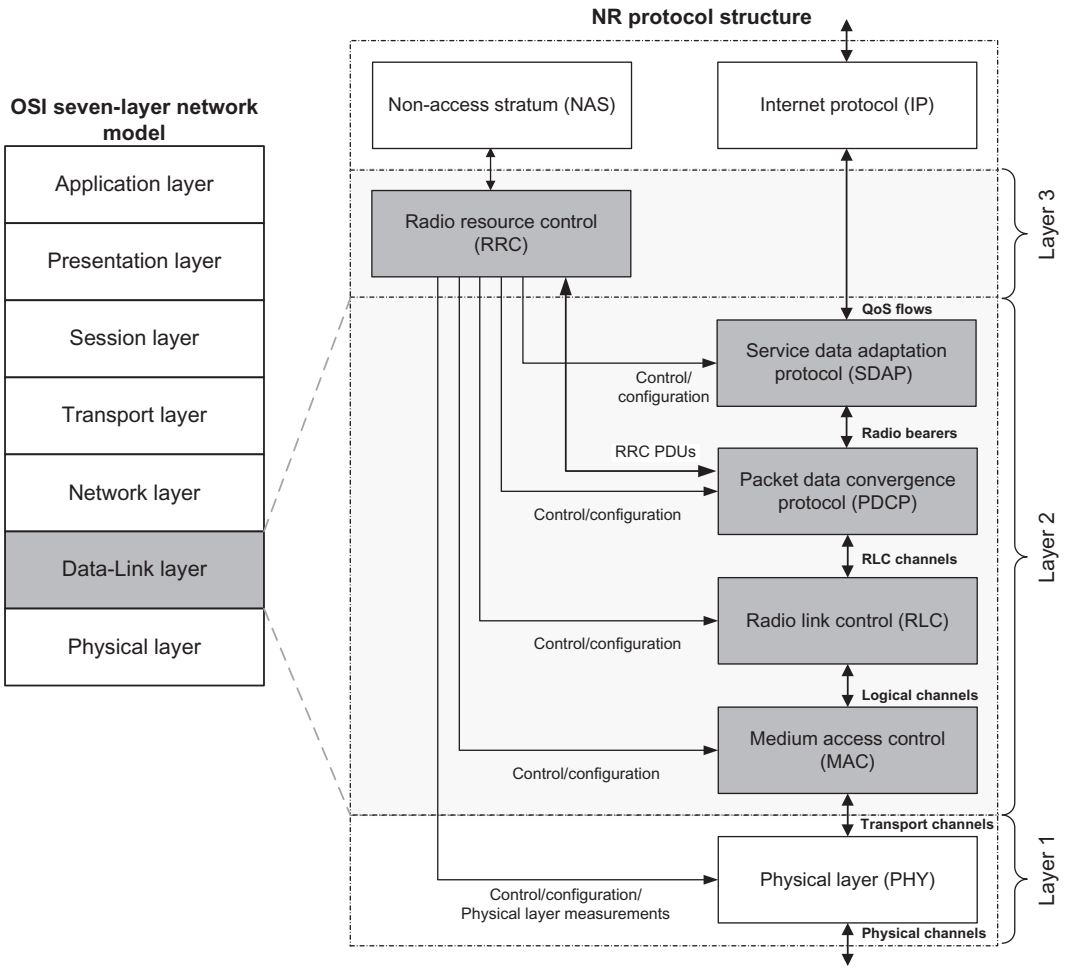


Figure 2.1
The layer 2/3 protocols in NR protocol stack [8].

transmission that enables the UE to request when specific system information is required instead of NG-RAN consuming radio resources to periodically broadcast system information, and the extension of the measurement reporting framework to support beam measurements for handover in a beamformed operation [8]. In NR layer 2 protocol structure, each sublayer provides certain services to the immediately adjacent layers by processing the incoming service data units (SDUs) and generating the proper protocol data units (PDUs). The functional processing of the protocol layers is further classified into user-plane and control-plane protocols (Fig. 2.2), where the reliability requirement of the control-plane information is much higher than that of the user-plane data.

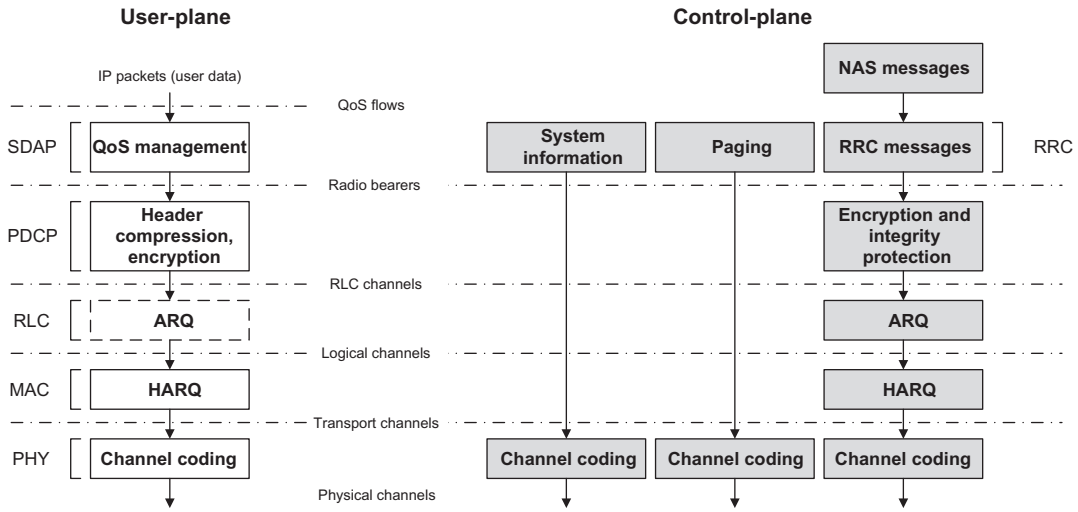


Figure 2.2
NR layer 2 user-plane and control-plane functional mapping and processing.

The service data adaptation protocol is a new sublayer in layer 2 which immediately interfaces with the network layer and provides a mapping between the QoS flows and data radio bearers (DRBs). It also marks the QoS flow identifiers (QFIs) in the downlink and uplink packets. A single-protocol entity of SDAP is configured for each individual PDU Session.

The services and functions of the PDCP sublayer on the user-plane include sequence numbering; header compression and decompression [only supports Robust Header Compression (ROHC) protocol]; transfer of user data; reordering and duplicate detection; in-sequence delivery; PDCP PDU routing (in case of split bearers); retransmission of PDCP SDUs; ciphering, deciphering, and integrity protection; PDCP SDU discard; PDCP re-establishment and data recovery for RLC acknowledged mode (AM); PDCP status reporting for RLC-AM; duplication of PDCP PDUs and duplicate discard indication to lower layers.

The main services and functions of the PDCP sublayer on the control-plane consist of sequence numbering; ciphering, deciphering, and integrity protection; transfer of control-plane data; reordering and duplicate detection; in-sequence delivery; duplication of PDCP PDUs and duplicate discard indication to lower layers [8]. The PDCP protocol performs (optional) IP-header compression, followed by ciphering, for each radio bearer. A PDCP header is added, carrying information required for deciphering in the other end, as well as a sequence number used for retransmission and in-sequence delivery.

The RLC sublayer supports three transmission modes, which include transparent mode (TM), unacknowledged mode (UM), and the acknowledged mode. The main difference between these

modes is the presence or absence of ARQ function to provide higher reliability required for RRC and NAS messages. The RLC configuration is per logical channel with no dependency on numerologies and/or transmission duration, and ARQ can operate on any of the numerologies and/or transmission durations that the logical channel is configured to support. The radio bearers in NR are classified into two groups of DRBs for user-plane data and signaling radio bearers (SRBs) for control-plane data. The RLC-TM mode is used for transport of SRB0, paging, and broadcast system information (SI), whereas for other SRBs, RLC-AM mode is used. For transport of DRBs, either RLC-UM or RLC-AM mode is used. Some of the services and functions of the RLC sublayer depend on the transmission mode. Those services include transfer of upper layer PDUs; sequence numbering independent of the one in PDCP (RLC-UM and RLC-AM); error correction through ARQ function (AM mode); segmentation (in RLC-AM and RLC-UM modes) and re-segmentation (in RLC-AM mode) of RLC SDUs; reassembly of SDU (in RLC-AM and RLC-UM modes); duplicate detection (in RLC-AM mode); RLC SDU discarding (in RLC-AM and RLC-UM modes); RLC re-establishment; and protocol error detection (in RLC-AM mode). The ARQ function within the RLC sublayer performs ARQ retransmission of RLC SDUs or SDU segments based on the RLC status reports. It further sends requests for RLC status reports when needed, and triggers an RLC status report after detecting a missing RLC SDU or SDU segment [8]. The RLC protocol performs segmentation of the PDCP PDUs, if necessary, and adds an RLC header containing a sequence number used for handling retransmissions. Unlike LTE, the NR RLC is not providing in-sequence delivery of data to higher layers due to the additional delay incurred by the reordering mechanism. In-sequence delivery can be provided by the PDCP layer, if necessary.

The services and functions of the MAC sublayer include mapping between logical and transport channels, as well as multiplexing/demultiplexing of MAC SDUs belonging to one or different logical channels on the transport blocks, which are mapped to the transport channels depending on the required physical layer processing. The MAC sublayer also performs scheduling of measurement reporting; error correction through HARQ [one HARQ entity per cell when carrier aggregation is utilized]; priority handling between user equipments through dynamic scheduling; and priority handling between logical channels of one UE through logical channel prioritization. A single MAC instantiation can support multiple numerologies, transmission timings, and cells. The mapping restrictions in logical channel prioritization can control which numerology(ies), cell(s), and transmission timing(s) a logical channel can use. The main difference between MAC and RRC control lies in the signaling reliability. The signaling corresponding to state transitions and radio bearer configurations should be performed by the RRC sublayer due to higher signaling reliability requirement. The RLC PDUs are delivered to the MAC sublayer, which multiplexes a number of RLC PDUs and adds a MAC header to form a transport block. It must be noted that in NR, the MAC headers are distributed across the MAC PDU, such that the MAC header corresponding to a particular RLC PDU is located immediately prior to it (see Fig. 2.3).

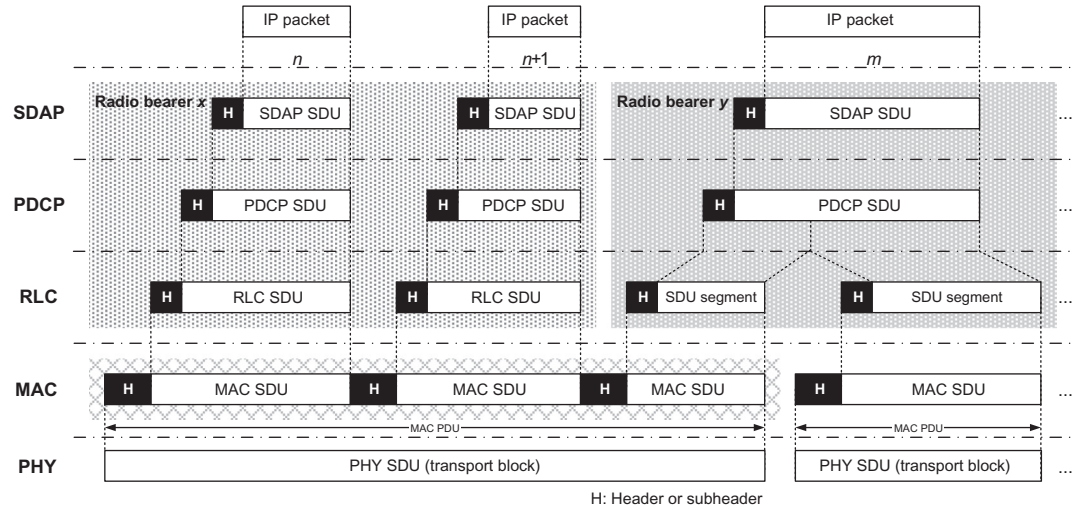


Figure 2.3
Layer 2 packet processing [8].

This is different compared to LTE, in which all header information are located at the beginning of the MAC PDU. The NR MAC PDUs can be assembled as soon as the RLC PDUs become available; thus there is no need to assemble the full MAC PDU before the header fields can be computed. This reduces the processing time and the overall latency [8,18].

The physical layer provides information transfer services to the MAC and higher layers. The physical layer transport services are described by how and with what characteristics data is transferred over the radio interface. This should be clearly distinguished from the classification of what is transported which relates to the concept of logical channels at the MAC sublayer.

It was mentioned earlier that each sublayer in layer 2 radio protocols receives SDUs from the previous layer, processes the information according to the configured functions and parameters of the sublayer, and generates PDUs that are delivered to the next layer. In this process, a unique header or subheader is attached to the SDU by each sublayer. An example is shown in Fig. 2.3, where a transport block is generated by MAC sublayer by means of concatenating two RLC PDUs from radio bearer x and one RLC PDU from radio bearer y. The RLC PDUs from radio bearer x each corresponds to one IP packet (n and $n + 1$) while the RLC PDU from radio bearer y is a segment of an IP packet (m) [8].

The services and functions of the RRC sublayer include broadcast of SI related to access stratum (AS) and NAS, as well as paging initiated by 5GC or NG-RAN (paging initiated by NG-RAN is a new NR feature). The RRC sublayer services further includes establishment, maintenance, and release of an RRC connection between the UE and NG-RAN that consist of addition, modification, and release of carrier aggregation; addition, modification, and release of

dual connectivity (DC) between LTE and NR. The security functions including key management, establishment, configuration, maintenance, and release of signaling and DRBs as well as mobility management, which comprises handover and context transfer; UE cell selection and reselection and control of cell selection and reselection; inter-RAT mobility; QoS management functions; UE measurement reporting and control of the reporting; detection of and recovery from radio link failure (RLF); and NAS message transfer are among other services provided by the RRC sublayer [8,15].

An NR UE at any time is in one of the three RRC states that are defined as follows [8]:

- *RRC_IDLE* which is characterized by PLMN selection; broadcast of system information; cell reselection; paging for mobile terminated data is initiated by 5GC; and discontinuous reception (DRX) for core-network paging configured by NAS.
- *RRC_INACTIVE* which is characterized by PLMN selection; broadcast of system information; cell reselection; paging initiated by NG-RAN (RAN paging); RAN-based notification area (RNA) is managed by NG-RAN; DRX for RAN paging configured by NG-RAN; 5GC and NG-RAN connection (both control and user-planes) establishment for the UEs; storage of UE AS context in NG-RAN and the UE; and NG-RAN knowledge of UE location at RNA-level.
- *RRC_CONNECTED* which is characterized by 5GC and NG-RAN connection (both control and user-planes) establishment for the UEs; storage of the UE AS context in NG-RAN and the UE; NG-RAN knowledge of UE location at cell level; transfer of unicast data between the UE and gNB; and network-controlled mobility including measurements.

The SI consists of a master information block (MIB) and a number of system information blocks (SIBs), which are divided into minimum SI and other SI. The minimum SI comprises basic information required for initial access and information for acquiring any other SI. The other SI encompasses all SIBs that are not broadcast in the minimum SI. Those SIBs can either be periodically broadcast on downlink shared channel (DL-SCH), broadcast on-demand on DL-SCH upon request from UEs in RRC_IDLE or RRC_INACTIVE states or sent in a dedicated manner on DL-SCH to UEs in RRC_CONNECTED state.

A UE is not required to acquire the content of the minimum SI of a cell/frequency that is considered for camping from another cell/frequency. This does not preclude the case that the UE applies stored SI from previously visited cell(s). A cell is barred, if a UE cannot determine/receive the full content of the minimum SI of that cell. In case of bandwidth adaptation the UE only acquires SI on the active BWP.

In the next sections, we will discuss layers 2 and 3 functions and procedures in more detail. We will further discuss UE states, state transitions, and important procedures such as idle,

inactive, and connected mode procedures, random-access procedure, mobility and power management, UE capability, and carrier aggregation.

2.2 Layer 2 Functions and Services

2.2.1 Medium Access Control Sublayer

The MAC sublayer performs logical channel multiplexing and controls HARQ retransmissions. It also handles scheduling functions and is responsible for multiplexing/demultiplexing data packets across multiple component carriers when carrier aggregation is configured. The services of MAC sublayer to the RLC sublayer are in the form of logical channels. A logical channel is defined by the type of information it carries and it is generally classified either as a control channel, for transmission of control and configuration information, or a traffic channel, for transmission of user data. The MAC sublayer provides services to the physical layer in the form of transport channels (see Fig. 2.4). A transport channel is defined by how and with what characteristics the information is transmitted over the radio interface. The information traversing a transport channel is organized in the form of transport blocks. In each physical layer transmission time interval (TTI), one transport block with dynamic size is transmitted over the radio interface to a device. In the case of spatial

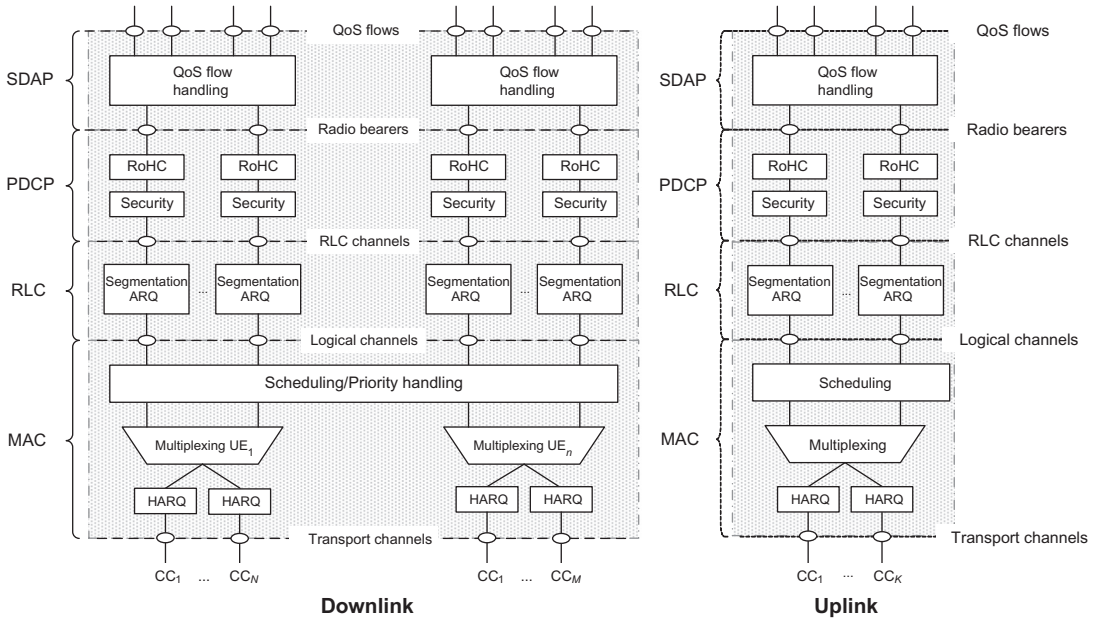


Figure 2.4
NR layer 2 functions in the downlink and uplink [8].

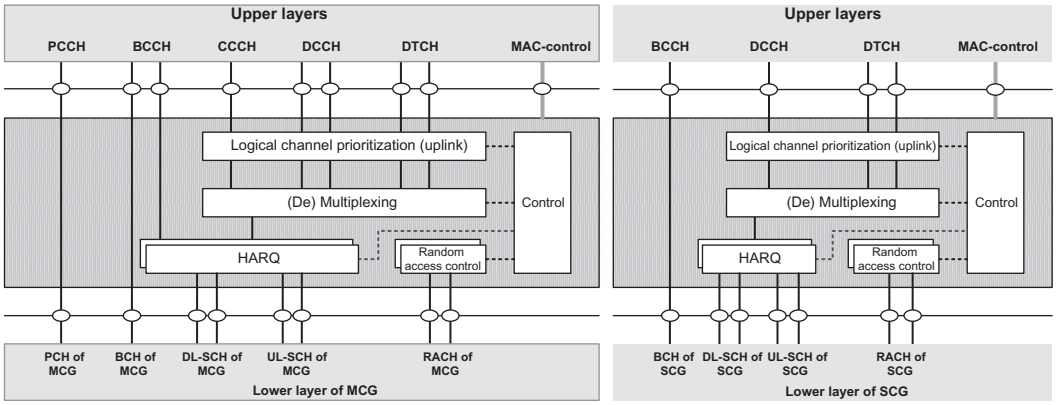


Figure 2.5
Example MAC structure with two MAC entities [9].

multiplexing when more than four layers are configured, there are two transport blocks per TTI. There is a transport format associated with each transport block, specifying how the transport block is transmitted over the radio interface. The transport format includes information about the transport block size, the modulation and coding scheme, and the antenna mapping. By varying the transport format, the MAC sublayer can realize different data rates, which is known as transport format selection.

The MAC entity in the UE manages the broadcast channel (BCH), DL-SCH, paging channel (PCH), uplink shared channel (UL-SCH)], and random-access channel (RACH). When the UE is configured with a secondary cell group (SCG), two MAC entities (instantiations) are configured where one of them is associated with master cell group (MCG) and another one is related to the SCG. Different MAC entities in the UE may operate independently, for example, the timers and parameters used in each MAC entity are independently configured. The serving cells, cell radio network temporary identifier (C-RNTI), radio bearers, logical channels, upper and lower layer entities, logical channel groups (LCGs), and HARQ entities are separately mapped to each MAC entity. If the MAC entity is configured with one or more SCells, there are multiple instances of DL-SCH, UL-SCH, and RACH per MAC entity; however, there is one instance of DL-SCH, UL-SCH, and RACH on the special cell (SpCell),[1] one DL-SCH, zero or one UL-SCH, and zero or one RACH for each SCell. If the MAC entity is not configured with any SCell, there is one instance of DL-SCH, UL-SCH, and RACH per MAC entity. Fig. 2.5 illustrates an example structure of the MAC entities when MCG and SCG are configured [9].

[1] In the context of dual connectivity, the SpCell refers to the primary cell of the MCG or the primary SCell of the SCG depending on whether the MAC entity is associated with the MCG or the SCG. Otherwise, the SpCell refers to the PCell. A SpCell supports PUCCH transmission and contention-based random-access procedure and it is always activated [9].

The MAC sublayer provides data transfer and radio resource allocation services to the upper layers and at the same time receives certain services from the physical layer which include data transfer, signaling of HARQ feedback, and scheduling request (SR), as well as conducting certain link-level measurements. The MAC sublayer provides a mapping between logical channels and transport channels; multiplexing of MAC SDUs from one or different logical channels to transport blocks to be delivered to the physical layer on transport channels; demultiplexing of MAC SDUs to one or different logical channels from transport blocks delivered from the physical layer on transport channels; scheduling measurement reporting; forward error correction through HARQ; and logical channel prioritization. The MAC sublayer further provides data transfer services via logical channels. To accommodate different types of data transfer services, various logical channels are defined, each supporting transfer of a particular type of information. Each logical channel is defined by the type of information which is transferred [9]. The MAC entity maps logical channels to transport channels in the uplink and downlink. This mapping depends on the multiplexing that is configured by RRC sublayer.

The priority handling among multiple logical channels, where each logical channel has its own RLC entity, is supported by multiplexing the logical channels to one transport channel. The MAC entity at the receiving side handles the corresponding demultiplexing and forwards the RLC PDUs to their respective RLC entity. To enable the demultiplexing function at the receiver, a MAC header is used. In NR, instead of putting the entire MAC header information at the beginning of a MAC PDU as LTE does, which implies that the assembly of a MAC PDU cannot start until the scheduling decision is available, the subheader corresponding to a certain MAC SDU is placed immediately in front of the SDU, as shown in Fig. 2.6. This allows the PDUs to be processed before a scheduling decision is received. If necessary, padding can be used to align the transport block size with those supported in NR.

A MAC subheader contains the identity of the logical channel (LCID) from which the RLC PDU originated, and the length of the PDU in bytes. There is also a flag indicating the size of the length indicator, as well as a reserved bit for future extension. In addition to multiplexing of different logical channels, the MAC sublayer can insert MAC control elements (CEs) in the transport blocks to be transmitted over the transport channels. A MAC CE is a form of in-band control signaling and is identified with reserved values in the LCID field, where the LCID value indicates the type of control information. Both fixed-length and variable-length MAC CEs are supported, depending on the use case. For downlink transmissions, MAC CEs are located at the beginning of the MAC PDU, whereas for uplink transmissions, the MAC CEs are located at the end, immediately before the padding (see Fig. 2.6). In some cases the size of padding can be zero. A MAC CE provides a faster way to send control signaling than RLC, without having to resort to the restrictions in terms of payload sizes and reliability offered by L1/L2 control signaling.

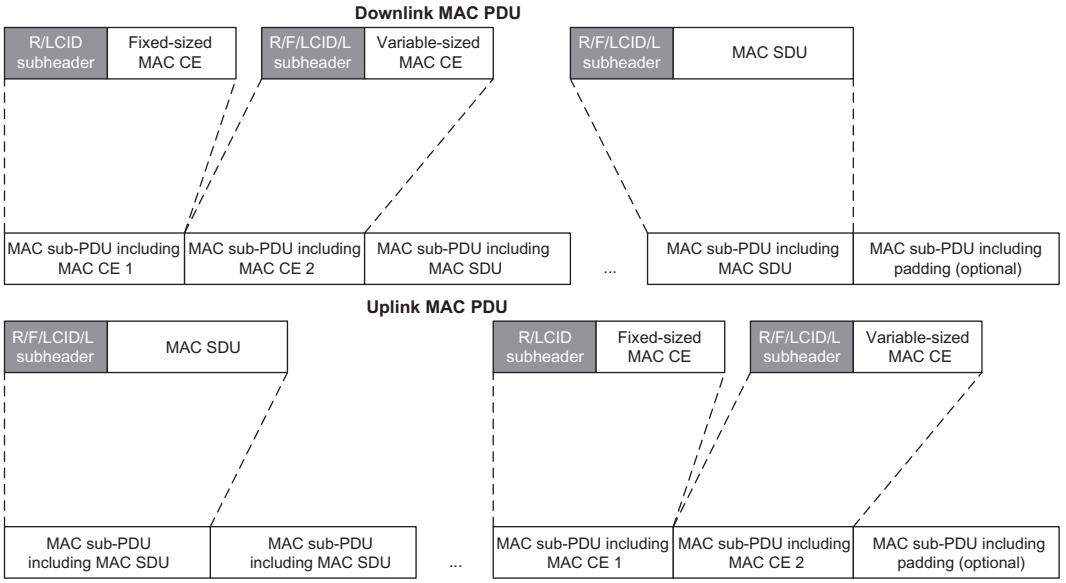


Figure 2.6
Example structures of downlink/uplink MAC PDUs [9].

A MAC PDU is a bit string that is octet-aligned and consists of one or more MAC sub-PDUs. Each MAC sub-PDU may consist of a subheader only (including padding); a subheader and a MAC SDU; a subheader and a MAC CE; or a subheader and padding. The MAC SDUs have variable sizes, where each MAC subheader corresponds to either a MAC SDU, a MAC CE, or padding. A MAC subheader typically consists of four header fields R/F/LCID/L (see Fig. 2.7). However, a MAC subheader for fixed-sized MAC CE, padding, and a MAC SDU containing uplink common control channel (CCCH), consists of two header fields R/LCID. MAC control elements are placed together. The downlink MAC subPDU(s) with MAC CE(s) is placed before any MAC subPDU with MAC SDU and MAC subPDU with padding. The uplink MAC subPDU(s) with MAC CE(s) is placed after all MAC subPDU(s) with MAC SDU and before the MAC subPDU with padding in the MAC PDU. Note that the size of padding can be zero. At most one MAC PDU can be transmitted per transport block per MAC entity. The aforementioned subheader fields are defined as follows [9]:

- *LCID*: The LCID field identifies the logical channel instance of the corresponding MAC SDU or the type of the corresponding MAC CE or padding. There is one LCID field per MAC subheader whose size is 6 bits.
- *L*: The length field indicates the length of the corresponding MAC SDU or variable-sized MAC CE in bytes. There is one L field per MAC subheader except for subheaders corresponding to fixed-sized MAC CEs, padding, and MAC SDUs containing uplink CCCH. The size of the L field is indicated by the F field.

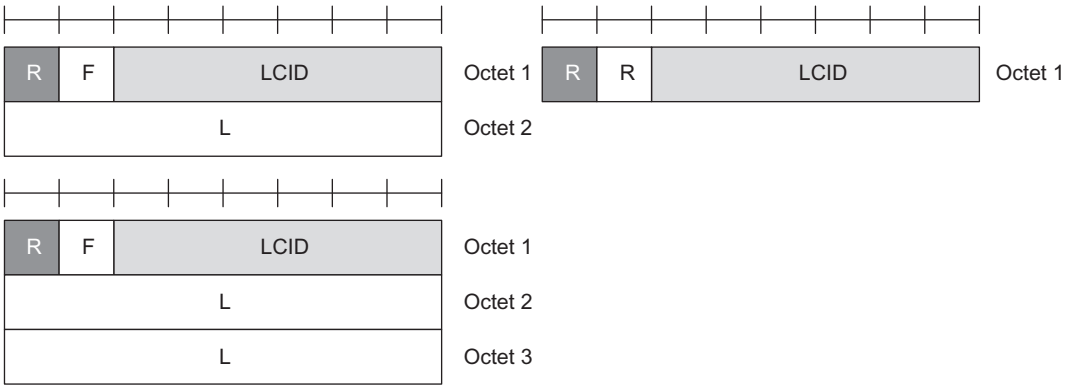


Figure 2.7
Structure of various MAC subheaders [9].

- *F*: The format field indicates the size of the length field. There is one F field per MAC subheader except for subheaders corresponding to fixed-sized MAC CEs, padding, and MAC SDUs containing uplink CCCH. The size of the F field is 1 bit, where the 8-bit and 16-bit Length fields are indicated by F = 0 and F = 1, respectively.
- *R*: Reserved bit, which is set to zero.

The MAC subheaders consist of the following fields:

- *E*: The extension field is a flag indicating whether the MAC sub-PDU, including the MAC subheader is the last MAC sub-PDU in the MAC PDU. It is set to one to indicate at least another MAC sub-PDU follows; otherwise, it is set to zero.
- *T*: The type field is a flag indicating whether the MAC subheader contains a random-access preamble identifier (RAPID) or a backoff indicator (BI). It is set to zero to indicate the presence of a BI field in the subheader; otherwise it is set to one to indicate the presence of a RAPID field in the subheader.
- *R*: Reserved bit, which is set to zero.
- *BI*: The BI field identifies the overload condition in the cell. The size of the BI field is 4 bit.
- *RAPID*: The RAPID field identifies the transmitted random-access preamble. The size of the RAPID field is 6 bits. If the RAPID in the MAC subheader of a MAC sub-PDU corresponds to one of the random-access preambles configured for SI request, MAC random-access response (RAR) is not included in the MAC sub-PDU. The MAC subheader is octet aligned.

There are a number of MAC CEs that are specified for the following purposes [9]:

- Duplication activation/deactivation
  - The duplication activation/deactivation MAC CE consists of one octet that is identified by a MAC PDU subheader with LCID index 56. It has a fixed size and contains eight D-fields, where the D_i field indicates the activation/deactivation status of the

PDCP duplication of the *i*th DRB associated with the RLC entities currently assigned to this MAC entity.

- SCell activation/deactivation
  - The SCell activation/deactivation MAC CE consists of either one or four octets that are identified by MAC PDU subheaders with LCID indices 58 or 57, respectively. It has a fixed size consisting of either a single octet or 4 octets and contains 7 or 31 C-fields (mapped to individual component carriers C_i) and one R-field.
- DRX
  - The DRX and long DRX command MAC CEs are identified by MAC PDU subheaders with LCID indices 60 and 59, respectively, both of which have a fixed size of zero bits.
- Timing advance command
  - The timing advance command MAC CE is identified by a MAC PDU subheader with LCID index 61. It has a fixed size of one octet and includes timing advance group (TAG) identity (TAG ID) which indicates the TAG ID of the addressed TAG (2 bit); and timing advance command which indicates the index value $TA = (0, 1, 2, \ldots, 63)$ that is used to control the amount of timing adjustment that MAC entity has to apply (6 bits).
- UE contention resolution identity
  - The UE contention resolution identity MAC CE is identified by a MAC PDU subheader with LCID index 62. It has a fixed 48 bit size and consists of a single field containing UE contention resolution identities.
- Semi-persistent (SP) CSI-RS/CSI-IM resource set activation/deactivation
  - The network may activate and deactivate the configured SP CSI-RS/CSI-IM resource sets of a serving cell by sending the SP CSI-RS/CSI-IM resource set activation/deactivation MAC CE. The configured SP CSI-RS/CSI-IM resource sets are initially deactivated upon configuration and after a handover.
- Aperiodic CSI trigger state sub-selection
  - The network may select among the configured aperiodic CSI trigger states of a serving cell by sending the aperiodic CSI trigger state sub-selection MAC CE.
- Transmission configuration indicator (TCI) states activation/deactivation for UE-specific PDSCH
  - The network may activate and deactivate the configured TCI states for physical downlink shared channel (PDSCH) of a serving cell by sending the TCI states activation/deactivation for UE-specific PDSCH MAC CE. The configured TCI states for PDSCH are initially deactivated upon configuration and after a handover.
- TCI state indication for UE-specific PDCCH
  - The network may indicate a TCI state for physical downlink control (PDCCH) reception for a CORESET of a serving cell by sending the TCI state indication for UE-specific PDCCH MAC CE.

- SP CSI reporting on physical uplink control channel (PUCCH) activation/deactivation
  - The network may activate and deactivate the configured SP CSI reporting on PUCCH of a serving cell by sending the SP CSI reporting on PUCCH activation/deactivation MAC CE. The configured SP CSI reporting on PUCCH is initially deactivated upon configuration and after a handover.
- SP SRS activation/deactivation
  - The network may activate and deactivate the configured SP SRS resource sets of a serving cell by sending the SP SRS activation/deactivation MAC CE. The configured SP SRS resource sets are initially deactivated upon configuration and after a handover.
- PUCCH spatial relation activation/deactivation
  - The network may activate and deactivate a spatial relation for a PUCCH resource of a serving cell by sending the PUCCH spatial relation activation/deactivation MAC CE.
- SP zero power (ZP) CSI-RS resource set activation/deactivation
  - The network may activate and deactivate the configured SP ZP CSI-RS resource set of a serving cell by sending the SP ZP CSI-RS resource set activation/deactivation MAC CE. The configured SP ZP CSI-RS resource sets are initially deactivated upon configuration and after a handover.
- Recommended bit rate
  - The recommended bit rate procedure is used to provide the MAC entity with information about the physical layer bit rate which the gNB recommends. An averaging window with default size of 2 seconds is applied. The gNB may transmit the recommended bit rate MAC CE to the receiver-side MAC entity to indicate the recommended bit rate for the UE for a specific logical channel and downlink/uplink direction.
- Single and multiple-entry power headroom report (PHR)
  - The single-entry PHR MAC CE is identified by a MAC PDU subheader with LCID index 57. It has a fixed size of two octets, including power headroom which indicates the PH level (6 bits) and $P_{CMAX_{lm}}$ denoting the maximum permissible UE transmit power for carrier l and serving cell m.
  - The multiple-entry PHR MAC CE is identified by a MAC PDU subheader with LCID index 56. It has a variable size and may include a bitmap, a Type 2 PH field and an octet containing the associated $P_{CMAX_{lm}}$ field for the SpCell of the other MAC entity, as well as a Type 1 PH field and an octet containing the associated $P_{CMAX_{lm}}$ for the primary cell (PCell). It may further include one or more Type x PH fields (x is either 1 or 3) and octets containing the associated $P_{CMAX_{lm}}$ fields for serving cells other than PCell indicated in the bitmap. The MAC entity determines whether power headroom value for an activated serving cell is based on a real transmission or a reference format by considering the configured grant(s) and the downlink control information (DCI) that has been received prior to the PDCCH occasion

in which the first uplink grant for a new transmission is received after a PHR has been triggered.

- Configured grant confirmation
  - The configured grant confirmation MAC CE is identified by a MAC PDU subheader with LCID index 55, which has a fixed size of zero bits.
- C-RNTI
  - The C-RNTI MAC CE is identified by a MAC PDU subheader with LCID index 58. It has a fixed size and consists of a single field containing the C-RNTI. The length of the field is 16 bits.
- Buffer status report (BSR)
  - The BSR MAC CEs consist of short BSR format (fixed size), long BSR format (variable size), short truncated BSR format (fixed size), and long truncated BSR format (variable size). The BSR formats are identified by MAC PDU subheaders with LCID indices 59, 60, 61, and 62 corresponding to short truncated BSR, long truncated BSR, short BSR, and long BSR MAC CEs, respectively. The fields in the BSR MAC CE contain an LCG ID, denoting the LCG ID which identifies the group of logical channels whose buffer status are being reported, and one or several LCG_i for the long BSR format, which indicate the presence of the buffer size field for the *i*th LCG. For the long truncated BSR format, this field indicates whether the *i*th LCG has data available. The buffer size field identifies the total amount of data available according to the data volume calculation procedure in references [10,11] across all logical channels of an LCG after the MAC PDU has been created, that is, after the logical channel prioritization procedure, which may result in setting the value of the buffer size field to zero. The amount of data is indicated in number of bytes. The size of the RLC and MAC headers are not considered in the buffer size calculation. The length of this field for the short BSR format and the short truncated BSR format is 5 bits and for the long BSR format and the long truncated BSR format is 8 bits. For the long BSR format and the long truncated BSR format, the buffer size fields are included in ascending order based on the LCG_i. For the long truncated BSR format the number of buffer size fields included is maximized, as long as it does not exceed the number of padding bits.

The structure and content of some MAC CEs are shown in Fig. 2.8.

As we mentioned earlier, a [transparent] MAC PDU may only consist of a MAC SDU whose size is aligned to a permissible transport block. This MAC PDU is used for transmissions on PCH, BCH, and DL-SCH including BCCH. Furthermore, a random-access response MAC PDU format is defined which consists of one or more MAC sub-PDUs and may contain some padding bits. Each MAC sub-PDU consists of a MAC sub-header with backoff indicator; a MAC sub-header with random access preamble identifier (RAPID), which is an acknowledgment for SI request; or a MAC sub-header with RAPID and MAC

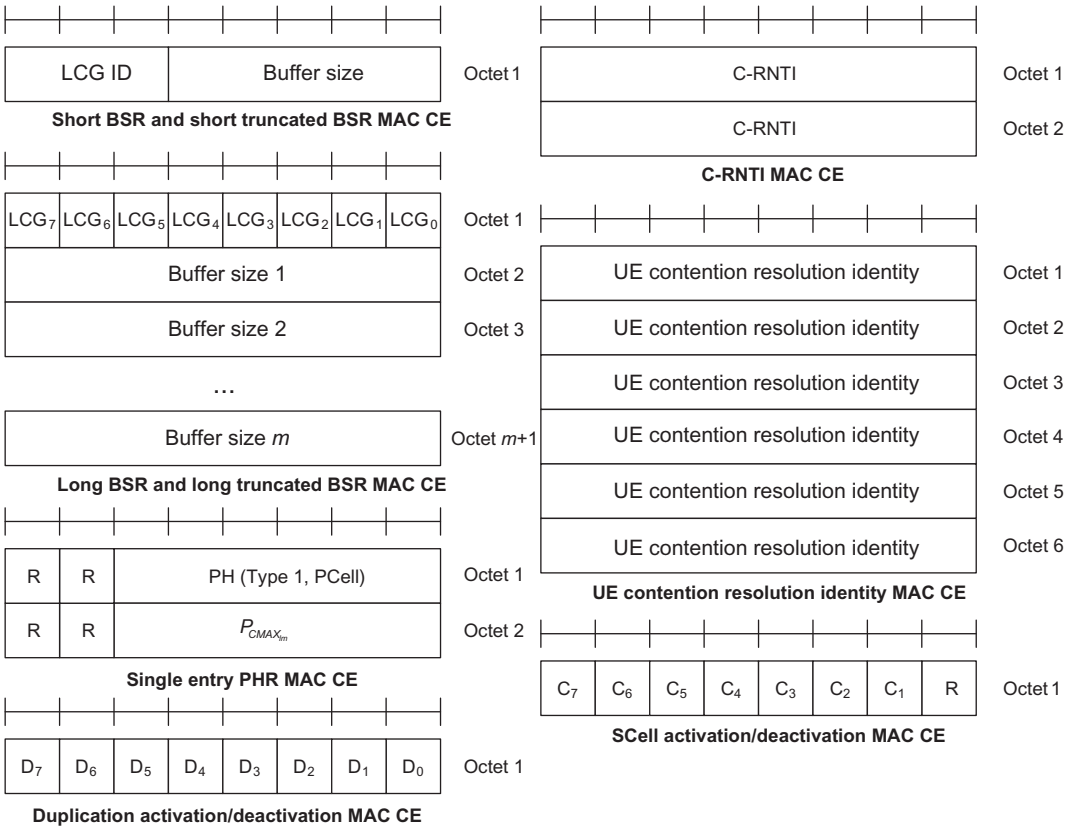


Figure 2.8
Some MAC control element formats [9].

RAR. A MAC sub-header with backoff indicator consists of five header fields E/T/R/R/BI (see Fig. 2.9). A MAC sub-PDU with backoff indicator is placed at the beginning of the MAC PDU, if included. The 'MAC sub-PDU(s) with RAPID' and 'MAC sub-PDU(s) with RAPID and MAC RAR' can be placed anywhere between MAC sub-PDU with backoff indicator and the padding bits. A MAC sub-header with RAPID consists of three header fields E/T/RAPID. Padding bits are placed at the end of the MAC PDU, if necessary. The presence and the length of the padding bits are based on the transport block size and the size of the MAC sub-PDU(s).

The MAC entity is further responsible for distributing IP packets corresponding to each flow across different component carriers or cells, when carrier aggregation is utilized. The carrier aggregation relies on independent processing of the component carriers in the physical layer, including control signaling, scheduling, and HARQ retransmissions. The MAC sublayer makes multi-carrier operation transparent to the upper sublayers/layers. The logical

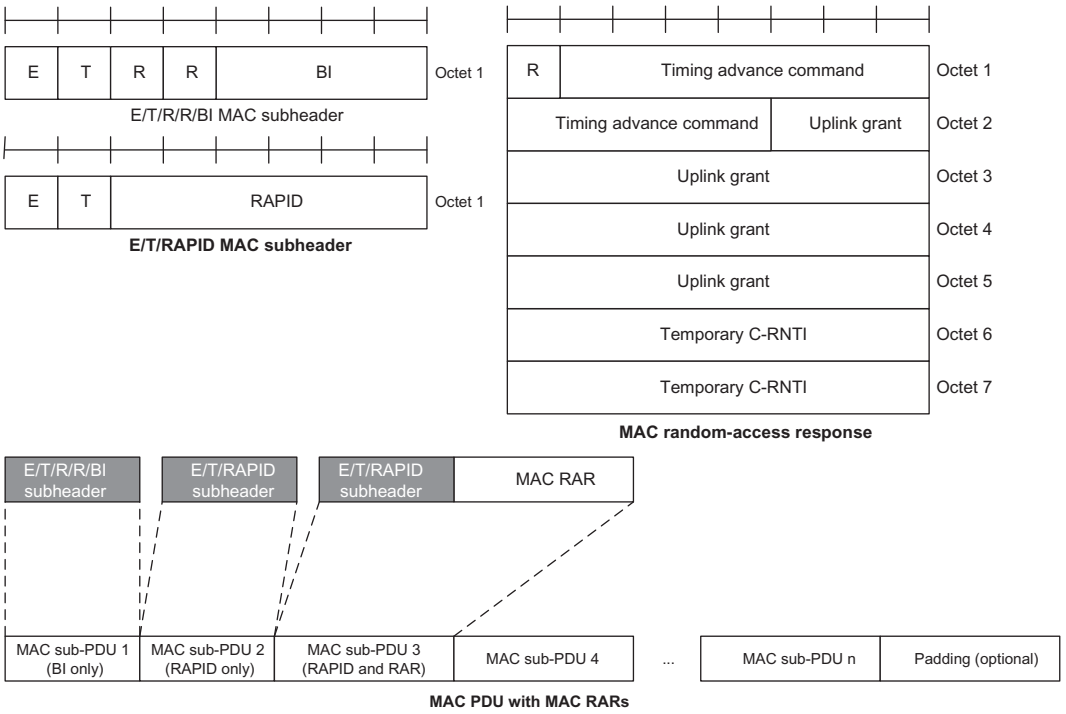


Figure 2.9
Structure of MAC subheaders and example MAC PDU with random-access response [9].

channels, including any MAC CEs, are multiplexed to form transport blocks per component carrier, where each component carrier has its own HARQ entity.

In order to efficiently utilize the radio resources, the MAC sublayer in gNB includes schedulers that allocate radio resources in the downlink/uplink to the active users in the cell. The default mode of operation of a gNB scheduler is dynamic scheduling, in which the gNB makes scheduling decisions, typically once per slot, and sends the scheduling information to a group of devices. While slot-based scheduling is a common practice, neither the scheduling decisions, nor the actual data transmission are required to start or end at the slot boundaries, which is important for low-latency applications and unlicensed spectrum operation. The downlink/uplink scheduling are independent and scheduling decisions can be independently made.

The downlink scheduler dynamically controls the radio resources allocated to active devices in order to efficiently share the DL-SCH. The selection of transport format which includes transport block size, modulation and coding scheme, and antenna mapping, as well as logical channel multiplexing for downlink transmissions are all managed by the gNB scheduler.

The uplink scheduler serves a similar purpose, that is, dynamically controlling the transmission opportunities of the active UEs and efficiently sharing of UL-SCH. The scheduling strategy of a gNB is vendor-dependent and implementation specific and is not specified by 3GPP. In general, the ultimate goal of all schedulers is to take advantage of the channel variations experienced by various devices and to schedule the transmissions/receptions to/from each device on the radio resources with advantageous channel conditions in time and frequency domains, that is, channel-dependent scheduling [8].

The downlink channel-dependent scheduling is supported through periodic/aperiodic channel state information reports sent by the devices to the gNB, which provide information on the instantaneous downlink channel quality in time and frequency domains, as well as information to determine an appropriate antenna/beam configuration. In the uplink, the channel state information required for channel-dependent scheduling can be obtained from the sounding reference signals transmitted by each device to the gNB. In order to assist the uplink scheduler, the device can further transmit BSR (measuring the data that is buffered in the logical channel queues in the UE) and PHR (measuring the difference between the nominal UE maximum transmit power and the estimated power for uplink transmission) to the gNB using MAC CEs. This information can only be transmitted, if the device has been given a valid scheduling grant. While dynamic scheduling is the default operation mode of many base station schedulers, there are cases where semi-persistent scheduling (SPS) is preferred due to reduced signaling overhead [17].

In the downlink, the gNB can dynamically allocate resources to the active UEs and identify them by their C-RNTIs on PDCCH(s). A UE always monitors the PDCCH(s) in order to find possible assignments according to its configured DRX cycles. When carrier aggregation is configured, the same C-RNTI applies to all serving cells. The gNB may preempt an ongoing PDSCH transmission to one UE with a delay-sensitive transmission to another UE. The gNB can configure the UEs to monitor interrupted transmission indications using INT-RNTI on a PDCCH. If a UE receives the interrupted transmission indication, it may assume that no useful information was intended for it by the resource elements included in the indication, even if some of those resource elements were already scheduled for that UE. Furthermore, the gNB can allocate downlink resources for the initial HARQ transmissions to the UEs configured with SPS. In that case, the RRC signaling defines the periodicity of the configured downlink assignments while PDCCH addressed to CS-RNTI can either signal and activate the configured downlink assignment or deactivate it, that is, a PDCCH addressed to CS-RNTI indicates that the downlink assignment can be implicitly reused according to the periodicity defined by the RRC signaling until it is deactivated [8].

The dynamically allocated downlink reception overrides the configured downlink assignment in the same serving cell, if they overlap in time. When carrier aggregation or bandwidth adaptation is configured, one configured downlink assignment can be signaled per

serving cell or per BWP, respectively. In each serving cell, the gNB can only configure one active downlink assignment for a UE at a time; however, it can simultaneously configure multiple active downlink assignments on different serving cells. Activation and deactivation of configured downlink assignments are independent among the serving cells [8].

In the uplink, the gNB can dynamically allocate resources to the active UEs and identify them by their C-RNTI on PDCCH(s). A UE always monitors the PDCCH(s) in order to find possible grants for uplink transmission according to its configured DRX cycles. When carrier aggregation is configured, the same C-RNTI applies to all serving cells. In addition, the gNB can allocate uplink resources for the initial HARQ transmissions to the UEs with the configured grants (alternatively known as semi-persistent scheduling). Two types of configured uplink grants are specified in NR. In Type 1 configured grant the RRC signaling directly provides the configured uplink grant (including the periodicity), whereas in Type 2 configured grant, the RRC signaling defines the periodicity of the configured uplink grant, while PDCCH addressed to CS-RNTI can either signal and activate the configured uplink grant or deactivate it. Therefore, a PDCCH addressed to CS-RNTI indicates that the uplink grant can be implicitly reused according to the periodicity defined through RRC signaling until it is deactivated. The dynamically allocated uplink transmission overrides the configured uplink grant in the same serving cell, if they overlap in time. The retransmissions other than repetitions are required to be explicitly allocated via PDCCH(s) [8]. If carrier aggregation or bandwidth adaptation is configured, one configured uplink grant can be signaled per serving cell or per BWP, respectively. In each serving cell, there is only one active configured uplink grant at a time. A configured grant for a serving cell in the uplink can either be Type 1 or Type 2. The activation and deactivation of configured grants for the uplink are independent among the serving cells for Type 2. In the case of supplemental uplink (SUL), a configured grant can only be signaled for one of the two uplink carriers of the cell [8].

When a downlink assignment is configured for SPS, the UE MAC entity will assume that the Nth downlink assignment occurs in the slot number that satisfies the following criterion [9]:

$$\left(N_{frame}^{slot} SFN + n_{slot}\right) = \left[\left(N_{frame}^{slot} SFN_{start-time} + slot_{start-time}\right) + NT_{SPS} N_{frame}^{slot}/10\right] \bmod \left(1024 N_{frame}^{slot}\right)$$

where $SFN_{start-time}$ and $slot_{start-time}$ are the system frame number (SFN) and slot of the first transmission of PDSCH where the configured downlink assignment was (re)initialized.

When a uplink grant is configured for a configured grant Type 1, the UE MAC entity will assume that the uplink grant repeats at symbols for which the following criterion is met [9]:

$$\left[\left(N_{frame}^{slot}N_{slot}^{symbol}SFN\right)+\left(n_{slot}N_{slot}^{symbol}\right)+n_{symbol}\right]=\left(N_{slot}^{symbol}T_{offset}+S+NT_{SPS}\right)\times \mathrm{mod}\left(1024N_{frame}^{slot}N_{slot}^{symbol}\right)\forall N\geq 0$$

When a uplink grant is configured for a configured grant Type 2, the UE MAC will assume that the uplink grant repeats with each symbol for which the following criterion is satisfied [9]:

$$\left[\left(N_{frame}^{slot}N_{slot}^{symbol}SFN\right)+\left(n_{slot}N_{slot}^{symbol}\right)+n_{symbol}\right]=\left[\left(N_{frame}^{slot}N_{slot}^{symbol}SFN_{start-time}+slot_{start-time}N_{slot}^{symbol}+symbol_{start-time}\right)+NT_{SPS}\right]\times \mathrm{mod}\left(1024N_{frame}^{slot}N_{slot}^{symbol}\right)\forall N\geq 0$$

where $SFN_{start-time}$, $slot_{start-time}$, and $symbol_{start-time}$ are the SFN, slot, and symbol corresponding to the first transmission opportunity of physical uplink shared channel (PUSCH), where the configured uplink grant was (re)initialized.

Measurement reports are required to enable the MAC scheduler to make scheduling decisions in the downlink and uplink. These reports include transport capabilities and measurements of UEs instantaneous radio conditions. The uplink BSRs are needed to support QoS-aware packet scheduling. In NR, uplink BSRs refer to the data that is buffered for an LCG in the UE. There are eight LCGs and two reporting formats in the uplink: a short format to report only one BSR (of one LCG) and a flexible long format to report several BSRs (up to eight LCGs). The uplink BSRs are transmitted using MAC CEs. When a BSR is triggered upon arrival of data in the transmission buffers of the UE, an SR is transmitted by the UE. The PHRs are needed to support power-aware packet scheduling. In NR, there are three types of reporting, that is, for PUSCH transmission, PUSCH and PUCCH transmission, and SRS transmission. In case of carrier aggregation, when no transmission takes place on an activated SCell, a reference power is used to provide a virtual report. The PHRs are transmitted using MAC CEs [8].

The HARQ functionality in the MAC sublayer ensures reliable transport of MAC PDUs between peer entities over the physical layer. A HARQ mechanism along with soft combining can provide robustness against transmission errors. A single HARQ process supports one or multiple transport blocks depending on whether the physical layer is configured for downlink/uplink spatial multiplexing. An asynchronous incremental redundancy HARQ protocol is supported in the downlink and uplink. The gNB provides the UE with the HARQ-ACK feedback timing either dynamically via DCI or semi-statically through an RRC configuration message. The UE may be configured to receive code-block-group−based transmissions where retransmissions may be scheduled to carry a subset of the code blocks

included in a transport block. The gNB schedules each uplink transmission and retransmission using the uplink grant on DCI [8].

The HARQ protocol in NR uses multiple parallel stop-and-wait processes. When a transport block is received at the receiver, it attempts to decode the packet and to inform the transmitter about the outcome of the decoding process through an ACK bit indicating whether the decoding process was successful or requesting the retransmission of the transport block. The HARQ-ACKs are sent by the receiver based on a specific timing relationship between UL HARQ-ACKs and DL HARQ processes or based on the position of the ACK bit in the HARQ-ACK codebook when multiple ACKs are transmitted simultaneously. An asynchronous HARQ protocol is used for both downlink and uplink, that is, an explicit HARQ process number is used to identify a particular process, since the retransmissions are scheduled in the same way as the initial transmission (explicit scheduling). The use of an asynchronous UL HARQ protocol, instead of a synchronous one that was used in LTE, was deemed to be necessary to support dynamic TDD where there is no fixed UL/DL allocation. It also provides more flexibility in terms of prioritization of data flows and devices and is further useful for operation in unlicensed spectrum. The NR supports up to 16 HARQ processes. The larger number of HARQ processes (compared to LTE) was motivated by the disaggregated RAN architectures and consideration for remote radio heads and fronthaul transport delay, as well as the use of shorter slot durations at high-frequency bands. It must be noted that the larger number of maximum HARQ processes does not imply a longer roundtrip delay, since the decoding will typically succeed after a few retransmissions depending on the channel conditions. Note that the PDCP sublayer can provide in-sequence delivery; thus this function is not provided by the RLC sublayer in order to reduce the latency.

A new feature of the NR HARQ protocol (compared to LTE) is the possibility for retransmission of code block groups, which is useful for very large transport blocks or when a transport block is partially preempted by another transmission. As part of the channel coding operation in the physical layer, a transport block is split into one or more code blocks with channel coding applied to each of the code blocks of up to 8448 bits in order to maintain a reasonable complexity. In practice and in the presence of burst errors, only a few code blocks in the transport block may be corrupted and the majority of code blocks are correctly received. In order to correctly deliver the transport blocks to the destination MAC sublayer, it is sufficient to only retransmit the erroneous code blocks. Furthermore, to avoid the excessive control signaling overhead due to individual code block addressing by HARQ mechanism, code block groups have been defined. If per-CBG retransmission is configured, feedback is provided per-CBG and only the erroneously received code block groups are retransmitted. The CBG-based retransmissions are transparent to the MAC sublayer and are handled in the physical layer. From MAC sublayer perspective, the transport block is not correctly received until all CBGs are correctly received. It is not possible to mix the CBGs

belonging to another transport block with retransmissions of CBGs belonging to the incorrectly received transport block in the same HARQ process [17].

2.2.2 Radio Link Control Sublayer

The RLC sublayer is located between PDCP and MAC sublayers as shown in Fig. 2.1. The RRC sublayer controls and configures the RLC functions. The RLC functions are performed by the RLC entities. For an RLC entity configured at the gNB, there is a peer RLC entity configured at the UE and vice versa. An RLC entity receives/delivers RLC SDUs from/to the upper layer and sends/receives RLC PDUs to/from its peer RLC entity via lower layers. The RLC sublayer can operate in one of the three modes of operation defined as transparent mode, unacknowledged mode, and acknowledged mode. Depending on the mode of operation, the RLC entity controls the usage of error correction, segmentation, resegmentation, reassembly, and duplicate detection of SDUs. An RLC entity in any mode can be configured either as a transmitting or a receiving entity. The transmitting RLC entity receives RLC SDUs from the upper layer and sends RLC PDUs to its peer receiving RLC entity via lower layers. The receiving RLC entity delivers RLC SDUs to the upper layer and receives RLC PDUs from its peer transmitting RLC entity via lower layers. Fig. 2.10 illustrates the

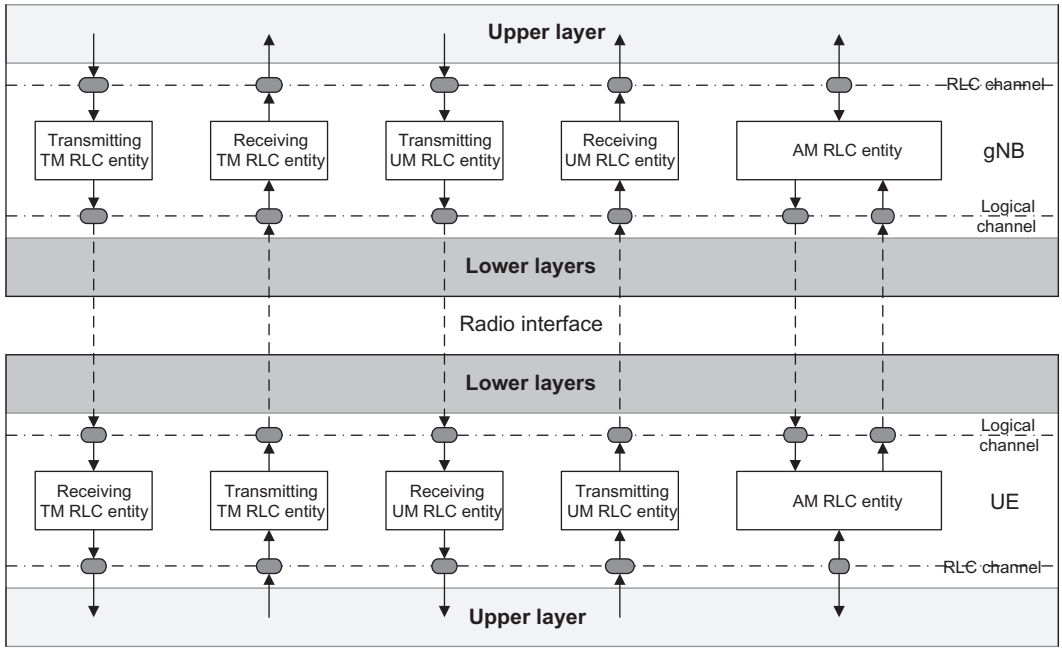


Figure 2.10
High-level architecture of RLC sublayer [10].

high-level architecture of the RLC sublayer. The [octet-aligned] RLC SDUs of variable sizes are supported for RLC entities. Each RLC SDU is used to construct an RLC PDU without waiting for notification from the MAC sublayer for a transmission opportunity. In the case of RLC-UM and RLC-AM entities, an RLC SDU may be segmented and transported using two or more RLC PDUs based on the notification(s) from the lower layer [10].

In RLC-TM, no RLC headers are added. The RLC-UM supports segmentation and duplicate detection, while the RLC-AM supports retransmission of erroneous packets. A key difference with LTE is that the NR RLC sublayer does not handle in-sequence delivery of SDUs to the upper layers. Eliminating the in-sequence delivery function from the RLC sublayer reduces the overall latency since the correctly received packets do not have to wait for retransmission of an earlier (missing) packet before being delivered to the higher layers. Another difference is the removal of concatenation function from the RLC protocol to allow RLC PDUs to be assembled prior to receiving the uplink scheduling grant. By eliminating the concatenation from RLC functions the RLC PDUs can be assembled in advance and upon receipt of the scheduling decision the device can forward a number of RLC PDUs to the MAC sublayer depending on the scheduled transport block size.

The RLC retransmission mechanism is used to provide error correction of data delivered to higher layers. The retransmission protocol operating between the RLC entities in the receiver and transmitter monitors the sequence numbers indicated in the headers of the incoming PDUs. The receiving RLC entity can identify missing PDUs based on the RLC sequence number which is independent of the PDCP sequence number. The status reports are sent to the transmitting RLC entity, requesting retransmission of the missing PDUs. Based on the received status report, the RLC entity at the transmitter can retransmit the missing PDUs. Even though the RLC sublayer is capable of correcting transmission errors, the error correction in most cases is handled by the MAC sublayer HARQ protocol. However, RLC sublayer and MAC sublayer retransmission mechanisms are meant for different purposes to achieve a highly reliable transmission for certain applications.

The HARQ mechanism allows fast retransmissions and feedback based on the success or failure of a downlink transmission after receiving a transport block. For the uplink transmissions, no explicit feedback needs to be transmitted because the receiver and scheduler are located in the same node. While in theory, it is possible to attain a very low block error rate using HARQ mechanism, in practice, it comes at a cost additional signaling and radio resources as well as increased power consumption. Therefore, the target block error rate in practice is less than 1%, which inevitably results in a residual error. For many applications such as voice over IP (VoIP), this residual error rate is sufficiently low and acceptable (because voice decoders can tolerate some level of frame erasure); however, there are use cases where lower block error rates are required, for example, transmission of RRC and NAS messages.

A sufficiently low block error rate is not only required for URLLC services, but it also is important from a system-level perspective in terms of sustaining data rate performance. The TCP protocol (at transport layer) requires virtually error-free delivery of packets to the peer TCP layer. As an example, to obtain sustained data rates in the excess of 100 Mbps in TCP/IP applications, a packet error rate of less than 10^{-5} is required, because the TCP protocol would consider packet errors due to network congestion and would trigger the congestion-avoidance mechanism to decrease the data rate, causing reduction of the overall data rate performance of the system. It must be noted that the infrequent transmission of RLC status reports compared to more frequent HARQ retransmission makes obtaining a reliability level of 10^{-5} using RLC ARQ retransmissions more practical [17]. Therefore, complementing MAC sublayer HARQ protocol with RLC sublayer ARQ mechanism would achieve relatively lower latency and reasonable feedback overhead to satisfy the stringent requirements of URLLC applications. The RLC sublayer ARQ mechanism retransmits RLC SDUs or RLC SDU segments based on RLC status reports, where polling may be used to request RLC status reports. An RLC receiver can also trigger RLC status report after detecting a missing RLC SDU or RLC SDU segment [8]. The RLC-TM entity is configured to transmit/receive RLC PDUs through BCCH, DL/UL CCCH, and paging control channel logical channels. The RLC-UM entity is configured to transmit/receive RLC PDUs via downlink or uplink dedicated traffic channel (DTCH). The RLC-AM entity can be configured to transmit/receive RLC PDUs through downlink or uplink dedicated control channel (DCCH) or DL/UL DTCH logical channels. Functional models of RLC-AM and RLC-UM entities are illustrated in Fig. 2.11, where the functions that are only related to RLC-AM mode (ARQ functions) are marked with a dark color.

The RLC PDUs and SDUs are bit strings that are octet-aligned. The TM data (TMD PDU) consists only of a data field and does not include any RLC headers. In RLC-AM and RLC-UM modes, a sequence number is generated and attached to the incoming SDUs using 6 or 12 bits for the RLC-UM and 12 or 18 bits for the RLC-AM. The sequence number is included in the RLC PDU header as shown in Fig. 2.12. If the SDU is not segmented, the RLC PDU consists of the RLC SDU and a header, which allows the RLC PDUs to be generated in advance as the header does not depend on the transport block size. However, depending on the transport block size after multiplexing at MAC sublayer, the size of the last RLC PDU in a transport block may not match the RLC SDU size, thus requiring dividing the SDU into multiple segments. If no segmentation is done, padding need to be used which would adversely impact the spectral efficiency. As a result, dynamically varying the number of RLC PDUs in a transport block along with segmentation to adjust the size of the last RLC PDU, ensures that the transport block is efficiently utilized. Segmentation is done by dividing the last preprocessed RLC SDU into two segments, the header of the first segment is updated, and a new header is added to the second segment as shown in Fig. 2.3. Each RLC SDU segment carries the same sequence number as the original SDU and the

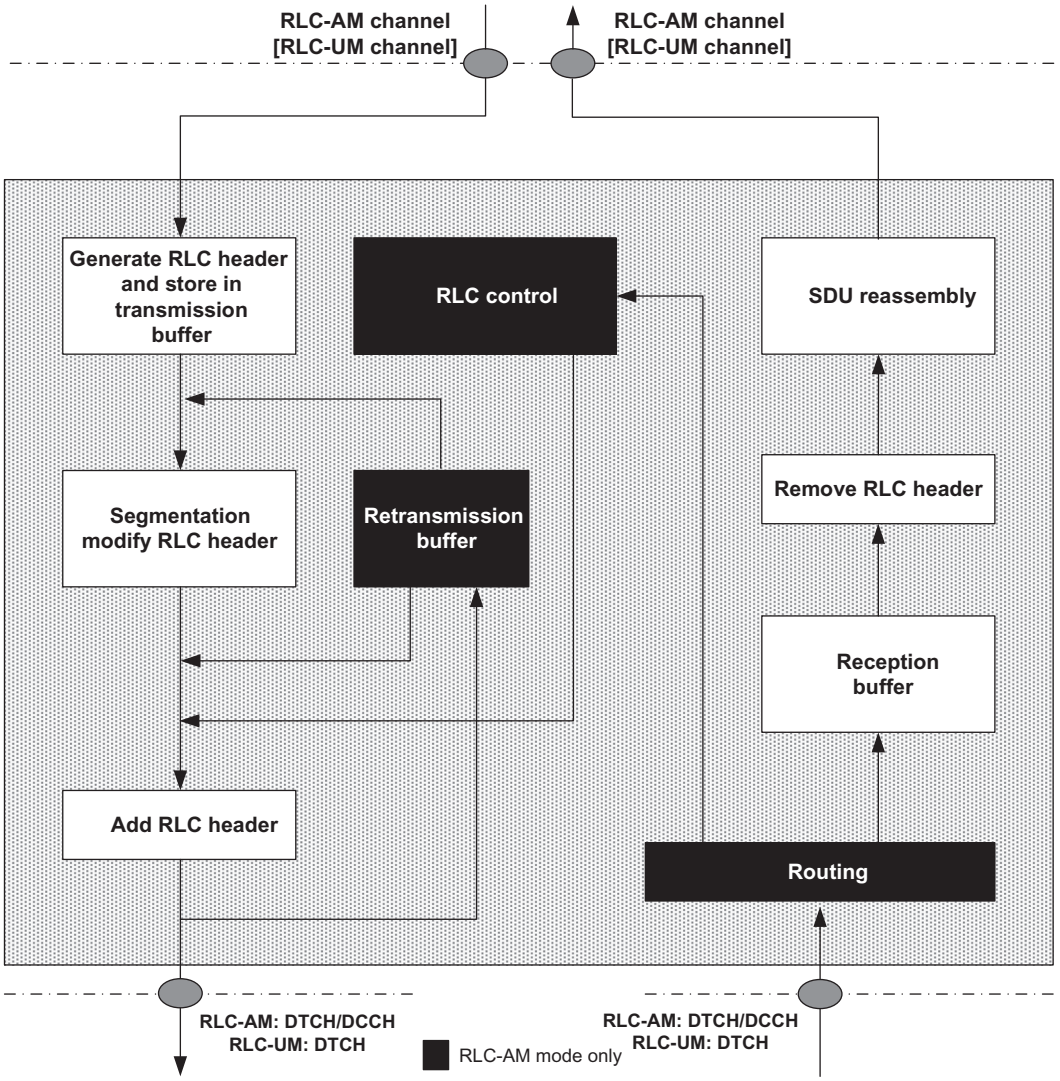


Figure 2.11
RLC-AM and RLC-UM entity models [10].

sequence number is part of the RLC header. To distinguish an unsegmented RLC PDU from a segmented one, a segmentation information field is included in the RLC header, indicating whether the PDU is a complete SDU, the first segment of the SDU, the last segment of the SDU, or a segment between the first and last segments of the SDU. Furthermore, in the case of a segmented SDU, a 16 bit segmentation offset (SO) is included in all segments except the first one to indicate which byte of the SDU is represented by the segment. The RLC header may further include a poll (P) bit, which is used to request status

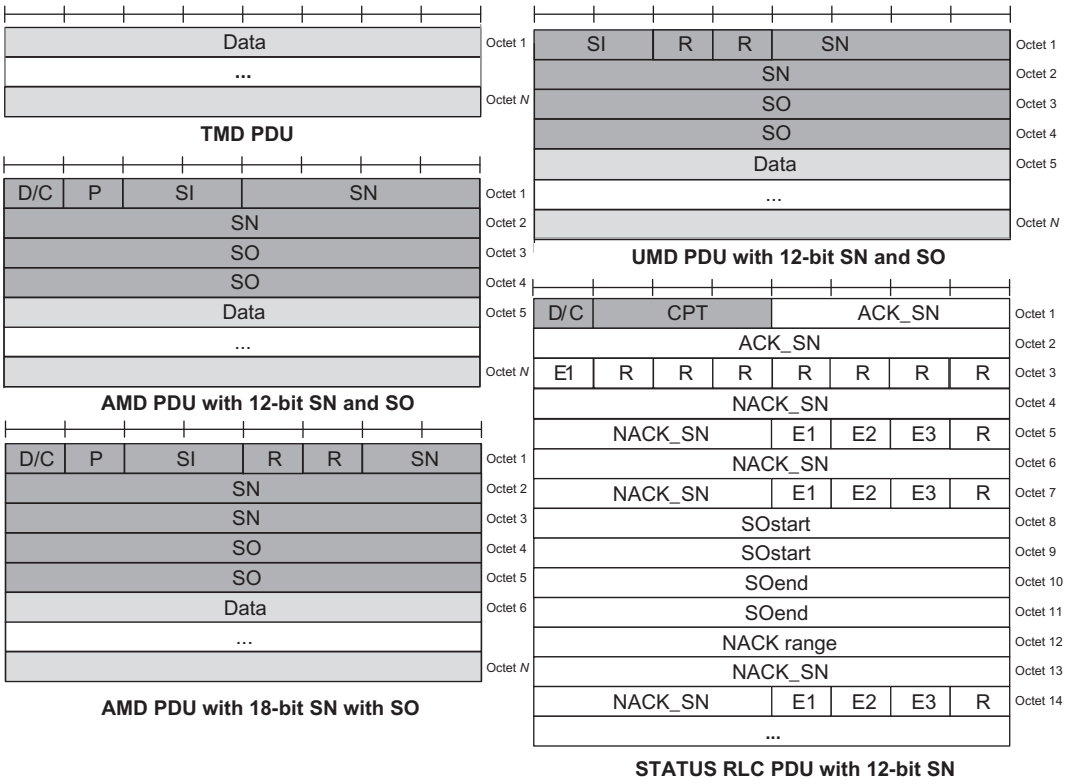


Figure 2.12
Example RLC PDU formats[10].

report for RLC-AM, and a data/control (D/C) indicator, indicating whether the RLC PDU contains data to/from a logical channel or control information required for RLC operation [10,17].

The retransmission of missing RLC PDUs is one of the main services of the RLC-AM entity. While MAC sublayer HARQ protocol provides some level of error correction, when combined with the ARQ retransmission mechanism, they can provide a higher level of reliability. The missing RLC PDUs can be detected by inspecting the sequence numbers of the received PDUs, and a retransmission requested from the transmitting side. The RLC-AM mode in NR is similar to LTE except the in-sequence delivery service is not supported in NR. As we mentioned earlier, eliminating the in-sequence delivery from the RLC helped reduce the overall latency. It further reduces the buffering requirements at the RLC sublayer. In LTE, which supports the in-sequence delivery service at RLC sublayer, an RLC SDU cannot be forwarded to higher layers unless all previous SDUs have been correctly received. A single missing SDU can significantly delay the delivery of the subsequent

SDUs. The RLC-AM is bidirectional which means the data may flow in both directions between the two peer entities (see Fig. 2.11), enabling the acknowledgement of the received RLC PDUs to the transmitting entity. The information about missing PDUs is provided by the receiving entity to the transmitting entity in the form of status reports. The status reports can either be transmitted autonomously by the receiver or requested by the transmitter. The PDUs in transit are tracked by the sequence number in the header. The transmitting and receiving RLC entities maintain two windows in RLC-AM, that is, the transmission and reception windows. The PDUs in the transmission window are only eligible for transmission, thus PDUs with sequence numbers less than the start of the window have already been acknowledged by the receiving RLC entity. In the same way, the receiving entity only accepts PDUs with sequence numbers within the reception window. The receiver discards any duplicate PDUs and delivers only one copy of each SDU to higher layers. The concept of RLC retransmissions is exemplified in Fig. 2.13, where transmitting and receiving RLC entities are illustrated. When operating in RLC-AM mode, each RLC entity has transmitting and receiving functionality; nevertheless, in this example we only show one direction as the other direction is identical [17].

In this example, the PDUs numbered k to $k+4$ are awaiting transmission in the transmission buffer. At time t_0, it is assumed that the PDUs with sequence number $SN \leq k$ have been transmitted and correctly received; however, only PDUs with sequence number $SN \leq k-1$ have been acknowledged by the receiver. The transmission window starts at k, that is, the first unacknowledged PDU, while the reception window starts at $k+1$, that is, the next PDU expected to be received. Upon reception of kth PDU, the SDU is reassembled and delivered to the higher layers. For a PDU containing an unsegmented SDU the reassembly function only involves header removal, but in the case of a segmented SDU, the SDU cannot be delivered to upper layers until the PDUs carrying all segments of the SDU arrive at the receiver. The transmission of PDUs continues such that at time t_1, PDUs $k+1$ and $k+2$ are transmitted but, at the receiving end, only PDU $k+2$ has arrived. As soon as a complete SDU is received, it is delivered to the higher layers; thus PDU $k+2$ is forwarded to the higher sublayer without waiting for the missing PDU $k+1$, which could be undergoing retransmission by the HARQ protocol. Therefore, the transmission window remains unchanged, since none of the PDUs with $SN \geq k$ have been acknowledged by the receiver. This could result in retransmission of these PDUs given that the transmitter is not aware of whether they have been correctly received. The reception window is not updated when PDU $k+2$ arrives because of the missing PDU $k+1$. At this point, the receiver starts the *t-Reassembly* timer. If the missing PDU $k+1$ is not received before the timer expires, a retransmission is requested. If the missing PDU arrives at time t_2 before the timer expires, the reception window is advanced, the reassembly timer is stopped, and PDU $k+1$ is delivered for reassembly with SDU $k+1$. The RLC sublayer is also responsible for duplicate detection using the same sequence number that is used for retransmission management.

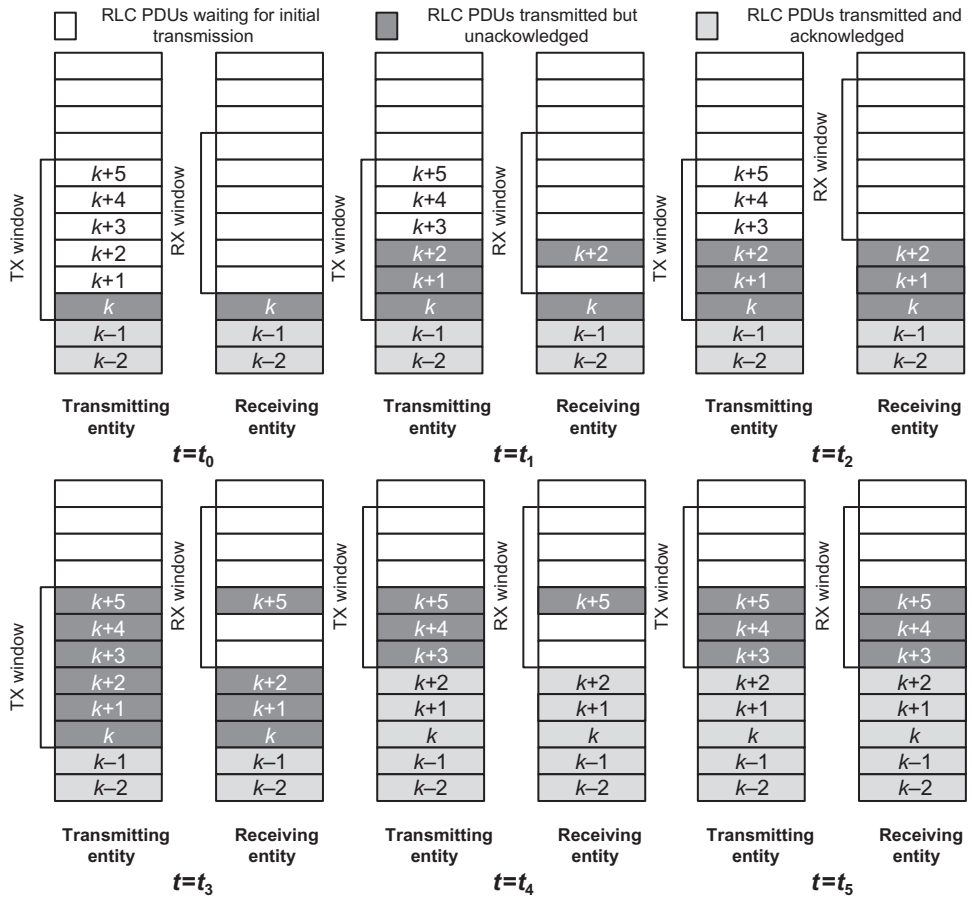


Figure 2.13
Example RLC-AM operation [17].

If PDU $k+2$ arrives again within the reception window, it will be discarded due to existence of another copy.

The transmission continues with PDUs $k+3$, $k+4$, and $k+5$, as shown in Fig. 2.13. At time t_3, PDUs with $SN \leq k+5$ have been transmitted, but only PDU $k+5$ has arrived and PDUs $k+3$ and $k+4$ are missing. Similar to the previous case, this causes the *t-Reassembly* timer to start. However, in this example no PDUs arrive prior to the expiration of the timer. The expiration of the timer at time t_4 triggers the receiver to send a control PDU containing a status report indicating the missing PDUs to its peer entity. The control PDUs have higher priority than data PDUs to avoid delayed transmission of the status reports, which would adversely impact the retransmission delay. Upon receipt of the status report at time t_5, the transmitter is informed that PDUs up to $k+2$ have been correctly

received; thus the transmission window is advanced. The missing PDUs $k + 3$ and $k + 4$ are retransmitted and are later correctly received. At time t_6, all PDUs, including the retransmissions, have been transmitted and successfully received. Since $k + 5$ was the last PDU in the transmission buffer the transmitter requests a status report from the receiver by setting a flag in the header of the last RLC data PDU. Upon reception of the PDU with the flag set, the receiver will respond by transmitting the requested status report, acknowledging all PDUs with $SN \leq k + 5$. The reception of the status report allows the transmitter to declare all PDUs as correctly received and to advance the transmission window. The status reports can be triggered for many reasons. However, to avoid transmission of excessive number of status reports, a *t-StatusProhibit* timer is used where status reports cannot be transmitted more than once within a time interval determined by the timer. In the above example, we assumed that each PDU carries an unsegmented SDU. Segmented SDUs are handled in the same way, but an SDU cannot be delivered to the higher layers until all segments have been received. The status reports and retransmissions operate on individual segments, and only the missing segments of a PDU are retransmitted [17].

2.2.3 Packet Data Convergence Protocol Sublayer

2.2.3.1 PDCP Services and Functions

The services and functions of the PDCP sublayer on the user-plane include sequence numbering; header compression and decompression; transfer of user data; reordering and duplicate detection; in-sequence delivery; PDCP PDU routing in multi-connectivity; retransmission of PDCP SDUs; ciphering, deciphering, and integrity protection; PDCP SDU discard; PDCP re-establishment and data recovery for RLC-AM; PDCP status reporting for RLC-AM; duplication of PDCP PDUs and duplicate discard indication to lower layers. The main services and functions of the PDCP sublayer on the control-plane consist of sequence numbering; ciphering, deciphering, and integrity protection; transfer of control-plane data; reordering and duplicate detection; in-sequence delivery; duplication of PDCP PDUs and duplicate discard indication to lower layers. The PDCP protocol performs (optional) IP-header compression, followed by ciphering, for each radio bearer. A PDCP header is added, carrying information required for deciphering in the other end, as well as a sequence number used for retransmission and in-sequence delivery [8].

In some services and applications such as VoIP, interactive gaming, and multimedia messaging, the data payload of the IP packet is almost the same size or even smaller than the header itself. Over the end-to-end connection comprising multiple hops, these protocol headers are extremely important, but over a single point-to-point link, these headers serve no useful purpose. It is possible to compress these headers, and thus save the bandwidth and use the expensive radio resources more efficiently. The header compression also provides other important benefits, such as reduction in packet loss and improved interactive response time. The payload header

compression is the process of suppressing the repetitive portion of payload headers at the sender side and restoring them at the receiver side of a low-bandwidth/capacity-limited link. The use of header compression has a well-established history in transport of IP-based payloads over capacity-limited wireless links where more bandwidth efficient transport methods are required. The Internet Engineering Task Force (IETF)[2] has developed several header compression protocols that are widely used in telecommunication systems. The header compression mechanism used in 3GPP LTE and NR standards is based on IETF RFC 5795, ROHC framework. The IP together with transport protocols such as TCP or UDP and application layer protocols (e.g., RTP) are described in the form of payload headers. The information carried in the header helps the applications to communicate over large distances connected by multiple links or hops in the network. This information consists of source and destination addresses, ports, protocol identifiers, sequence numbers, error checksums, etc. Under nominal conditions, most of the information carried in packet-headers remains the same or changes in specific patterns. By observing the fields that remain constant or change in specific patterns, it is possible either not to send them in each packet, or to represent them in a smaller number of bits than would have been originally required. This process is referred to as header compression.

The PDCP sublayer, at the transmit side, performs encryption of IP packets to protect user privacy, and additionally for the control-plane messages, performs integrity protection to ensure that control messages originate from the correct source and can be authenticated. At the receiver side the PDCP performs the corresponding decryption and decompression operation. The PDCP further discards duplicate PDUs and performs in-sequence delivery of the packets. Upon handover, the undelivered downlink PDUs will be forwarded by the PDCP sublayer of the source gNB to peer entity of the target gNB for delivery to the UE. The PDCP entity in the device will also handle retransmission of all uplink packets that were not delivered to the gNB given that the HARQ buffers are flushed upon handover. In this case, some PDUs may be received in duplicate form, possibly from both the source and the target gNBs. In this case, the PDCP will remove any duplicates. The PDCP entity can also be configured to perform reordering to ensure in-sequence delivery of SDUs to higher layer protocols. In some cases, the PDUs can be duplicated and transmitted on multiple cells, increasing the likelihood of their correct reception at the receiving end, which can be useful for services requiring very high reliability. At the receiving end, the PDCP duplicate removal functionality removes any duplicates. The PDCP plays an important role in dual connectivity, where a device is connected to two cells, that is, an MCG and an SCG. The two cell groups can be handled by different gNBs. A radio bearer is typically handled by one of the cell groups, but there is also the possibility for split bearers, in which case one

[2] IETF: https://www.ietf.org/.

radio bearer is handled by both cell groups. In this case the PDCP entity is responsible for distributing the data between the MCG and the SCG [17].

The PDCP entities are located in the PDCP sublayer. Several PDCP entities may be defined for a UE. Each PDCP entity carries the data of one radio bearer, which is associated with either the control-plane or the user-plane. Fig. 2.14 illustrates the functional block diagram of a PDCP entity. For split bearers, the routing function is performed in the transmitting PDCP entity. The PDCP sublayer provides services to RRC and SDAP sublayers. The PDCP provides transfer of user-plane and control-plane data; header compression; ciphering; integrity protection services to the higher layer protocols. The maximum size of a data or control PDCP SDU supported in NR is 9000 bytes [8]. It must be noted that an NR system provides protection against eavesdropping and modification attacks. Signaling traffic (RRC messages) is encrypted and integrity protected. User-plane traffic (IP packets) is encrypted and can be integrity protected. User-plane integrity protection is a new feature (relative to LTE) that is useful for small-data transmissions, and particularly

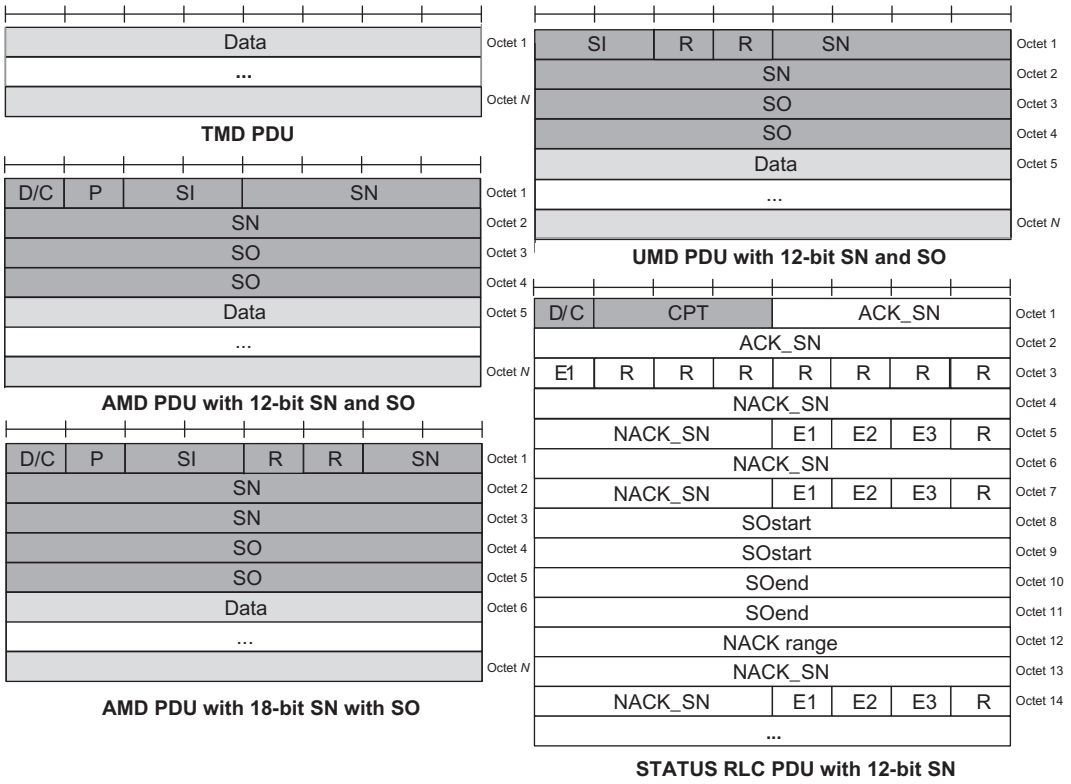


Figure 2.14
Functional block diagram of the PDCP entities [11].

for constrained IoT devices. Data traffic including voice calls, Internet traffic, and text messages are protected using encryption. The device and the network mutually authenticate each other and use integrity-protected signaling. This setup makes nearly impossible for an unauthorized party to decrypt and read the information that is communicated over the air. Although integrity protection of user-plane data is supported in 5G networks, this feature is not used in E-UTRA-NR-DC (EN-DC) scenarios since LTE systems only provide integrity protection of control-plane messages.

When duplication is configured for a radio bearer by RRC, a secondary RLC entity is added to the radio bearer to handle the duplicated PDCP PDUs, where the logical channels corresponding to the primary and the secondary RLC entities are referred to as the *primary logical channel* and the *secondary logical channel*, respectively. Therefore, the duplication function at PDCP sublayer consists of submitting the same PDCP PDUs twice, that is, to the primary RLC entity and to the secondary RLC entity. With two independent transmission paths, packet duplication would improve the transmission reliability and reduce the latency, which is especially advantageous to URLLC services [8]. The PDCP control PDUs are not duplicated and are always submitted to the primary RLC entity. When configuring duplication for a DRB, the RRC sublayer sets the initial state to be either activated or deactivated. After the configuration, the state can be dynamically controlled by means of a MAC CE. In dual connectivity, the UE applies the MAC CE commands regardless of their origin (MCG or SCG). When duplication is configured for an SRB, the state is always active and cannot be dynamically controlled. When activating duplication for a DRB, NG-RAN ensures that at least one serving cell is activated for each logical channel of the DRB. When the deactivation of SCells leaves no activated serving cells for the logical channels of the DRB, NG-RAN ensures that duplication is also deactivated [8].

When duplication is activated, the original PDCP PDU and the corresponding duplicate are not transmitted on the same carrier. The primary and secondary logical channels can either belong to the same MAC entity (referred to as CA duplication) or to different ones (referred to as DC duplication). In CA duplication, logical channel mapping restrictions are applied to the MAC sublayer to ensure that the primary and secondary logical channels are not sent on the same carrier. When duplication is deactivated for a DRB, the secondary RLC entity is not reestablished, the HARQ buffers are not flushed, and the transmitting PDCP entity should indicate to the secondary RLC entity to discard all duplicated PDCP PDUs. In addition, in case of CA duplication, the logical channel mapping restrictions of the primary and secondary logical channels are relaxed as long as duplication remains deactivated. When an RLC entity acknowledges the transmission of a PDCP PDU, the PDCP entity notifies the other RLC entity to discard that PDU. When the secondary RLC entity

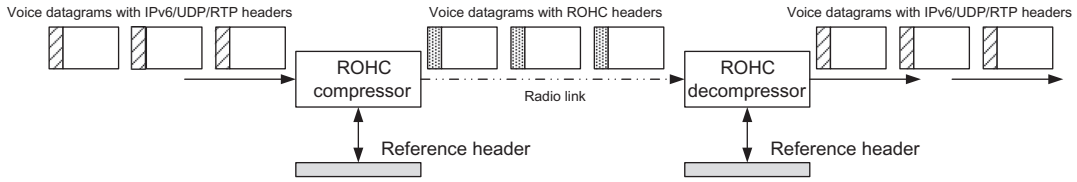


Figure 2.15
Example ROHC compression/decompression of RTP/UDP/IP headers for communication over a radio link [16].

reaches the maximum number of retransmissions for a PDCP PDU, the UE informs the gNB, but does not trigger RLF[3] [8].

2.2.3.2 Header Compression Function

There are multiple header compression algorithms referred to as profiles in 3GPP specifications [11]. Each profile is specific to a particular network layer, transport layer, or upper layer protocol combination, for example, TCP/IP or RTP/UDP/IP. The PDCP entities associated with DRBs carrying user-plane data can be configured by upper layers to use header compression. Each PDCP entity uses at most one ROHC compressor instance and at most one ROHC decompressor instance.

As we mentioned earlier, the ROHC algorithm reduces the size of transmitted RTP/UDP/IP header by removing the redundancies. This mechanism starts by classifying header fields into different classes according to their variation pattern. The fields that are classified as *inferred* are not sent. The *static* fields are sent initially and then are not sent anymore and the fields with varying information are always sent. The ROHC mechanism is based on a context,[4] which is maintained, by both ends, that is, the compressor and the decompressor (see Fig. 2.15). The context encompasses the entire header and ROHC information. Each context has a context ID, which identifies the flows. The ROHC scheme operates in one of the following three operation modes [16]:

[3] The UE declares a RLF when one of the following criteria is met: (1) Expiry of a timer started after indication of radio problems from the physical layer (if radio problems are recovered before the timer expires, the UE stops the timer), random-access procedure failure, or RLC failure. After an RLF is declared, the UE stays in RRC_CONNECTED and selects a suitable cell and then initiates RRC connection reestablishment. The UE further enters RRC_IDLE, if a suitable cell cannot be found within a certain time after RLF was declared [8].

[4] The context of the compressor is the state it uses to compress a header. The context of the decompressor is the state it uses to decompress a header. Either of these or the combinations of these two is usually referred to as "context." The context contains relevant information from previous headers in the packet stream, such as static fields and possible reference values for compression and decompression. Moreover, additional information describing the packet stream is also part of the context, for example, information about how the IP identifier field changes and the typical interpacket increase in sequence numbers or timestamps [16].

1. *Unidirectional mode (U)*, where the packets are only sent in one direction from compressor to decompressor. This mode makes ROHC usable over links where a return path from decompressor to compressor is unavailable or undesirable.
2. *Optimistic mode (O)* is a bidirectional mode similar to the unidirectional mode, except that a feedback channel is used to send error recovery requests and (optionally) acknowledgements of significant context updates from the decompressor to compressor. The O-mode aims to maximize compression efficiency and sparse usage of the feedback channel.
3. *Reliable mode (R)* is a bidirectional mode which differs in many ways from the previous two modes. The most important differences include intensive use of feedback channel and a strict logic at both compressor and decompressor that prevents loss of context synchronization between compressor and decompressor except for very high residual bit error rates.

The U-mode is used when the link is unidirectional or when feedback is not possible. For bidirectional links, O-mode uses positive feedback packets (ACK) and R-mode use positive and negative feedback packets [ACK and NACK]. The ROHC mechanism always starts header compression using U-mode even if it is used over a bidirectional link and it does not send retransmissions when an error occurs; thus the erroneous packet is dropped. The ROHC feedback is used only to indicate to the compressor side that there was an error and probably the context is damaged. After receiving a negative feedback the compressor always reduces its compression level.

The ROHC compressor has three compression states defined as follows [16]:

1. *Initialization and refresh (IR)*, where the compressor has been just created or reset and full packet-headers are sent.
2. *First order (FO)*, where the compressor has detected and stored the static fields such as IP addresses and port numbers on both sides of the connection.
3. *Second order (SO)*, where the compressor is suppressing all dynamic fields such as RTP sequence numbers and sending only a logical sequence number and partial checksum to make the other side generate the headers based on prediction and verify the headers of the next expected packet.

Each compression state uses a different header format in order to send the header information. The IR compression state establishes the context, which contains static and dynamic header information. The FO compression state provides the change pattern of dynamic fields. The SO compression state sends encoded values of *sequence number* (SN) and *timestamp* (TS), forming the minimal size packets (Figs. 2.6–2.15). Using this header format, all header fields can be generated at the other end of the radio link using the previously established change pattern. When some updates or errors occur, the compressor returns to

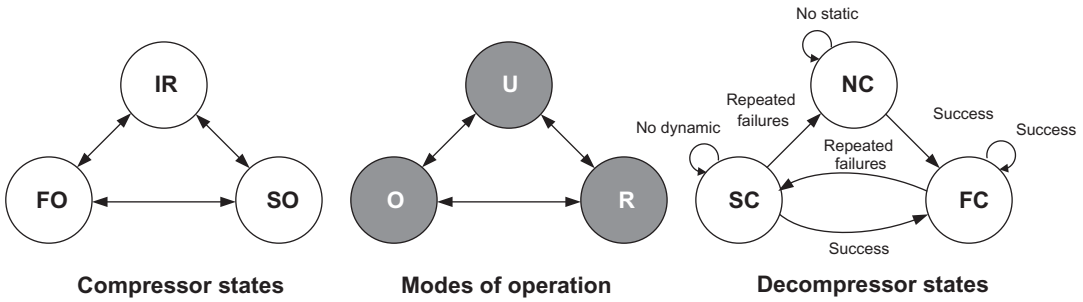


Figure 2.16
ROHC state machines [16].

upper compression states. It only transitions to the SO compression state after retransmitting the updated information and reestablishing the change pattern in the decompressor.

In the U-mode the feedback channel is not used. To increase the compression level an optimistic approach is used for compressor to ensure that the context has been correctly established at decompressor side. This means that compressor uses the same header format for a number of packets. Since the compressor does not know whether the context is lost, it also uses two timers, to be able to return to the FO and IR compression states. The decompressor works at the receiving end of the link and decompresses the headers based on the header fields' information of the context. Both the compressor and the decompressor use a context to store all the information about the header fields. To ensure correct decompression the context should be always synchronized. The decompressor has three states as follows: (1) no context (NC) where there is NC synchronization, (2) static context (SC) where the dynamic information of the context has been lost, and (3) full context (FC) when the decompressor has all the information about header fields. In FC state the decompressor transitions to the initial states as soon as it detects corruption of the context. The decompressor uses the "*k* out of *n*" rule by looking at the last *n* packets with CRC failures. If *k* CRC failures have occurred, it assumes the context has been corrupted and transitions to an initial state (SC or NC). The decompressor also sends feedback according to the operation mode (Fig. 2.16).

The values of the ROHC compression parameters that determine the efficiency and robustness are not defined in ROHC specification and are not negotiated initially but are stated as implementation dependent. The values of these parameters remain fixed during the compression process. The compression parameters are defined as follows [16]:

- *L*: In U-mode and O-mode, the ROHC compressor uses a confidence variable *L* in order to ensure the correct transmission of header information.
- *Timer_1* (*IR_TIMEOUT*): In U-mode, the compressor uses this timer to return to the IR compression level and periodically resends static information.

- *Timer_2* (*FO_TIMEOUT*): The compressor also uses another timer in U-mode, and this timer is used to go downward to FO compression level, if the compressor is working in SO compression level.
- *Sliding window width (SWW)*: The compressor, while compressing header fields such as *sequence number* (SN) and *TS*, utilizes window-based least significant bit (*W_LSB)* encoding that uses a sliding window of width equal to *SWW*.
- *W_LSB* encoding is used to compress those header fields whose change pattern is known. When using this encoding, the compressor sends only the least significant bits. The decompressor uses these bits to construct the original value of the encoding fields.
- *k* and *n*: The ROHC decompressor uses a "*k* out of *n*" failure rule, where *k* is the number of packets received with an error in the last *n* transmitted packets. This rule is used in the state machine of the decompressor to assume the damage of context and move downwards to a state after sending a NACK to the compressor, if bidirectional link is used. The decompressor does not assume context corruption and remains in the current state until *k* packets arrive with error in the last *n* packets.

2.2.3.3 Ciphering and Integrity Protection Functions

The RRC confidentiality protection is provided by the PDCP sublayer between a UE and the serving gNB. The user-plane security policy indicates whether the user-plane confidentiality and/or user-plane integrity protection is activated for all DRBs belonging to the PDU session. The input parameters to the 128-bit NR encryption algorithm (NEA) (or alternatively encryption algorithm for 5G), which is used for ciphering, are 128-bit cipher key referred to as KEY (K_{RRCenc}), 32-bit COUNT (PDCP COUNT), 5-bit radio bearer identity BEARER, 1-bit direction of the transmission, that is, DIRECTION, and the length of the keystream required identified as LENGTH. The DIRECTION bit is set to zero for uplink and one for downlink. Fig. 2.17 illustrates the use of the ciphering algorithm NEA to encrypt plain text by applying a keystream using a bit-wise binary addition of the *plaintext* block and the *keystream* block. The *plaintext* block may be recovered by generating the same *keystream* block using the same input parameters and applying a bit-wise binary addition with the *ciphertext* block. Based on the input parameters, the algorithm generates the output keystream block *keystream* which is used to encrypt the input *plaintext* block to produce the output *ciphertext* block. The input parameter LENGTH only denotes the length of the *keystream* block and not its content [2,11]. The ciphering algorithm and key to be used by the PDCP entity are configured by upper layers and the ciphering method is applied according to the security architecture of 3GPP system architecture evolution (SAE, which is the LTE system architecture). The ciphering function is activated by upper layers. After security activation, the ciphering function is applied to all PDCP PDUs indicated by upper layers for downlink and uplink transmissions. The COUNT value is composed of a hyperframe number (HFN) and the PDCP SN. The size of the HFN part in bits is equal to 32

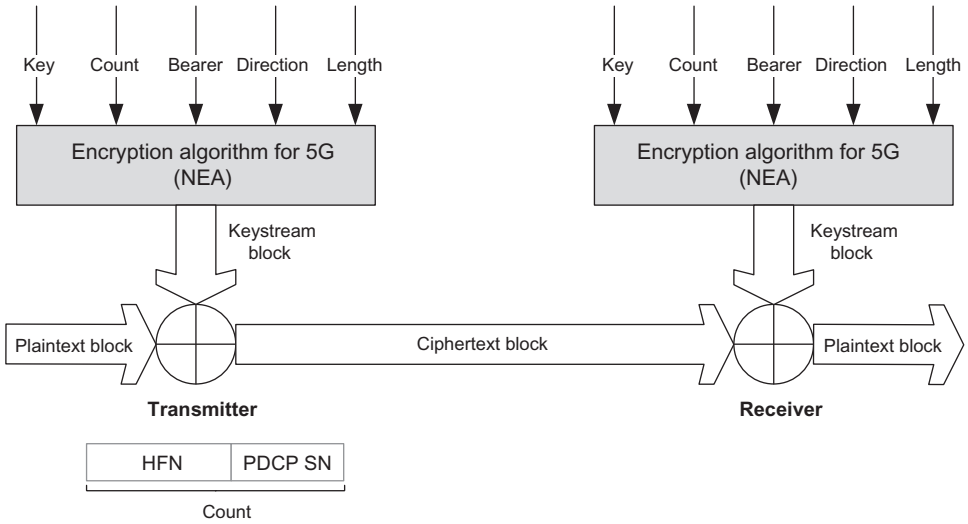


Figure 2.17
Illustration of the ciphering and deciphering procedures [2,11].

minus the length of the PDCP SN. The PDCP does not allow COUNT to wrap around in the downlink and uplink; thus it is up to the network to prevent it from happening [2,8,11].

As shown in Fig. 2.18, the input parameters to the 128-bit NR integrity protection algorithm (NIA) (or alternatively integrity protection algorithm for 5G) are the RRC messages denoted as MESSAGE, 128-bit integrity key K_{RRCint} referred to as KEY, 5-bit bearer identity BEARER, 1-bit direction of transmission denoted as DIRECTION, and a bearer specific direction-dependent 32-bit input COUNT which corresponds to the 32-bit PDCP COUNT. The RRC integrity checks are performed both in the UE and the gNB. If the gNB or the UE receives a PDCP PDU which fails the integrity check with faulty or missing message authentication code (MAC-I) after the start of integrity protection, the PDU will be discarded. The DIRECTION bit set to zero for uplink and one for downlink. The bit length of the MESSAGE is LENGTH. Based on these input parameters, the sender computes a 32-bit message authentication code (MAC-I/NAS-MAC)[5] using the integrity protection algorithm

[5] In cryptography a MAC-I is a cryptographic checksum on data that uses a session key to detect both accidental and intentional modifications of the data. A MAC (not to be confused with medium access control MAC) requires two inputs: a message and a secret key known only to the originator of the message and its intended recipient(s). This allows the recipient of the message to verify the integrity of the message and authenticate that the message sender has the shared secret key. Any mismatch between the sender's and receiver's calculated MAC-I values would invalidate the message. There are four types of message authentication codes: unconditionally secure, hash function-based, stream cipher-based, and block cipher-based. In the past, the most common approach to creating a message authentication code was to use block ciphers; however, hash-based MACs which use a secret key in conjunction with a cryptographic hash function to produce a hash, have become more widely used.

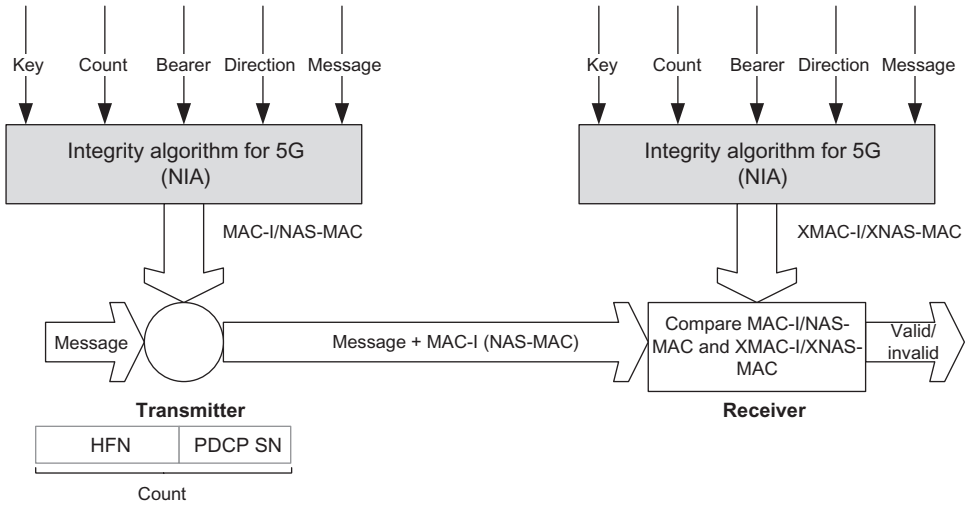


Figure 2.18
Integrity protection and verification procedures [2,11].

NIA. The message authentication code is then appended to the message when sent. For integrity protection algorithms, the receiver computes the expected message authentication code (XMAC-I/XNAS-MAC) on the message received in the same way that the sender computed its message authentication code on the message sent and verifies the integrity of the message by comparing it to the received message authentication code, that is, MAC-I/NAS-MAC. The integrity protection algorithm and key to be used by the PDCP entity are configured by upper layers and the integrity protection method is applied according to security architecture of 3GPP SAE [2]. The integrity protection function is activated by upper layers. Following the security activation, the integrity protection function is applied to all PDUs including and subsequent to the PDU indicated by upper layers for downlink and uplink transmissions. As the RRC message which activates the integrity protection function is itself integrity protected with the configuration included in that RRC message, the message must be decoded by RRC before the integrity protection verification can be performed for the PDU in which the message was received. The parameters that are required by PDCP for integrity protection are defined in reference [11] and are input to the integrity protection algorithm.

The PDCP data PDU is used to convey user-plane and control-plane data, as well as MAC-I in addition to the PDU header. The PDCP control PDU is used to transport PDCP status report and/or interspersed ROHC feedback in addition to the PDU header. A PDCP SDUs and PDUs are octet-aligned bit strings. A compressed or uncompressed SDU is included in a PDCP data PDU. Fig. 2.19 shows the format of the PDCP data PDU with 12-bit SN, which is applicable to SRBs. The figure further shows the format of the PDCP data PDU

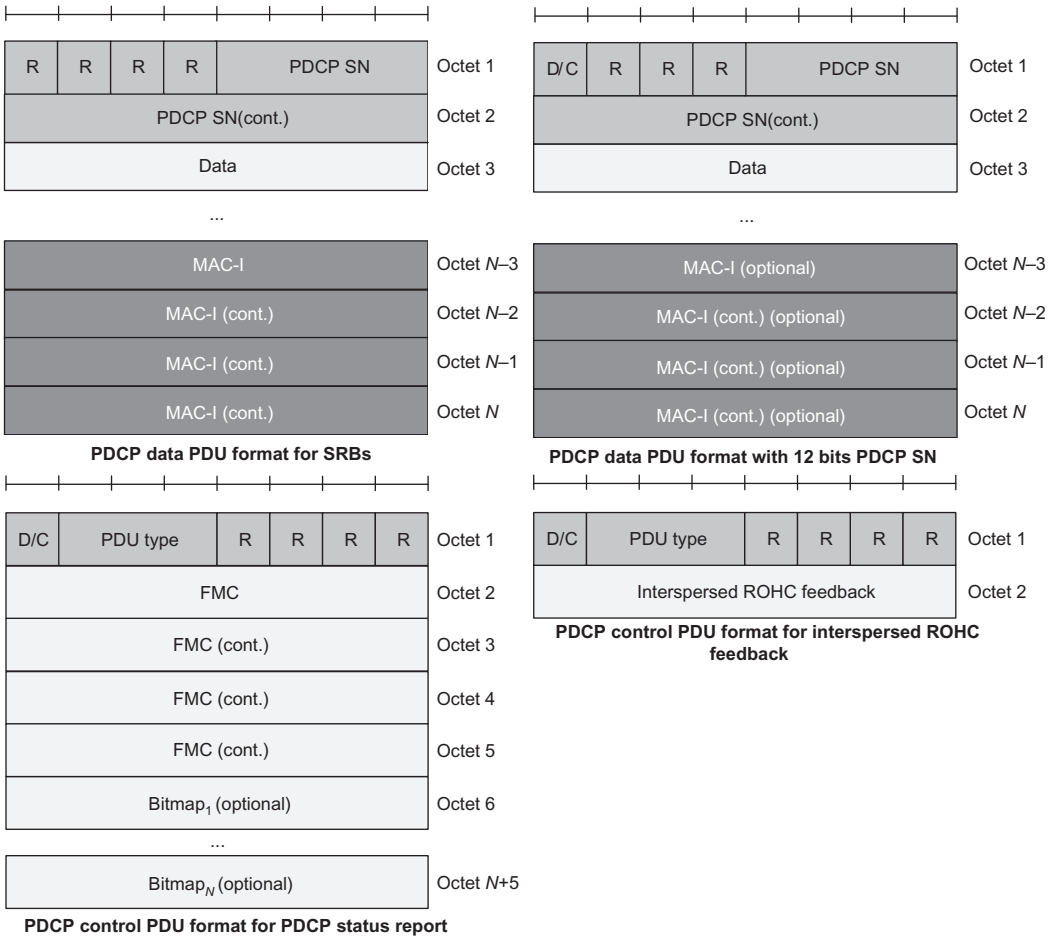


Figure 2.19
PDCP PDU formats [11].

with 12-bit SN for RLC-UM and RLC-AM DRBs. The structure of PDCP control PDU carrying one PDCP status report, which is applicable to RLC-AM DRBs, as well as the structure of PDCP control PDU transporting one interspersed ROHC feedback, which is applicable to RLC-UM and AM DRBs are shown in Fig. 2.19.

In PDCP PDU formats, the sequence number (SN) is a 12- or 18-bit number which is configured by RRC. The data field is a variable-size field which includes uncompressed user-plane/control-plane data or compressed user-plane data. As we stated earlier, the header compression only applies to user-plane data. The MAC-I field carries a message authentication code. For SRBs the MAC-I field is always present; however, if integrity protection is not configured, the MAC-I field is still present in PDCP PDU but is padded with zeros.

For DRBs, the MAC-I field is present only when the DRB is configured with integrity protection, which is unique to NR. The D/C field indicates whether the corresponding PDCP PDU is a PDCP data PDU or a PDCP control PDU. The PDU Type identifies the type of control information included in the corresponding PDCP control PDU, which can be a status report, interspersed ROHC feedback, or reserved. The first missing COUNT indicates the COUNT value of the first missing PDCP SDU within the reordering window. The Bitmap field indicates which SDUs are missing and which SDUs have been correctly received in the receiving entity. The interspersed ROHC feedback has a variable length and contains one ROHC packet with only feedback, that is, a ROHC packet which is not associated with a PDCP SDU. When an interspersed ROHC feedback is generated by the header compression protocol, the transmitting PDCP sends the corresponding PDCP control PDU to the lower layers without associating a PDCP SN or performing ciphering. The receiving PDCP entity delivers the corresponding interspersed ROHC feedback to the header compression protocol without performing deciphering [11].

2.2.4 Service Data Adaptation Protocol Sublayer

The main services and functions of SDAP sublayer include mapping between a QoS flow and a DRB and marking QFI in downlink and uplink IP packets. A single-protocol entity of SDAP is configured for each individual PDU session. The SDAP sublayer was introduced in NR because of the new QoS framework compared to LTE QoS management when connected to the 5G core. However, if the gNB is connected to the EPC, which is the case for non-standalone deployments, the SDAP service/functionality is not used. As we mentioned earlier, the NG-RAN architecture supports disaggregated gNB where gNB functions are split into a central unit (gNB-CU) and one or more distributed units (gNB-DU) connected via F1 interface. In the case of a split gNB, the RRC, PDCP, and SDAP protocols, described in more detail below, reside in the gNB-CU and the remaining protocol entities (RLC, MAC, and PHY) will be located in the gNB-DU. The interface between the gNB (or the gNB-DU) and the device is denoted as the Uu interface. In the example shown in Fig. 2.4, the SDAP protocol maps the IP packets to different radio bearers, that is, IP packets n and $n + 1$ are mapped to radio bearer x and IP packet m is mapped to radio bearer y. The SDAP mapping function between a QoS flow and a DRB is due to the new QoS framework which is used in the new radio. The SDAP further marks the QFIs in the downlink due to the use of reflective QoS[6] and in the uplink due to the use of new QoS framework. A single SDAP entity (as shown in Fig. 2.20) is configured for each individual PDU session, except for the dual connectivity scenario where two entities can be configured.

[6] Reflective QoS flow to DRB mapping is a QoS flow to DRB mapping scheme where a UE monitors the QoS flow to DRB mapping rule in the downlink and applies it to in the uplink [3].

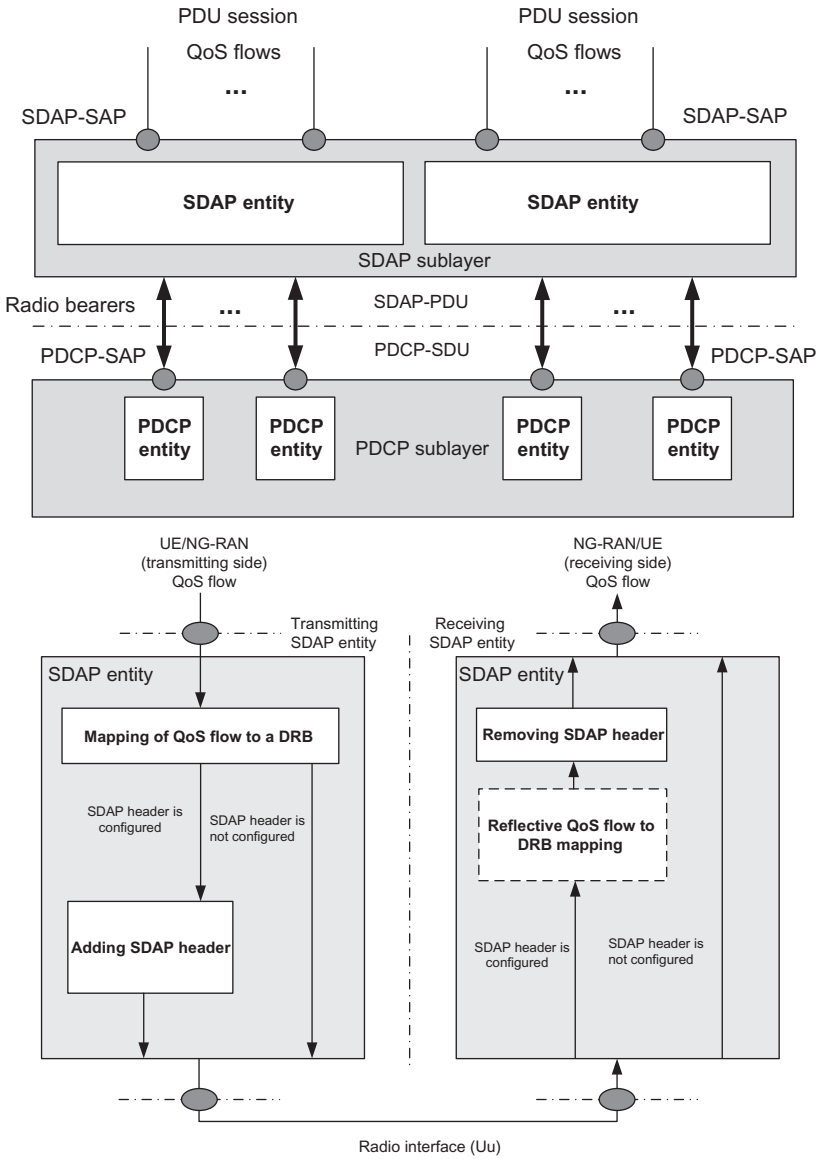


Figure 2.20
High-level SDAP sublayer functional architecture [3].

Fig. 2.20 illustrates one possible structure for the SDAP sublayer; however, the actual implementations may vary. The SDAP sublayer is configured by RRC. It maps the QoS flows to DRBs. One or more QoS flows may be mapped onto one DRB. However, in the uplink, one QoS flow is mapped to only one DRB at a time. The SDAP entities are located in the SDAP sublayer. Several SDAP entities may be defined for a UE. One SDAP entity is

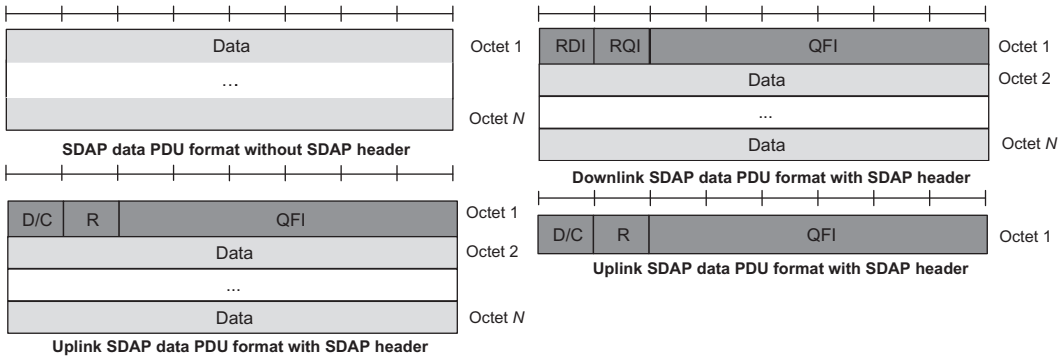


Figure 2.21
SDAP PDU formats [3].

configured for each individual PDU session. An SDAP entity receives/delivers SDAP SDUs from/to upper layers and transmits/receives SDAP data PDUs to/from its peer SDAP entity via lower layers. In the transmitting side, when an SDAP entity receives an SDAP SDU from upper layers, it constructs the corresponding SDAP data PDU and submits it to lower layers. In the receiving side, when an SDAP entity receives an SDAP data PDU from lower layers, it retrieves the corresponding SDAP SDU and delivers it to upper layers. Fig. 2.20 illustrates the functional block diagram of the SDAP entity for the SDAP sublayer [3]. Note that the reflective QoS flow to DRB mapping is performed at UE, if downlink SDAP header is configured. The SDAP sublayer transfers user-plane data and exclusively provides services to the user-plane upper layers. As shown in Fig. 2.20, the SDAP sublayer supports transfer of user-plane data, mapping between a QoS flow and a DRB for both downlink and uplink; marking QFI in both downlink and uplink packets; and reflective QoS flow to DRB mapping for the uplink SDAP data PDUs.

The SDAP data PDU is used to convey SDAP header and user-plane data. An SDAP PDU or SDU is a bit string that is octet-aligned (Fig. 2.21). An SDAP SDU is included in an SDAP PDU. An SDAP data PDU may only consist of a data field with no SDAP header. As shown in Fig. 2.21, the SDAP data PDU in the downlink or uplink consists of a header and data fields, where the headers for the downlink and uplink are different. For each downlink SDAP data PDU, in which reflective QoS indication (RQI) is set to one, the SDAP entity would inform the NAS layer of the RQI and QFI. The end-marker control PDU is used by the SDAP entity at UE to inform the gNB that the SDAP SDU QoS flow mapping to the DRB on which the end-marker PDU is transmitted, has stopped. The D/C bit specifies whether the SDAP PDU is an SDAP data PDU or an SDAP control PDU. The QFI field identifies the QFI to which the SDAP PDU belongs. The RQI bit indicates whether NAS should be informed of the updated SDF to QoS flow mapping rules and the reflective QoS flow to DRB mapping indication bit implies whether QoS flow to DRB mapping rule should be updated [3].

2.3 Layer 3 Functions and Services

2.3.1 Radio Resource Control Sublayer

The RRC sublayer consists of control-plane set of protocols for connection control and setup, system configuration, mobility management, and security establishment. It is further responsible for broadcast of SI including NAS common control information and information applicable to UEs in RRC_IDLE and RRC_INACTIVE states, for example, cell selection/reselection parameters, neighboring cell information, as well as information applicable to UEs in RRC_CONNECTED state, for example, common channel configuration information [14]. The RRC connection control functions include paging; establishment/modification/suspension/resumption/release of RRC connections including assignment/modification of UE identity; establishment/modification/suspension/resumption/release of SRBs except SRB0; access barring; initial security activation including initial configuration of AS integrity protection (SRBs, DRBs), and AS ciphering (SRBs, DRBs); mobility management including intra-frequency and inter-frequency handover, associated security handling such as key or algorithm change, specification of RRC context information transferred between network nodes; establishment/modification/suspension/resumption/release of radio bearers carrying user data (DRBs); radio configuration control including assignment/modification of ARQ configuration, HARQ configuration, and DRX configuration.

In case of dual connectivity, the RRC sublayer provides cell management functions including change of primary second cell (PSCell), addition/modification/release of SCG cell(s). In case of carrier aggregation, RRC sublayer provides cell management functions including addition/modification/release of SCell(s). The RRC sublayer further provides QoS control including assignment/modification of SPS configuration and DL/UL configured grant configuration, assignment/modification of parameters for uplink rate-control in the UE, that is, allocation of a priority and a prioritized bit rate for each resource block. The RRC sublayer also handles recovery from RLF condition; inter-RAT mobility including security activation, transfer of RRC context information; measurement configuration and reporting which includes establishment/modification/release of measurement configuration (e.g., intra-frequency, inter-frequency, and inter-RAT measurements); setup and release of measurement gaps; and measurement reporting [14].

The RRC messages are sent to the UEs using SRBs, based on the same set of protocol layers that are used for user-plane packet processing except the SDAP sublayer. The SRBs are mapped to the CCCH during connection setup and to the DCCH once the connection is established. The control-plane and user-plane data can be multiplexed in the MAC sublayer and transmitted to the device within the same TTI. The MAC CEs can be used for control of radio resources in some specific cases where low latency is more important than ciphering, integrity protection, and reliable transport of data.

Table 2.1: Radio resource control (RRC) functions in standalone and non-standalone NR operation.

Services	Functions	Non-standalone Architecture	Standalone Architecture	Differences With LTE RRC
System information	Broadcast of minimum system information	√	√	–
	Broadcast of other system information		√	Introduction of on-demand and area provision
Connection control	Bearer and cell settings	√	√	Introduction of split SRB[a] and direct SRB[b]
	Connection establishment with the core network		√	Introduction of RRC_INACTIVE state
	Paging		√	Introduction of RAN level paging
	Access control		√	Introduction of unified access control
Mobility	Handover		√	–
	Cell selection/reselection		√	–
Measurement	Downlink quality measurements/reporting	√	√	Introduction of beam measurements
	Cell identifier measurement/reporting	√	√	–

[a]Split SRB is a bearer for duplicating RRC messages generated by the master node for terminals in dual connectivity scenarios and transmitting via the secondary node.
[b]Direct SRB is a bearer whereby the secondary node can send RRC messages directly to the terminals in dual connectivity scenarios.

Table 2.1 shows the functional classification of NR RRC, the relevance of each function to standalone and non-standalone operation, and the similarities and differences with the LTE RRC functions.

SRBs are defined as radio bearers that are used only for the transmission of RRC and NAS messages. More specifically, the new radio has specified four types of SRBs which includes SRB0 for RRC messages using the CCCH logical channel; SRB1 for RRC messages which may include a piggybacked NAS message, as well as for NAS messages prior to the establishment of SRB2 using DCCH logical channel; SRB2 for NAS messages using DCCH logical channel (SRB2 has a lower priority than SRB1 and may be configured by the network after security activation); and SRB3 for specific RRC messages when UE is in EN-DC mode using DCCH logical channel. In the downlink, piggybacking of NAS messages is used only for bearer establishment/modification/release. In the uplink, piggybacking of NAS messages is used only for transferring the initial NAS messages during connection setup and connection resume. The NAS messages transferred via SRB2 are also contained

in RRC messages, which do not carry any RRC protocol control information. Once security is activated, all RRC messages on SRB1, SRB2, and SRB3, including those containing NAS messages, are integrity protected and ciphered by PDCP sublayer. The NAS independently applies integrity protection and ciphering to the NAS messages [14].

2.3.2 System Information

The system information consists of an MIB and a number of SIBs, which are divided into minimum SI and other SI. The minimum SI comprises basic information required by the UEs for initial access and acquiring any other SI. The minimum SI itself consists of MIB which contains cell barred status information and essential physical layer information of the cell required for the UEs to receive further SI, for example, CORESET#0 configuration. The MIB is periodically broadcast on BCH. The minimum SI further includes SIB1 which defines the scheduling of other SIBs and contains information required for initial access. The SIB1 is also referred to as remaining minimum system information (RMSI) and is periodically broadcast on DL-SCH or sent in a dedicated manner on DL-SCH to UEs in RRC_CONNECTED state. The other SI encompasses all SIBs that are not broadcast as part of minimum SI. Those SIBs can either be periodically broadcast on DL-SCH, broadcast on-demand on DL-SCH, that is, upon request from the UEs in RRC_IDLE or RRC_INACTIVE state or sent in a dedicated manner on DL-SCH to the UEs in RRC_CONNECTED state. The other SI is divided into the following SIBs [8]:

- SIB2 contains cell reselection information related to the serving cell.
- SIB3 contains information about the serving frequency and intra-frequency neighbor cells relevant for cell reselection, including cell reselection parameters common for a frequency as well as cell-specific reselection parameters.
- SIB4 contains information about other NR frequencies and inter-frequency neighbor cells relevant for cell reselection including cell reselection parameters common for a frequency as well as cell-specific reselection parameters.
- SIB5 contains information about E-UTRA frequencies and E-UTRA neighbor cells relevant for cell reselection, including cell reselection parameters common for a frequency as well as cell-specific reselection parameters.
- SIB6 contains an Earthquake and Tsunami Warning System (ETWS)[7] primary notification.
- SIB7 contains an ETWS secondary notification.

[7] ETWS is a public warning system developed to satisfy the regulatory requirements for warning notifications related to earthquake and/or tsunami events. The ETWS warning notifications can either be a primary notification (short notification) or secondary notification (providing detailed information) [8].

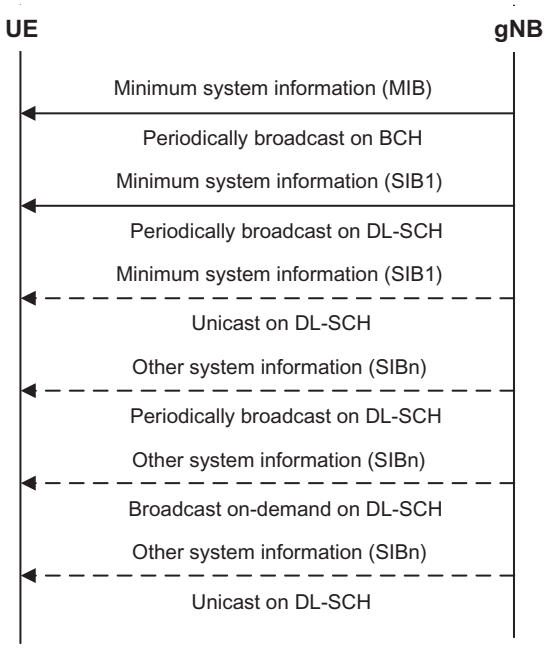


Figure 2.22
System information provisioning [8].

- SIB8 contains a Commercial Mobile Alert System (CMAS)[8] warning notification.
- SIB9 contains information related to Global Positioning System (GPS) time and coordinated universal time.

Fig. 2.22 summarizes SI provisioning. For a cell/frequency that is considered for camping by the UE, the UE is not required to acquire the contents of the minimum SI of that cell/frequency from another cell/frequency layer. This does not preclude the case where the UE applies stored SI from previously visited cell(s). The UE would consider a cell as barred, if it cannot determine the full content of the minimum SI broadcast by that cell. The UE only acquires SI on the active BWP, when using bandwidth adaptation.

The MIB is mapped to BCCH and is exclusively carried on BCH; however, other SI messages are mapped to BCCH and are dynamically carried on DL-SCH. The scheduling of SI messages is part of other SI and is signaled via SIB1. The UEs in RRC_IDLE or RRC_INACTIVE state may request other SI which would trigger a random-access procedure, wherein the corresponding Msg3 includes the SI request message unless the requested SI is associated with a subset of physical RACH (PRACH) resources, in that case Msg1 is

[8] CMAS is a public warning system developed for the delivery of multiple, concurrent warning notifications [8].

used. When Msg1 is used, the minimum granularity of the request is one SI message (i.e., a set of SIBs), one RACH preamble and/or PRACH resource can be used to request multiple SI messages and the gNB acknowledges the request in Msg2. When Msg3 is used, the gNB acknowledges the request in Msg4. The other SI may be broadcast at a configurable periodicity for a certain duration of time. The other SI may also be broadcast when it is requested by a UE in RRC_IDLE or RRC_INACTIVE state [8]. A UE would be allowed to camp on a cell, if it acquires the minimum SI broadcast by that cell. It must be noted that not all cells in a network broadcast the minimum SI; thus the UE would not be able to camp on those cells.

The SI may be changed at the specific radio frames according to a modification period. The SI may be transmitted a number of times with the same content within the modification period defined by its scheduling. The modification period is configured by the SI. When the network parameters change (some of the SI), it first notifies the UEs about this change, that is, this may be done within a modification period. In the next modification period, the network transmits the updated SI. Upon receiving a change notification the UE acquires the new SI from the beginning of the next modification period. The UE applies the previously acquired SI until the UE acquires the new SI. The short message transmitted with P-RNTI via DCI on PDCCH is used to inform UEs in RRC_IDLE, RRC_INACTIVE, or RRC_CONNECTED state about an SI change. If the UE receives a short message with SI change indication, it means that the SI will change at the next modification period boundary [8].

As we mentioned earlier, the SI is divided into the MIB and a number of SIBs. The MIB is always transmitted on the BCH with a periodicity of 80 ms and is repeated within the 80 ms. The MIB includes parameters that are needed to acquire SIB1 from the cell.

The SIB1 is transmitted on the DL-SCH with a periodicity of 160 ms and variable transmission repetition periodicity within 160 ms. The default repetition period of SIB1 is 20 ms; however, the actual repetition periodicity is up to network implementation. For synchronization signal/PBCH block (SSB) and CORESET multiplexing pattern 1, SIB1 repetition transmission period is 20 ms. For SSB and CORESET multiplexing pattern 2/3, SIB1 repetition period is the same as the SSB period. The SIB1 includes information regarding the availability and scheduling (e.g., mapping of SIBs to SI message, periodicity, and SI-window size) of other SIBs with an indication whether the SIBs are only provided on-demand, and in that case the configuration needed by the UE to perform the SI request. The SIB1 is cell-specific.

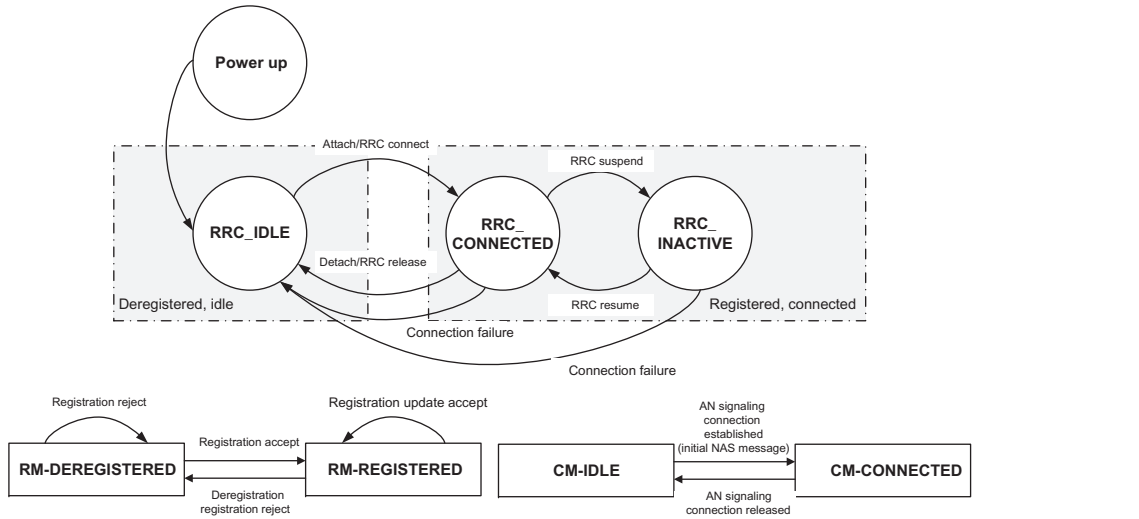
Other SIBs are carried in SI messages, which are transmitted on the DL-SCH. The SIBs with the same periodicity can only be mapped to the same SI message. Each SI message is transmitted within periodically occurring time domain windows referred to as SI-windows with same length for all SI messages. Each SI message is associated with an SI-window and

the SI-windows of different SI messages do not overlap. That is, within one SI-window only the corresponding SI message is transmitted. Any SIB except SIB1 can be configured to be cell-specific or area-specific, using an indication in SIB1. The cell-specific SIB is applicable only within the cell that provides the SIB, while the area-specific SIB is applicable within an area referred to as SI area, which consists of one or several cells and is identified by s*ystemInformationAreaID*. For a UE in RRC_CONNECTED state, the network can provide SI through dedicated signaling using the *RRCReconfiguration* message, for example, if the UE has an active BWP with no common search space configured to monitor SI or paging. For PSCell and SCells, the network provides the required SI by dedicated signaling, that is, within an *RRCReconfiguration* message. Nevertheless, the UE acquires the MIB of the PSCell in order to obtain the SFN timing of the SCG which may be different from that of MCG. Upon change of the relevant SI for SCell, NG-RAN releases and adds the concerned SCell. The physical layer imposes a limit on the maximum size of a SIB. The maximum SIB1 or *SI message* size is 2976 bits [14].

We will explain in Chapter 3 that the PDCCH (physical downlink control channel) monitoring occasions for SI message are determined based on the search space indicated by *searchSpaceOtherSystemInformation* parameter, if the latter parameter is not set to zero. The PDCCH monitoring occasions for SI message, which are not overlapping with uplink symbols (determined according to *tdd-UL-DL-ConfigurationCommon*) are sequentially numbered from one in the SI-window. The PDCCH monitoring occasion(s) $[xN + K]$ for SI message in SI-window correspond to the Kth transmitted synchronization signal block [see Chapter 3 for description], where $x = 0, 1, \ldots, X - 1$, $K = 1, 2, \ldots, N$, N is the number of actual transmitted synchronization signal blocks determined according to *ssb-PositionsInBurst* in SIB1 and $X = \lfloor \textit{Number of PDCCH monitoring occasions in SI} - \textit{window}/N \rfloor$ [14].

2.3.3 User Equipment States and State Transitions

The operation of the RRC sublayer is guided by a state machine which defines the states that a UE may be present in at any time during its operation in the network. Apart from RRC_CONNECTED and RRC_IDLE states, which are similar to those of LTE, the NR has introduced a new RRC state referred to as RRC_INACTIVE state. As shown in Fig. 2.23, when a UE is powered up, it is in disconnected and idle mode. However, the UE can transition to the connected mode with initial access and RRC connection establishment. If there is no UE activity for a short period of time, the UE can suspend its active session and transition to the inactive mode. Nevertheless, it can resume its session by moving to the connected mode. 5G applications and services have different characteristics. To meet the requirements of different services, it was imperative to reduce the control-plane latency by introducing a new RRC state machine and a dormant state. The URLLC services are



UE status	OFF	Attach	Idle/registered	Connected to EPC	Active
EMM state	Deregistered		Registered		
ECM state	Idle				Connected
RRC state	Idle	Connected	Idle	Connected	
Mobility	-	UE-based	UE-based	Network-based	

LTE

UE status	OFF	Attach	Connected/inactive	Connected/active
RM state	Deregistered		Registered	
CM state	-	Connected		
RRC state	-	Connected	Inactive	Connected
Mobility	-	UE-based	UE-based/NW assisted	Network-based

NR

Figure 2.23
NR UE states and comparison to LTE [8,14].

characterized by transmission of frequent/infrequent small packets that require very low latency and high reliability; thus the devices must stay in a low-activity state, and intermittently transmit uplink data and/or status reports with small payloads to the network. There is further a need for periodic/aperiodic downlink small packet transmissions. A UE can move to RRC idle mode from RRC connected or RRC inactive state.

A UE needs to register with the network to receive services that requires registration. Once registered, the UE may need to update its registration with the network either periodically, in order to remain reachable (periodic registration update); or upon mobility (mobility registration update); or to update its capabilities or renegotiate protocol parameters (mobility registration update). As we discussed in Chapter 1, the mobility state of a UE in 5G core network (CN) can be either RM-REGISTERED or RM-DEREGISTERED depending on whether the UE is registered with 5GC. The registration management (RM) states are used in the UE and the Access and Mobility Management Function (AMF) to reflect the registration status of the UE in the selected PLMN. In the RM-DEREGISTERED state, the UE is not registered with the network. The UE context in AMF holds no valid location or routing information for the UE; thus the UE is not reachable by the AMF. However, some parts of UE context may still be stored in the UE and the AMF to avoid performing an authentication procedure in each registration procedure. In the RM-REGISTERED state, the UE is registered with the network and can receive services that require registration with the network [1]. Fig. 2.23 shows the UE RM states and transition between the two states.

Two connection management states are used to reflect the NAS signaling connectivity status of the UE with the CN, namely, CM-IDLE and CM-CONNECTED. A UE in CM-IDLE state has no NAS signaling connection established with the AMF over N1, whereas in CM-CONNECTED state the UE has a NAS signaling connection with the AMF over N1. A NAS signaling connection uses the RRC connection between the UE and the NG-RAN to encapsulate NAS messages exchanged between the UE and the CN in the RRC messages. The NR RRC protocol states consist of three states, where in addition to RRC_IDLE and RRC_CONNECTED states, a third state has been introduced, RRC_INACTIVE, as a primary sleeping state prior to transition to RRC_IDLE state in order to save UE power and to allow fast connection setup [1,8].

In RRC_IDLE state there is no UE context, that is, the parameters necessary for communication between the device and the network, in the radio access network, and the device is not registered to a specific cell. From the CN perspective, the device is in the CM-IDLE state where no data transfer may be performed since the device is in a sleep mode to conserve the battery. The idle mode UEs periodically wake up to receive paging messages, if any, from the network. In this mode, the mobility is managed by the device through cell reselection. The uplink synchronization is not maintained in the idle mode, and thus the UE is required to perform a random-access procedure in order to transition to the connected mode. As part of transitioning to the RRC_CONNECTED state, the UE context is established in the device and the

network. From the CN perspective the device is in the CM-CONNECTED state and registered with the network. The cell to which the device belongs is known and a temporary identity for the device, that is, C-RNTI is used to identify the UE in NG-RAN while in the connected mode. The connected mode is intended for data transfer to/from the device; however, a DRX cycle can be configured during inactive times to reduce the UE power consumption. Since there is an already-established UE context in the gNB in the connected mode, transition from DRX mode and starting data transfer is relatively fast, requiring no connection setup. In this mode, the mobility is managed by the network. The device provides neighbor cell measurements to the network, and the network would instruct the device to perform a handover when necessary. The UE may lose uplink synchronization and may need to perform random-access procedure to be synchronized.

The LTE system only supports idle and connected modes. In practice, the idle mode serves as the primary means to reduce the device power consumption during operation in the network. However, intermittent transmission of small packets by some delay-sensitive applications results in frequent state transitions which would cause signaling overhead and additional delays. Therefore to reduce the signaling overhead and the latency, a third state has defined in NR. As shown in Fig. 2.23, in RRC_INACTIVE state, the UE context is maintained in the device and the gNB and the CN connection is preserved (i.e., the device is in CM-CONNECTED state). As a result, the transitions to or from the RRC_CONNECTED state for data transfer becomes more efficient and faster. In this mode, the UE is configured with sleep periods similar to the idle mode, and mobility is controlled through cell reselection without involvement of the network. The characteristics of NR UE states are summarized in Table 2.2. One important difference between the UE states in NR is the way that the mobility is handled. In the idle and inactive states, the mobility is

Table 2.2: Characteristics of RRC states in NR [19].

<table>
<tr><th>RRC_IDLE</th><th>RRC_INACTIVE</th><th>RRC_CONNECTED</th></tr>
<tr><td colspan="2">UE-controlled mobility based on network configuration (cell reselection)</td><td rowspan="5">Network-controlled mobility within NR and to/from LTE
DRX configured by gNB
Neighbor cell measurements
Network can transmit and/or receive data to/from UE
NG-RAN knows the cell which the UE belongs to</td></tr>
<tr><td>DRX configured by NAS</td><td>DRX configured by NAS or gNB</td></tr>
<tr><td colspan="2">Broadcast of system information</td></tr>
<tr><td>Paging (CN-initiated)</td><td>Paging (CN-initiated or NG-RAN-initiated)</td></tr>
<tr><td>UE has an CN ID that uniquely identifies it w/in a tracking area</td><td>NG-RAN knows the RNA which the UE belongs to</td></tr>
<tr><td>No UE context stored in gNB</td><td colspan="2">UE and NG-RAN have the UE AS context stored, and the 5GC-NG-RAN connection (both control/user-planes) is established for the UE</td></tr>
</table>

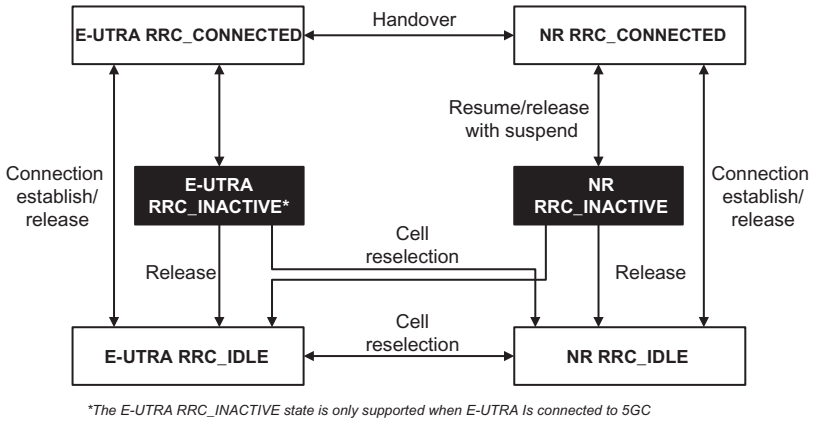


Figure 2.24
UE state machine and state transitions between NR/5GC, LTE/EPC, and LTE/5GC [1].

handled by the device through cell reselection, while in the connected state, the mobility is managed by the network based on measurements.

Fig. 2.24 illustrates an overview of UE RRC state machine and state transitions in NR. A UE has only one RRC state in NR at a given time. The UE is either in RRC_CONNECTED state or in RRC_INACTIVE state when an RRC connection has been established; otherwise, if no RRC connection is established, the UE is in RRC_IDLE state. The RRC states can further be characterized as follows (see the summary in Table 2.2):

- *RRC_IDLE*: In this state, a UE-specific DRX may be configured by upper layers. A UE-controlled mobility based on network configuration will be used. The UE monitors short messages transmitted with P-RNTI over downlink control channel. It also monitors the PCH for CN paging using 5G-S-TMSI and performs neighboring cell measurements and cell (re)selection. It further acquires SI and can send SI request (if configured).
- *RRC_INACTIVE*: In this state, a UE-specific DRX may be configured by upper layers or by RRC layer. A UE-controlled mobility based on network configuration is used. The UE stores the inactive AS context and an RNA is configured by the RRC sublayer. The UE monitors short messages transmitted with P-RNTI over downlink control channel. It monitors the PCH for CN paging using 5G-S-TMSI and RAN paging using full I-RNTI and performs neighboring cell measurements and cell (re)selection. The UE further periodically performs RNA updates (RNAUs) when moving outside the configured RNA. It also acquires SI and can send SI request (if configured).
- *RRC_CONNECTED*: In this state, the UE stores the AS context. The network transfers unicast data to/from the UE. At the lower layers, the UE may be configured with a UE-specific DRX. For carrier-aggregation-capable UEs, the network may use one or

more SCells, aggregated with the SpCell to increase operation bandwidth. For the UEs supporting dual connectivity, the network may use one SCG, aggregated with the MCG, for increased operational bandwidth. Network-controlled mobility within NR and to/from E-UTRA is used in this mode. The UE monitors short messages transmitted with P-RNTI over downlink control channel. It further monitors control channels associated with the shared data channel to determine if data is scheduled for it. The UE provides channel quality and feedback information to the gNB and conducts neighbor cell measurements and measurement reporting and acquires the SI.

2.3.3.1 Idle Mode Procedures

The RRC_IDLE state and RRC_INACTIVE state procedures can be divided into three processes, namely, PLMN selection, cell selection/reselection, and location registration and RNA update. The PLMN selection, cell reselection procedures, and location registration are common for both RRC_IDLE state and RRC_INACTIVE state, whereas RNA update is only applicable to RRC_INACTIVE state. When the UE selects a new PLMN, it transitions from RRC_INACTIVE to RRC_IDLE state. When a UE is powered on, a PLMN is selected by NAS and a number of RATs associated with the selected PLMN are identified for cell selection. The NAS provides a list of equivalent PLMNs that the AS must use for cell selection/reselection.

During cell selection the UE searches for a suitable cell within the selected PLMN. The UE would select a cell, if certain criteria are met. This procedure is known as camping on the cell in 3GPP terminology. The UE then registers with the network by means of NAS registration procedure in the tracking area of the selected cell. A successful location registration would make the selected PLMN as the registered PLMN. If the UE finds a more suitable cell, based on the cell reselection criteria, it reselects that cell and camps on it. If the new cell does not belong to at least one tracking area to which the UE is registered, another location registration is performed. In RRC_INACTIVE state, if the new cell does not belong to the configured RNA, an RNA update procedure is performed. The UE usually searches for higher priority PLMNs at regular time intervals and continues to search for a more suitable cell, if another PLMN has been selected by NAS. If the UE loses the coverage of the registered PLMN, either a new PLMN is automatically selected (automatic mode), or the user is notified of the available PLMNs so that a manual selection can be performed (manual mode).

The purpose of camping on a cell in RRC_IDLE or RRC_INACTIVE state is to enable the UE to receive SI from the PLMN, when registered. If the network needs to send a message or deliver data to the registered UE, it knows the set of tracking areas (in RRC_IDLE state) or RNA (in RRC_INACTIVE state) in which the UE is camped. The network can then send a paging message to the UE on the control channels of all cells in the corresponding set of (tracking) areas. It further enables the UE to receive ETWS and CMAS notifications. During the PLMN selection, the UE scans all RF channels in the NR bands according to its

capabilities to find available PLMNs. On each carrier, the UE searches for the strongest cell and acquires the SI in order to find which PLMN(s) the cell belongs to. If the UE can detect one or several PLMN identities in the strongest cell, each detected PLMN is reported to the NAS as a high-quality PLMN provided that the measured reference signal received power value is greater than or equal to −110 dBm [12].

As we mentioned earlier, the PLMN selection in NR is based on the 3GPP PLMN selection, and that is, cell selection is required upon transition from RM-DEREGISTERED to RM-REGISTERED, from CM-IDLE to CM-CONNECTED, and from CM-CONNECTED to CM-IDLE based on the following principles [8]:

- The UE NAS layer identifies a PLMN and equivalent PLMNs.
- Cell selection is always based on cell-defining SSBs (CD-SSBs)[9] located on the synchronization raster.
- The UE searches the designated NR frequency bands and for each carrier frequency identifies the strongest cell consistent with the CD-SSB. It then detects the SI broadcast from that cell to identify its PLMN(s).
- The UE may conduct a full search in initial cell selection or make use of the stored information to shorten the search, that is, stored information cell selection.
- The UE searches for a suitable cell, and if it is not able to identify a suitable cell, it may proceed with an acceptable cell. When a suitable cell or an acceptable cell is found, the UE camps on that cell and starts the cell reselection procedure.
- A suitable cell is one for which the measured cell attributes satisfy the cell selection criteria; the cell PLMN is the selected PLMN, registered, or an equivalent PLMN; the cell is not barred or reserved, and the cell is not part of a tracking area which is in the list of *forbidden tracking areas for roaming*. An acceptable cell is one for which the measured cell attributes satisfy the cell selection criteria and the cell is not barred.
- Upon transition from RRC_CONNECTED or RRC_INACTIVE to RRC_IDLE, the UE should camp on a cell following cell selection depending on the frequency assigned by the RRC sublayer in the state transition message.
- The UE should attempt to find a suitable cell in the manner described for stored information or initial cell selection. If no suitable cell is found on any frequency or RAT, the UE should attempt to find an acceptable cell.
- In multi-beam operation, the cell quality is derived amongst the beams corresponding to the same cell.

[9] Within the frequency span of a carrier, multiple SSBs can be transmitted. The PCIs of SSBs transmitted in different frequency locations do not have to be unique, that is, different SSBs in the frequency domain can have different PCIs. However, when an SSB is associated with an RMSI, the SSB corresponds to an individual cell, which has a unique NR cell global identifier (NCGI). Such an SSB is referred to as CD-SSB. A PCell is always associated with a CD-SSB located on the synchronization raster.

A UE in RRC_IDLE may perform cell reselection according to the following procedure [8]:

- Cell reselection is always based on CD-SSBs located on the synchronization raster.
- The UE measures the attributes of the serving and neighbor cells to facilitate the reselection process.
- The UE would only need information on the carrier frequencies of the inter-frequency neighbor cells, when conducting search and measurement.
- Cell reselection identifies the cell that the UE should camp on. It is based on cell reselection criteria which involves measurements conducted on the serving and the neighbor cells. It must be noted that intra-frequency cell reselection is based on ranking of the cells and inter-frequency cell reselection is based on absolute priorities, where a UE would camp on the highest priority frequency available.
- A neighbor cell list (NCL) can be provided by the serving cell to facilitate cell selection in specific cases for intra- and inter-frequency neighbor cells. The NCL contains cell-specific cell reselection parameters (e.g., cell-specific offset) for specific neighbor cells. Black lists can be provided to prevent the UE from reselecting to specific intra- and inter-frequency neighboring cells.
- In multi-beam operations, the cell quality is derived amongst the beams corresponding to the same cell.

2.3.3.2 Inactive Mode Procedures

RRC_INACTIVE is a state where a UE remains in CM-CONNECTED state while roaming within an area configured by NG-RAN known as RNA without notifying NG-RAN. In RRC_INACTIVE state, the last serving gNB maintains the UE context and the UE-associated NG connection with the serving AMF and user-plane function (UPF). If the last serving gNB receives downlink data from the UPF or downlink UE-associated signaling from the AMF (except the *UE Context Release Command* message) while the UE is in RRC_INACTIVE state, it pages the UE in the cells corresponding to the RNA and may send an XnAP RAN Paging[10] to neighbor gNB(s), if the RNA includes cells of neighboring gNB(s). Upon receiving the *UE Context Release Command* message while the UE is in RRC_INACTIVE state, the last serving gNB may send the paging message in the cells corresponding to the RNA and may send an XnAP RAN Paging to neighbor gNB(s), if the RNA includes the cells of neighbor gNB(s). Upon receiving the *NG RESET* message while the UE is in RRC_INACTIVE state, the last serving gNB may page the involved UE(s) in

[10] The purpose of the RAN Paging procedure is to enable the NG-RAN $node_1$ to request paging of a UE in the NG-RAN $node_2$. The procedure uses non-UE-associated signaling. The RAN paging procedure is triggered by the NG-RAN $node_1$ by sending the RAN paging message to the NG-RAN $node_2$, in which the necessary information such as UE RAN Paging identity is provided.

the cells corresponding to the RNA and may send an XnAP RAN Paging to neighbor gNB (s), if the RNA includes the cells of neighbor gNB(s) [8].

The AMF provides to the NG-RAN node the *CN assistance information* to assist the NG-RAN node's decision on whether the UE can be moved to the RRC_INACTIVE state. The *CN assistance information* includes the registration area configured for the UE, the *periodic registration update* timer, and the *UE identity index* value, and may further include the UE-specific DRX, an indication if the UE is configured with mobile initiated connection only mode by the AMF, and the *expected UE behavior*. The UE registration area is considered by the NG-RAN node, when configuring the RNA. The UE-specific DRX and *UE identity index* value are used by the NG-RAN node for RAN paging. The *periodic registration update* timer is taken into account by the NG-RAN node to configure *periodic RNA update* timer. The NG-RAN node further considers the *expected UE behavior* to assist the UE RRC state transition decision [8].

During transition to RRC_INACTIVE state, the NG-RAN node may configure the UE with a periodic *RNA update* timer value. If the UE attempts to access a gNB other than the last serving gNB, the receiving gNB triggers the *XnAP retrieve UE context* procedure to obtain the UE context from the last serving gNB and may also trigger a data forwarding procedure including tunnel information for recovery of data from the last serving gNB. Upon successful UE context retrieval, the receiving gNB performs the slice-aware admission control in case of receiving slice information and becomes the serving gNB and further triggers the *NGAP path switch request* and applicable RRC procedures. After the path switch procedure, the serving gNB triggers release of the UE context at the last serving gNB by means of the *XnAP UE context release* procedure. If the UE attempts to access a gNB other than the last serving gNB and the receiving gNB does not find a valid UE context, the receiving gNB can establish a new RRC connection instead of resumption of the previous RRC connection.

A UE in the RRC_INACTIVE state is required to initiate RNAU procedure, when it moves out of the configured RNA. When receiving RNAU request from the UE, the receiving gNB triggers the XnAP retrieve UE context procedure to obtain the UE context from the last serving gNB and may decide to move the UE back to RRC_INACTIVE state, RRC_CONNECTED state, or RRC_IDLE state. In case of periodic RNA update, if the last serving gNB decides not to relocate the UE context, it fails the retrieve UE context procedure and directly moves the UE back to RRC_INACTIVE state or to RRC_IDLE state by an encapsulated *RRCRelease* message. Table 2.3 provides the functional split between the UE NAS and AS procedures in RRC_IDLE and RRC_INACTIVE states.

The UE may use DRX in RRC_IDLE and RRC_INACTIVE states in order to reduce power consumption. The UE monitors one paging occasion (PO) per DRX cycle. A PO is a set of

Table 2.3: Functional split between access stratum and non-access stratum in RRC_IDLE and RRC_INACTIVE states [12].

RRC_IDLE and RRC_INACTIVE State Procedure	UE Non-access Stratum	UE Access Stratum
PLMN selection	• Maintain a prioritized list of PLMNs • Select a PLMN using automatic or manual mode • Request AS to select a cell belonging to this PLMN. For each PLMN, associated RAT(s) may be set • Evaluate reports of available PLMNs from AS for PLMN selection • Maintain a list of equivalent PLMN identities	• Search for available PLMNs, if the associated RAT(s) are set for the PLMN, search among those RAT(s) • Perform measurements to support PLMN selection • Synchronize to a broadcast channel to identify PLMNs • Report available PLMNs with the associated RAT(s) to NAS on request from NAS or autonomously
Cell selection	• Control cell selection by indicating RAT(s) associated with the selected PLMN to be used initially in the search of a cell in the cell selection process • Maintain a list of *forbidden tracking areas* and provide the list to AS	• Perform the required measurements to support cell selection • Detect and synchronize to a broadcast channel • Receive and process broadcast information • Forward NAS system information to NAS • Search for a suitable cell. The cells broadcast one or more PLMN identity in the system information. Search among the associated RATs for that PLMN • Respond to NAS whether such cell is found • If a cell is found which satisfies cell selection criteria, camp on that cell
Cell reselection	• Maintain a list of equivalent PLMN identities and provide the list to AS • Maintain a list of *forbidden tracking areas* and provide the list to AS	• Perform the required measurements to support cell reselection • Detect and synchronize to a broadcast channel • Receive and process broadcast information • Forward NAS system information to NAS • Change cell if a more suitable cell is found
Location registration	• Register the UE as active after power up • Register the UE's presence in a registration area regularly or when entering a new tracking area • De-register UE when shutting down • Maintain a list of *forbidden tracking areas*	• Report registration area information to NAS
RAN notification area update	N/A	• Register the UE's presence in a RAN-based notification area and periodically or when entering a new RNA

PDCCH monitoring occasions and can consist of multiple time slots (e.g., subframe or OFDM symbols) where paging DCI can be sent. One paging frame (PF) is one radio frame and may contain one or multiple PO(s) or starting point of a PO. In multi-beam operations, the UE can assume that the same paging message is repeated in all transmitted beams, and thus the selection of the beam(s) for the reception of the paging message is up to UE implementation. The paging message is the same for both RAN-initiated paging and CN-initiated paging. The UE initiates *RRC Connection Resume* procedure upon receiving RAN-initiated paging. If the UE receives a CN-initiated paging in RRC_INACTIVE state, the UE transitions to RRC_IDLE state and informs the NAS. The PF and PO for paging are determined by the following expressions. The *SFN* for the PF is determined by $(SFN + PF\_offset)\text{mod}T = (T/N)(UE\_ID\text{mod}N)$ where index i_s indicating the index of the PO is determined by $i_s = \lfloor UE\_ID/N \rfloor \text{mod}N_s$. The parameters of the latter equations are defined as follows: T denotes the DRX cycle of the UE where T is determined by the shortest of the UE-specific DRX value, if configured by RRC or upper layers and a default DRX value broadcast in SI. If UE-specific DRX is not configured by RRC or by upper layers, the default value is applied; N is the number of total PFs in T; N_s denotes the number of POs for a PF; PF_offset is the offset used for PF determination; and $UE\_ID = 5G{-}S{-}TMSI\text{mod}1024$ [12].

A UE in RRC_INACTIVE state performs cell reselection similar to the procedure earlier defined for the RRC_IDLE state. The UE in the RRC_INACTIVE state can be configured by the last serving NG-RAN node with an RNA, where the RNA can cover a single or multiple cells and is contained within the CN registration area, as well as an RNA update that is periodically sent by the UE and is also sent when the cell reselection procedure of the UE selects a cell that does not belong to the configured RNA [8].

2.3.3.3 Connected Mode Procedures

In the RRC_CONNECTED state, the device has a connection established to the network. The goal of connected-state mobility is to ensure that this connectivity is sustained without interruption or noticeable degradation as the device moves across the network. To satisfy this goal, the device continuously searches for and conducts measurements on new cells both at the current carrier frequency (intra-frequency measurements) and at different carrier frequencies (inter-frequency measurements) that the device has been configured to do. Such measurements are conducted on the SSB in the same way as for initial access and cell selection/reselection in idle and inactive modes. However, the measurements can also be conducted on configured CSI-RS. In the connected mode, the handover is network-controlled, and the UE does not make any decision on handover to a different cell. Based on different triggering conditions such as the relative power of a measured SSB relative to that of the current cell, the device reports the result of the measurements to the network. The network then makes a decision as to whether the device has to be handed-over to a

new cell. It should be noted that the reporting is provided through RRC signaling and not layer 1 measurement and reporting framework used for beam management. Apart from some cases in small cell network architectures where the cells are relatively synchronized, the device must perform a new uplink synchronization with respect to the target cell prior to handover. To obtain synchronization to a new cell, the UE has to perform a contention-free random-access procedure using resources specifically assigned to the device with no risk of collision with the goal of attaining synchronization to the target cell. Thus, only first two steps of the random-access procedure are needed which includes the preamble transmission and the corresponding random-access response providing the device with updated transmission timing [17].

In RRC_CONNECTED state, a network-controlled mobility scheme is used which has two variants: cell-level mobility and beam-level mobility.

In *cell-level mobility* explicit RRC signaling is used to trigger a handover. For inter-gNB handover, the signaling procedures consist of four components as follows [8]:

1. The source gNB initiates handover and issues a *Handover Request* message over Xn interface.
2. The target gNB performs admission control and provides the RRC configuration as part of the *Handover ACK* message.
3. The source gNB provides the RRC configuration to the UE in the *Handover Command*. The *Handover Command* message includes at least the cell ID and all information required to access the target cell so that the UE can access the target cell without detecting its SI. In some cases, the information required for contention-based and contention-free random-access procedure can be included in the *Handover Command* message. The access information to the target cell may include beam-specific information.
4. The UE moves the RRC connection to the target gNB and replies with a *Handover Complete* message.

The handover mechanism triggered by RRC signaling requires the UE to at least reset the MAC entity and reestablish RLC entity. The NR supports RRC-managed handovers with and without PDCP entity reestablishment. For DRBs using RLC-AM mode, PDCP can either be reestablished along with security key change or initiate a data recovery procedure without key change. For DRBs using RLC-UM mode and for SRBs the PDCP entity can be reestablished either together with security key change or to remain as it is without key change. Data forwarding, in-sequence delivery, and duplication avoidance during handover can be guaranteed, when the target gNB uses the same DRB configuration as the source gNB. The NR further supports timer-based handover failure procedure where the RRC connection reestablishment procedure is used for recovering from handover failure.

In *beam-level mobility*, explicit RRC signaling is not required to trigger handover. The gNB provides the UE via RRC signaling the measurement configuration containing configurations of SSB/CSI-RS resources and resource sets, reports, and trigger states for triggering channel and interference measurements and reports. The beam-level mobility is performed at lower layers by means of physical layer and MAC layer control signaling, and the RRC is not required to know which beam is being used at any given time. The SSB-based beam-level mobility is based on the SSB associated with the initial DL BWP and can only be configured for the initial DL BWPs and for DL BWPs containing the SSB associated with the initial DL BWP. For other DL BWPs, the beam-level mobility can only be performed based on CSI-RS.

2.3.4 User Equipment Capability

The UE capabilities in NR do not rely on UE categories. Unlike LTE, the NR UE categories are associated with fixed peak data rates and defined for marketing purposes; thus they are not signaled to the network. Instead, the network determines the uplink and downlink data rate supported by a UE from the supported band combinations and from the baseband capabilities such as modulation scheme, and the number of MIMO layers. In order to limit signaling overhead, the gNB can ask the UE to provide NR capabilities for a restricted set of bands. When responding, the UE can skip a subset of the requested band combinations when the corresponding UE capabilities are the same.

The NR defines an approximate (peak) data rate for a given number of aggregated carriers in a band or band combination as follows [13]:

$$D_{NR} = 10^{-6} \sum_{j=1}^{J} \left[v_{layer}^{(j)} Q_m^{(j)} f^{(j)} R_{\max} \frac{12 N_{PRB}^{BW(j)}(\mu)}{T_s(\mu)} \left(1 - \alpha^{(j)}\right) \right] \text{(Mbps)}$$

where J is the number of aggregated component carriers in a band or band combination; $R_{\max} = 948/1024$; $v_{layer}^{(j)}$ is the maximum number of supported layers given by higher layer parameter *maxNumberMIMO-LayersPDSCH* for the downlink and higher layer parameters *maxNumberMIMO-LayersCB-PUSCH* and *maxNumberMIMO-LayersNonCB-PUSCH* for the uplink; $Q_m^{(j)}$ is the maximum supported modulation order given by higher layer parameter *supportedModulationOrderDL* for the downlink and higher layer parameter *supportedModulationOrderUL* for the uplink; $f^{(j)}$ is the scaling factor given by higher layer parameter *scalingFactor* which can take the values of 1, 0.8, 0.75, and 0.4; μ is the numerology (an OFDM parameter); $T_s(\mu)$ is the average OFDM symbol duration in a subframe for numerology μ which is given as $T_s(\mu) = 0.0142^{\mu}$ for normal cyclic prefix; $N_{PRB}^{BW(j)}(\mu)$ is the

maximum resource block allocation in bandwidth $BW^{(j)}$ with numerology μ where $BW^{(j)}$ is the UE supported maximum bandwidth in the given band or band combination; and $\alpha^{(j)}$ is the estimated overhead which takes the values 0.14 [for downlink in frequency range (FR) 1], 0.18 (for downlink in FR2), 0.08 (for uplink in FR1), and 0.10 (for uplink in FR2). Note that only one of the uplink or supplemental uplink carriers with the higher data rate is counted for a cell operating SUL.

The approximate maximum data rate can be computed as the maximum of the approximate data rates computed using the above expression for each of the supported band or band combinations. For LTE in the case of dual connectivity, the approximate data rate for a given number of aggregated carriers in a band or band combination is computed as $D_{MR-DC} = 10^{-3} \sum_{j=1}^{J} TBS_j$ (Mbps), where J is the number of aggregated LTE component carriers in multi-radio dual connectivity (MR-DC) band combination and TBS_j is the total maximum number of DL-SCH transport block bits received within a 1 ms TTI for the jth component carrier based on the UE supported maximum MIMO layers for the jth carrier, and based on the modulation order and the number of physical resource blocks (PRBs) in the bandwidth of the jth carrier. The approximate maximum data rate can be calculated as the maximum of the approximate data rates computed using the latter equation for each of the supported band or band combinations. For MR-DC, the approximate maximum data rate is computed as the sum of the approximate maximum data rates from NR and LTE [13].

The total layer 2 buffer size is another UE capability attribute that is defined as the sum of the number of bytes that the UE is capable of storing in the RLC transmission windows and RLC reception and reordering windows and also in PDCP reordering windows for all radio bearers. The total layer 2 buffer size in MR-DC and NR-DC scenario is the maximum of the calculated values based on the following equations [13]:

$$MaxULDataRate\_MN \times RLCRTT\_MN + MaxULDataRate\_SN \times RLCRTT\_SN + MaxDLDataRate\_SN \times RLCRTT\_SN + MaxDLDataRate\_MN \times (RLCRTT\_SN + X2/Xndelay + QueuinginSN)$$

$$MaxULDataRate\_MN \times RLCRTT\_MN + MaxULDataRate\_SN \times RLCRTT\_SN + MaxDLDataRate\_MN \times RLCRTT\_MN + MaxDLDataRate\_SN \times (RLCRTT\_MN + X2/Xndelay + QueuinginMN)$$

In other scenarios, the total layer 2 buffer size is calculated as $MaxDLDataRate \times RLCRTT + MaxULDataRate \times RLCRTT$. It must be noted that the additional layer 2 buffer required for preprocessing of data is not taken into account in above formula. The total layer 2 buffer size is determined as the maximum layer 2 buffer size of all calculated ones for each band combination and the applicable Feature Set combination in the supported MR-DC or NR band combinations. The RLC RTT for NR cell group corresponds to the smallest subcarrier spacing (SCS) numerology supported in the band combination and the applicable Feature Set combination. The NR

specifications specify *X2/Xn delay + Queuing in SN* = , if SCG is NR, and 55 ms if SCG is LTE. The NR specifications define *X2/Xn delay + Queuing in MN = 25 ms*, if MCG is NR, and 55 ms if MCG is LTE. They further specify RLC RTT for LTE cell group as 75 ms and the RLC RTT for NR cell group ranging from 20 to 50 ms depending on the subcarrier spacing [13].

The maximum supported data rate for integrity-protected DRBs (see Section 2.2.3.3) is a UE capability indicated at NAS layer, with a minimum value of 64 kbps and a maximum value of the highest data rate supported by the UE. In case of failed integrity check (i.e., due to a faulty or missing MAC-I) the corresponding PDU is discarded by the receiving PDCP entity.

2.4 Discontinuous Reception and Power-Saving Schemes

The UE monitors physical downlink control channel (PDCCH) while in RRC_CONNECTED state. This activity is controlled by the DRX and bandwidth adaptation schemes configured for the UE. When bandwidth adaptation is configured, the UE only has to monitor PDCCH on the active BWP, that is, it does not have to monitor PDCCH on the entire downlink frequency of the cell. A BWP inactivity timer (independent from the DRX inactivity timer) is used to switch the active BWP to the default one. The latter timer is restarted upon successful PDCCH decoding and the switching to the default BWP happens when it expires. When DRX is configured, the UE is not required to continuously monitor the PDCCH. The DRX mechanism is characterized by the following parameters [8]:

- *On-duration*: The time interval during which the UE would expect to receive the PDCCH. If the UE successfully decodes the PDCCH, it stays awake and starts the inactivity timer.
- *Inactivity timer*: The time interval during which the UE waits for successful decoding of the PDCCH, starting from the last successful decoding of a PDCCH. If the decoding fails, the UE can go back to sleep. The UE restarts the inactivity timer following a single successful decoding of a PDCCH for the first transmission only (i.e., not for retransmissions).
- *Retransmission-timer*: The time interval until a retransmission can be expected.
- *Cycle*: It specifies the periodic repetition of the on-duration followed by a possible period of inactivity.
- *Active-time*: The total time duration that the UE monitors PDCCH. This includes the on-duration of the DRX cycle, the time that the UE is performing continuous reception while the inactivity timer is running, and the time when the UE is performing continuous reception while awaiting a retransmission opportunity.

Due to bursty nature of the packet data traffic, which is characterized by intermittent periods of transmission activity followed by longer periods of inactivity, and to reduce the UE power consumption, NR supports a DRX scheme similar to that of LTE. Bandwidth adaptation and dynamic carrier activation/deactivation are two other power-saving mechanisms supported in NR. The underlying mechanism for DRX is a configurable DRX cycle in the device. When a DRX cycle is configured, the device monitors the downlink control channel only during the active-time and sleeps, with its receiver circuitry switched off, during the inactivity time, leading to a significant reduction in UE power consumption. The longer the DRX inactive time, the lower the power consumption. However, this would have certain implications for the scheduler, since the device is only reachable when it is active according to the DRX cycle configured for it. In many cases, if the device has been scheduled and is engaged in receiving or transmitting data, it is likely that it will be scheduled again soon; thus waiting until the next activity period according to the DRX cycle would result in additional delays. Therefore, to reduce the delays, the device remains in the active state for a configurable period of time after being scheduled. This is realized by an inactivity timer started by the UE every time that it is scheduled where the UE remains awake until the time expires (Fig. 2.25). Since NR supports multiple numerologies the time unit of the DRX timers is specified in milliseconds in order to avoid associating the DRX periodicity to a certain numerology (Fig. 2.26).

The NR HARQ retransmissions are asynchronous in both downlink and uplink. If the device has been scheduled a transmission in the downlink that it cannot decode, a typical gNB behavior is to retransmit the data at a later time. In practice, the DRX scheme has a configurable timer which is started after an erroneously received transport block and is used to wake up the UE receiver when it is likely for the gNB to schedule a retransmission. The value of the timer is preferably set to match the (implementation-specific) roundtrip time in the HARQ protocol. The above mechanism is a (long) DRX cycle in conjunction with the device remaining awake for a period of time after being scheduled. However, in some services such voice over IP, which is characterized by periods of regular transmission, followed by periods of no activity, a second (short) DRX cycle can be optionally configured in addition to the long DRX cycle.

The RRC entity controls the DRX operation by configuring the following parameters [14]:

- *drx-onDurationTimer*: The duration at the beginning of a DRX cycle.
- *drx-SlotOffset*: The delay before starting the *drx-onDurationTimer*.
- *drx-InactivityTimer*: The duration after the PDCCH occasion in which a PDCCH indicates a new uplink/downlink transmission for the MAC entity.
- *drx-RetransmissionTimerDL* (per-DL HARQ process except for the broadcast process): The maximum duration until a downlink retransmission is received.

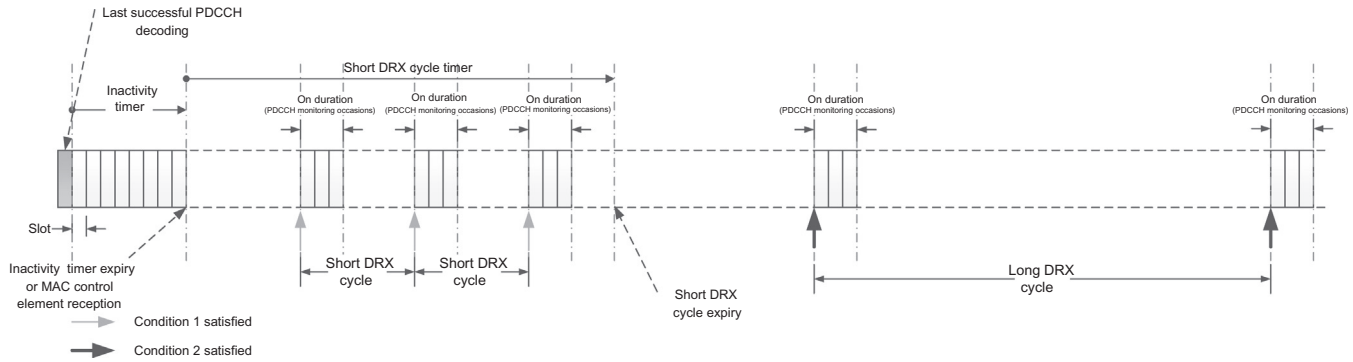


Figure 2.25
Illustration of DRX mechanism [8].

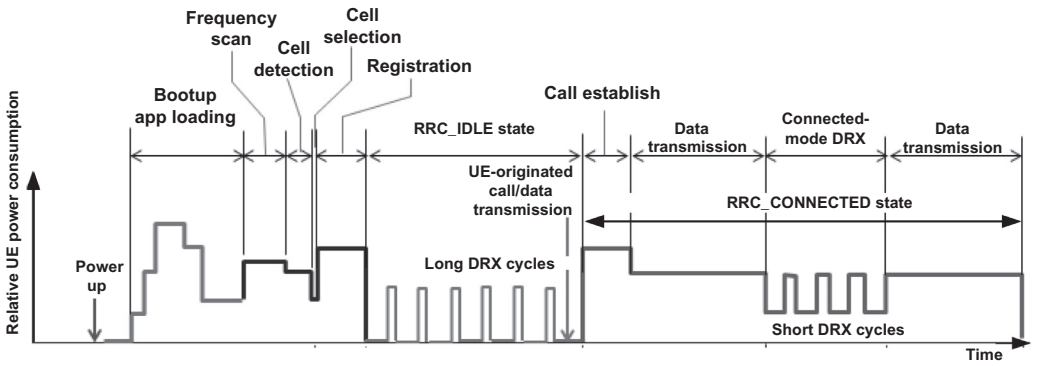


Figure 2.26
Example UE power consumption when transitioning through various RRC states [20].

- *drx-RetransmissionTimerUL* (per-UL HARQ process): The maximum duration until a grant for uplink retransmission is received.
- *drx-LongCycleStartOffset*: The long DRX cycle and *drx-StartOffset* which define the subframe where the long and short DRX cycle starts.
- *drx-ShortCycle* (optional): The short DRX cycle.
- *drx-ShortCycleTimer* (optional): The duration in which the UE follows the short DRX cycle.
- *drx-HARQ-RTT-TimerDL* (per-DL HARQ process except for the broadcast process): The minimum duration before a downlink assignment for HARQ retransmission is expected by the MAC entity.
- *drx-HARQ-RTT-TimerUL* (per-UL HARQ process): The minimum duration before an uplink HARQ retransmission grant is expected by the MAC entity.

2.5 Mobility Management, Handover, and UE Measurements

The NR performs load balancing through handover and redirection mechanisms upon RRC release and through use of inter-frequency and inter-RAT absolute priorities as well as inter-frequency *Qoffset* parameters (see Section 2.5.2). The measurements performed by a UE for connected mode mobility are classified into three types, namely, intra-frequency NR, inter-frequency NR, and inter-RAT measurements for LTE. For each measurement type, one or several measurement objects can be defined (a measurement object defines the carrier frequency to be monitored). For each measurement object, one or several reporting configurations can be defined (a reporting configuration defines the reporting criteria). Three reporting criteria are used: (1) event-triggered reporting, (2) periodic reporting, and (3) event-triggered periodic reporting. The association between a measurement object and a reporting configuration is created by a

measurement identity (a measurement identity associates one measurement object and one reporting configuration of the same RAT). By using several measurement identities (one for each measurement object, reporting configuration pair), it is possible to associate several reporting configurations to one measurement object and to associate one reporting configuration to several measurement objects [8]. The measurements identity is used when reporting results of the measurements. Measurement quantities are considered separately for each RAT.

Measurement commands are used by NG-RAN to instruct the UE to start, modify, or stop measurements. Handover can be performed within the same RAT and/or CN, or it can involve a change of the RAT and/or CN. Inter-system fallback toward LTE RAN is performed for load balancing when 5GC does not support emergency services or voice services. Depending on certain criteria such as CN interface availability, network configuration, and radio conditions, the fallback procedure results in either connected-state mobility (handover procedure) or idle state mobility (redirection) [8].

2.5.1 Network-Controlled Mobility

The mobility of the UEs in RRC_CONNECTED state is controlled by the network (network-controlled mobility), which is classified into two types of mobility, namely, cell-level mobility and beam-level mobility. The *cell-level mobility* requires explicit RRC signaling in order to be triggered, which results in handover. The main steps of the inter-gNB handover signaling procedures are illustrated in Fig. 2.27. The inter-gNB handover comprises the following steps [8]:

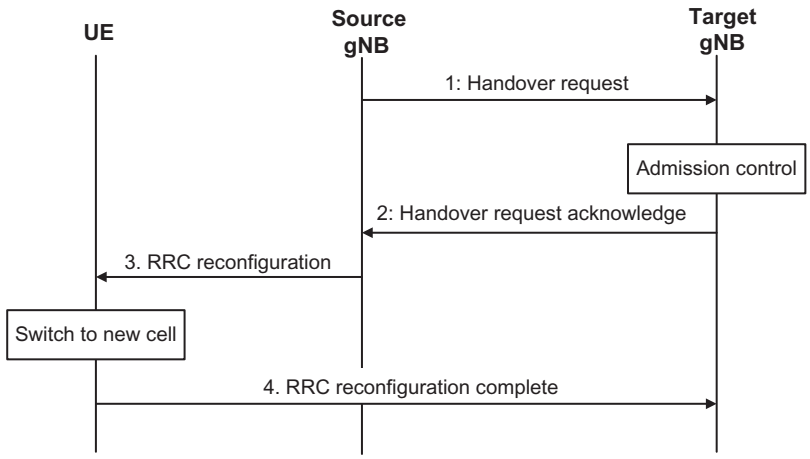


Figure 2.27
Inter-gNB handover procedure [8].

1. The source gNB initiates handover and issues a *Handover Request* over Xn interface.
2. The target gNB performs admission control and provides the RRC configuration as part of the *Handover ACK*.
3. The source gNB provides the RRC configuration to the UE in the *Handover Command* message, which includes the cell ID and all information required to access the target cell, so that the UE can access the target cell without detecting that cell's SI. In some cases, the information required for contention-based and contention-free random-access procedure can be included in the *Handover Command* message. The access information to the target cell may include beam-specific information.
4. The UE moves the RRC connection to the target gNB and replies with the *Handover Complete* message.

The user data can be sent in step 4, if the grant allows. The handover mechanism triggered by RRC signaling requires the UE to reset the MAC entity and reestablish RLC entity. The RRC-triggered handovers with and without PDCP entity reestablishment are both supported in NR. For DRBs using RLC-AM mode, the PDCP entity can either be reestablished along with a security key change or initiate a data recovery procedure without a key change. For DRBs using RLC-UM mode and for SRBs, the PDCP entity can either be reestablished in conjunction with a security key change or to remain as it is without a key change. Data forwarding, in-sequence delivery, and duplication avoidance at handover can be guaranteed when the target gNB uses the same DRB configuration as the source gNB. Timer-based handover failure procedure is supported in NR where an RRC connection reestablishment procedure is used for recovering from handover failure [8].

The *beam-level mobility* does not require explicit RRC signaling in order to be triggered. The gNB provides the UE, via RRC signaling, with measurement configuration containing configurations of SSB/CSI resources and resource sets, as well as trigger states for triggering channel and interference measurements and reports. Beam-level mobility is then managed at lower layers through physical layer and MAC sublayer control signaling. The RRC sublayer is not required to know about the beam that is used at any given time. The SSB-based beam-level mobility is based on the SSB associated with the initial DL BWP and can only be configured for the initial DL BWPs and for DL BWPs containing the SSB associated with the initial DL BWP. For other DL BWPs the beam-level mobility can only be performed based on CSI-RS measurements [8].

2.5.1.1 Control-Plane Handover Procedures

The intra-NR handover includes the preparation and execution phases of the handover procedure performed without 5GC involvement, that is, preparation messages are directly exchanged between the gNBs. The release of the resources at the source gNB during the handover completion phase is triggered by the target gNB. Fig. 2.28 shows the basic

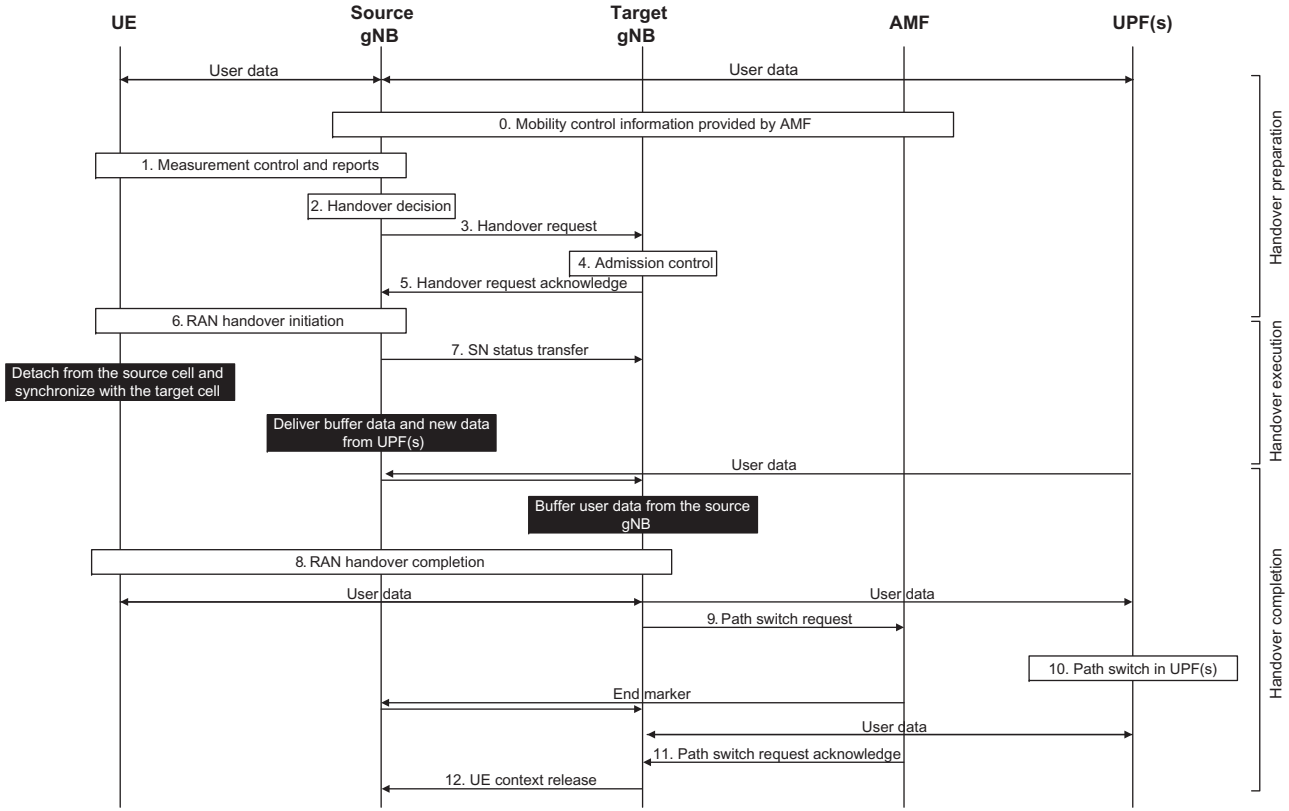


Figure 2.28
Intra-AMF/UPF handover in NR [8].

handover procedure where the AMF and the UPF entities do not change. The processing steps of this handover are as follows [8]:

1. The UE context within the source gNB contains information concerning roaming and access restrictions, which were provided either at connection establishment or at the last tracking area update.
2. The source gNB configures the UE measurement procedures and the UE reports according to the measurement configuration.
3. The source gNB decides to handover the UE, based on measurement reports and radio resource management (RRM) information.
4. The source gNB issues a *Handover Request* message to the target gNB passing a transparent RRC container with necessary information to prepare for the handover at the target gNB. The information includes target cell ID, K_{gNB^*}, C-RNTI of the UE in the source gNB, RRM configuration including UE inactive time, basic AS-configuration including *antenna info and DL carrier frequency*, the current QoS flow to DRB mapping rules applied to the UE, the SIB1 from source gNB, the UE capabilities for different RATs, and PDU Session related information, and can further include the UE reported measurement information including beam-related information. The PDU Session related information includes the slice information (if supported) and QoS flow level QoS profile(s). After issuing a *Handover Request,* the source gNB should not reconfigure the UE, including performing reflective QoS flow to DRB mapping.
5. Admission control may be performed by the target gNB. Slice-aware admission control is performed, if the slice information is sent to the target gNB. If the PDU sessions are associated with non-supported slices, the target gNB would reject those PDU sessions.
6. The target gNB prepares the handover with L1/L2 and sends the *Handover Request Acknowledge* to the source gNB, which includes a transparent container to be sent to the UE as an RRC message to perform the handover.
7. The source gNB triggers the Uu handover by sending an *RRCReconfiguration* message to the UE, containing the information required to access the target cell, that is, the target cell ID, the new C-RNTI, and the target gNB security algorithm IDs for the selected security algorithms. It can also include a set of dedicated RACH resources, the association between RACH resources and SSB(s), the association between RACH resources and UE-specific CSI-RS configuration(s), common RACH resources, and the SI of the target cell.
8. The source gNB sends the *SN Status Transfer* message to the target gNB.
9. The UE synchronizes to the target cell and completes the RRC handover procedure by sending *RRCReconfigurationComplete* message to target gNB.

10. The target gNB sends a *Path Switch Request* message to the AMF to trigger 5GC to switch the downlink data path toward the target gNB and to establish an NG-C interface instance toward the target gNB.
11. 5GC switches the downlink data path toward the target gNB. The UPF sends one or more end-marker packets on the old path to the source gNB per PDU session/tunnel and then can release any user-plane/TNL resources toward the source gNB.
12. The AMF confirms the *Path Switch Request* message with the *Path Switch Request Acknowledge* message.
13. Upon reception of the *Path Switch Request Acknowledge* message from the AMF the target gNB sends the *UE Context Release* message to inform the source gNB about the success of the handover. The source gNB can then release radio and control-plane-related resources associated to the UE context. Any ongoing data forwarding may continue.

The RRM configuration can include both beam measurement information (for layer 3 mobility) associated to SSB(s) and CSI-RS(s) for the reported cell(s), if both types of measurements are available. Also, if carrier aggregation is configured, the RRM configuration can include the list of best cells on each frequency for which measurement information is available. The RRM measurement information can also include the beam measurement for the listed cells that belong to the target gNB.

The common RACH configuration for beams in the target cell is only associated to the SSB(s). The network can have dedicated RACH configurations associated to the SSB(s) and/or have dedicated RACH configurations associated to CSI-RS(s) within a cell. The target gNB can only include one of the following RACH configurations in the handover command to enable the UE to access the target cell:

1. Common RACH configuration
2. Common RACH configuration + dedicated RACH configuration associated with SSB
3. Common RACH configuration + dedicated RACH configuration associated with CSI-RS

The dedicated RACH configuration allocates RACH resource(s) along with a quality threshold to use them. When dedicated RACH resources are provided, they are prioritized by the UE, and the UE does not switch to contention-based RACH resources as long as the quality threshold of those dedicated resources is met. The order to access the dedicated RACH resources is up to UE implementation.

2.5.1.2 User-Plane Handover Procedures

The user-plane aspects of intra-NR handover for the UEs in RRC_CONNECTED state include the following principles to avoid loss of user data during handover [8]. During handover preparation, the user-plane tunnels can be established between the source gNB

and the target gNB. During handover execution, the user data can be forwarded from the source gNB to the target gNB. Packet forwarding should be done in order, as long as the packets are received at the source gNB from the UPF or the source gNB buffer has not been emptied. During handover completion, the target gNB sends a path switch request message to the AMF to inform it that the UE has been granted access and the AMF then triggers path switch related 5GC internal signaling and actual path switch of the source gNB to the target gNB in UPF. The source gNB should continue forwarding data, as long as packets are received at the source gNB from the UPF or the source gNB buffer has not been emptied.

For RLC-AM bearers, in-sequence delivery, and duplication avoidance, the PDCP SN is maintained on a per DRB basis and the source gNB informs the target gNB about the next downlink PDCP SN to allocate to a packet which does not have a PDCP sequence number yet, neither from source gNB nor from the UPF. For security synchronization, the HFN is maintained and the source gNB provides the target gNB one reference HFN for the uplink and one for the downlink, that is, HFN and corresponding SN. In both UE and target gNB, a window-based mechanism is used for duplication detection and reordering. The occurrence of duplicates over the air interface in the target gNB is minimized by means of PDCP-SN-based reporting at the target gNB by the UE. In the uplink, the reporting is optionally configured on as per DRB basis by the gNB and the UE initially starts by transmitting those reports when granted resources are in the target gNB. In the downlink, the gNB can decide when and for which bearers a report is sent, and the UE does not wait for the report to resume UL transmission [8].

The target gNB retransmits and prioritizes all downlink data forwarded by the source gNB excluding the PDCP SDUs for which the reception was acknowledged through PDCP-SN-based reporting by the UE, that is, the target gNB should initially send all forwarded PDCP SDUs with PDCP SNs, then all forwarded downlink PDCP SDUs without SNs before sending new data from 5GC. Lossless delivery, when a QoS flow is mapped to a different DRB at handover, requires that the old DRB to be configured in the target cell. For in-order delivery in the downlink, the target gNB should first transmit the forwarded PDCP SDUs on the old DRB before transmitting new data from 5G CN on the new DRB. In the uplink, the target gNB should not deliver data of the QoS flow from the new DRB to 5G CN before receiving the end-marker on the old DRB from the UE [8].

The UE retransmits in the target gNB all uplink PDCP SDUs starting from the oldest PDCP SDU that has not been acknowledged at RLC sublayer in the source, excluding PDCP SDUs for which the reception was acknowledged through PDCP-SN-based reporting by the target. For RLC-UM bearers the PDCP SN and HFN are reset in the target gNB; no PDCP SDUs are retransmitted in the target gNB; and the target gNB prioritizes all downlink SDAP SDUs forwarded by the source gNB over the data from the CN. To minimize the

losses when a QoS flow is mapped to a different DRB at handover, the old DRB needs to be configured in the target cell. For in-order delivery in the downlink, the target gNB should first transmit the forwarded PDCP SDUs on the old DRB before transmitting new data from 5G CN on the new DRB. In the uplink, the target gNB should not deliver data of the QoS flow from the new DRB to 5G CN before receiving the end-marker on the old DRB from the UE. The UE does not retransmit any PDCP SDU in the target cell for which transmission had been completed in the source cell.

The source NG-RAN node may request downlink data forwarding per QoS flow to be established for a PDU session and may provide information on how it maps QoS flows to DRBs. The target NG-RAN node decides whether data forwarding per QoS flow should be established for a PDU session. If lossless handover is desired and the QoS flow to DRB mapping, applied at the target NG-RAN node, allows employing data forwarding with the same QoS flow to DRB mapping that was used in the source NG-RAN node for a DRB and if all QoS flows mapped to that DRB are accepted for data forwarding, the target NG-RAN node establishes a downlink forwarding tunnel for that DRB. For a DRB for which SN status preservation is important, the target NG-RAN node may decide to establish an uplink data forwarding tunnel.

The target NG-RAN node may also decide to establish a downlink forwarding tunnel for each PDU session. In this case, the target NG-RAN node provides information related to the QoS flows for which data forwarding has been accepted and the corresponding uplink TNL information for data forwarding tunnels to be established between the source and the target NG-RAN nodes [8].

2.5.2 UE-Based Mobility

The PLMN selection in NR is based on 3GPP PLMN selection rules. Cell selection is required upon transition from RM-DEREGISTERED to RM-REGISTERED, from CM-IDLE to CM-CONNECTED, and from CM-CONNECTED to CM-IDLE. The UE NAS layer identifies a selected PLMN and equivalent PLMNs. Cell selection is always based on CD-SSBs located on the synchronization raster. The UE scans the NR frequency bands and for each carrier frequency identifies the strongest cell and the associated CD-SSB. It then detects broadcast SI of the cell to identify its PLMN(s). The UE may scan each carrier in certain order during initial cell selection or take advantage of stored information to expedite the search during stored information cell selection. The UE then tries to identify a suitable cell. If it is not able to identify a suitable cell, it then tries to identify an acceptable cell. When a suitable cell or an acceptable cell is found, the UE camps on it and begins the cell reselection procedure. A suitable cell is a cell whose measured cell attributes satisfy the cell selection criteria; the cell PLMN is the selected PLMN, registered or an equivalent PLMN; the cell is not barred or

reserved, and the cell is not part of a tracking area which is in the list of forbidden tracking areas for roaming. An acceptable cell is a cell whose measured cell attributes satisfy the cell selection criteria, and the cell is not barred.

Upon transition from RRC_CONNECTED or RRC_INACTIVE to RRC_IDLE state, a UE may camp on a cell as a result of cell selection according to the frequency assigned to the UE via RRC signaling in the state transition message. The UE may attempt to find a suitable cell in the above manner described for stored information or initial cell selection. If no suitable cell is found on any frequency or RAT, the UE may attempt to find an acceptable cell.

In multi-beam operations, the cell quality is derived among the beams that are corresponding to the same cell [8].

The cell selection criterion S is considered fulfilled, if the following criteria are satisfied: $Srxlev > 0$ AND $Squal > 0$ where $Srxlev = Q_{rxlevmeas} - \left(Q_{rxlevmin} + Q_{rxlevminoffset}\right) - P_{compensation} - Qoffset_{temp}$ and $Squal = Q_{qualmeas} - \left(Q_{qualmin} + Q_{qualminoffset}\right) - Qoffset_{temp}$. The signaled values $Q_{rxlevminoffset}$ and $Q_{qualminoffset}$ are only applied when a cell is evaluated for cell selection as a result of a periodic search for a higher priority PLMN, while camped normally in a visited PLMN (VPLMN). During the periodic search for higher priority PLMN, the UE may check the S criterion of a cell using stored parameter values that were obtained from a different cell within the higher priority PLMN [12]. The cell selection parameters are described in Table 2.4.

The following rules are used by the UE to limit the required measurements. If the serving cell fulfills $Srxlev > S_{IntraSearchP}$ AND $Squal > S_{IntraSearchQ}$, the UE may skip intra-frequency measurements; otherwise, the UE must perform intra-frequency measurements. The UE must apply the following rules for NR inter-frequency and inter-RAT frequency measurements, which are identified in the SI and for which the UE has priority. For an NR inter-frequency or inter-RAT frequency with a reselection priority higher than the reselection priority of the current NR frequency, the UE is required to perform measurements on higher priority NR inter-frequency or inter-RAT frequencies [6]. For an NR inter-frequency with an equal or lower reselection priority than the reselection priority of the current NR frequency and for inter-RAT frequency with lower reselection priority than the reselection priority of the current NR frequency, if the serving cell fulfills $Srxlev > S_{nonIntraSearchP}$ AND $Squal > S_{nonIntraSearchQ}$ criterion, the UE may skip measurements of NR inter-frequencies or inter-RAT frequency cells of equal or lower priority; otherwise, the UE is required to perform measurements on NR inter-frequencies or inter-RAT frequency cells of equal or lower priority [6,8].

Table 2.4: Cell selection parameters [7,12].

Parameter	Description
Srxlev	Cell selection receive-level value (dB)
Squal	Cell selection quality value (dB)
$Qoffset_{temp}$	An offset temporarily applied to a cell (dB)
$Q_{rxlevmeas}$	Measured cell receive-level value (RSRP)
$Q_{qualmeas}$	Measured cell quality value (RSRQ)
$Q_{rxlevmin}$	Minimum required receive-level in the cell (dBm). If the UE supports SUL frequency for this cell, *Qrxlevmin* is obtained from *RxLevMinSUL*, if present, in *SIB1*, *SIB2*, and *SIB4*. If $Q_{rxlevminoffsetcellSUL}$ is present in *SIB3* and *SIB4* for the candidate cell, this cell-specific offset is added to the corresponding *Qrxlevmin* to achieve the required minimum receive-level in the candidate cell; otherwise, *Qrxlevmin* is obtained from *q-RxLevMin* in *SIB1*, *SIB2*, and *SIB4*. If $Q_{rxlevminoffsetcell}$ is present in *SIB3* and *SIB4* for the candidate cell, this cell-specific offset is added to the corresponding *Qrxlevmin* to achieve the required minimum receive-level in the candidate cell
$Q_{qualmin}$	Minimum required quality level in the cell (dB). If $Q_{qualminoffsetcell}$ is signaled for the concerned cell, this cell-specific offset is added to achieve the required minimum quality level in the candidate cell
$Q_{rxlevminoffset}$	An offset to the signaled $Q_{rxlevmin}$ taken into account in the *Srxlev* evaluation as a result of periodic search for a higher priority PLMN while camped normally in a VPLMN
$Q_{qualminoffset}$	Offset to the signaled $Q_{qualmin}$ taken into account in the *Squal* evaluation as a result of periodic search for a higher priority PLMN while camped normally in a VPLMN
$P_{compensation}$	If the UE supports the additional P_{max} in the *NR-NS-PmaxList*, if present, in *SIB1*, *SIB2*, and *SIB4*, $\max(P_{EMAX1} - P_{PowerClass}, 0) - [\min(P_{EMAX2}, P_{PowerClass}) - \min(P_{EMAX1}, P_{PowerClass})]$(dB) else $\max(P_{EMAX1} - P_{PowerClass}, 0)$(dB)
P_{EMAX1}, P_{EMAX2}	Maximum transmit-power level that a UE may use when transmitting in the uplink in the cell (dBm) is defined as P_{EMAX}. If the UE supports SUL frequency for this cell, P_{EMAX1}andP_{EMAX2} are obtained from the *p-Max* for SUL in *SIB1* and *NR-NS-PmaxList* for SUL, respectively, in *SIB1*, *SIB2*, and *SIB4*; otherwise, P_{EMAX1} and P_{EMAX2} are obtained from the *p-Max* and *NR-NS-PmaxList*, respectively, in *SIB1*, *SIB2*, and *SIB4* for regular uplink
$P_{PowerClass}$	Maximum RF output power of the UE (dBm) according to the UE power class

When evaluating *Srxlev* and *Squal* of non-serving cells for reselection evaluation purposes, the UE must use the parameters that are provided by the serving cell and for the verification of cell selection criterion, the UE must use the parameters provided by the target cell for cell reselection. The NAS can control the RAT(s) in which the cell selection should be performed, for instance by indicating RAT(s) associated with the selected PLMN, and by maintaining a list of forbidden registration area(s) and a list of equivalent PLMNs. The UE must select a suitable cell based on RRC_IDLE or RRC_INACTIVE state measurements and cell selection criteria. In order to expedite the cell selection process, the previously stored information for other RATs may be used by the UE. After camping on a cell, the UE is required to regularly search for a better cell according to the cell reselection criteria. If a better cell is found, the UE proceed to select that cell. The change of cell may imply a change of RAT.

The UE mobility state is determined, if the parameters (T_{CRmax}, N_{CR_H}, N_{CR_M}, and $T_{CRmaxHyst}$) are broadcast in SI for the serving cell. The **state detection criteria** are based on the following principles [12]:

- Normal-mobility state criterion: If the number of cell reselections during time period T_{CRmax} is less than N_{CR_M}.
- Medium-mobility state criterion: If the number of cell reselections during time period T_{CRmax} is greater than or equal to N_{CR_M} but less than or equal to N_{CR_M}.
- High-mobility state criterion: If the number of cell reselections during time period T_{CRmax} is greater than N_{CR_M}.

The UE must not attempt to make consecutive reselections, where a cell is reselected again immediately after one reselection for mobility state detection criteria.

The **state transitions** of the UE are determined based on the following criteria [12]:

- If the criteria for high-mobility state is detected, the UE must transition to the high-mobility state.
- If the criteria for medium-mobility state is detected, the UE must transition to medium--mobility state.
- If criteria for either medium- or high-mobility state is not detected during time period $T_{CRmaxHyst}$, the UE must transition to normal-mobility state.

If the UE is in high- or medium-mobility state, it must apply the speed-dependent scaling rules [12].

2.5.3 Paging

Paging allows the network to reach UEs in RRC_IDLE and RRC_INACTIVE states, and to notify the UEs in RRC_IDLE, RRC_INACTIVE, and RRC_CONNECTED states of SI change, as well as to send ETWS/CMAS notifications. While in RRC_IDLE state, the UE monitors the PCHs for CN-initiated paging. The UEs in RRC_INACTIVE state also monitor PCHs for RAN-initiated paging. To ensure limiting the adverse impact of paging procedure on the battery consumption, a UE does not need to continuously monitor the PCHs since NR defines a UE-specific paging DRX. The UE in RRC_IDLE or RRC_INACTIVE state is only required to monitor PCHs during paging occasions in a DRX cycle. The paging DRX cycles are configured by the network as follows [8]. For CN-initiated paging, a default cycle is broadcast in SI and also a UE-specific cycle is configured via NAS signaling. For RAN-initiated paging, a UE-specific cycle is configured via RRC signaling. The UE applies the shortest of the DRX cycles that are configured for it, that is, a UE in RRC_IDLE uses the shortest of the CN-initiated paging cycles, whereas a UE in RRC_INACTIVE applies the shortest of CN-initiated and RAN-initiated paging cycles (see Fig. 2.29).

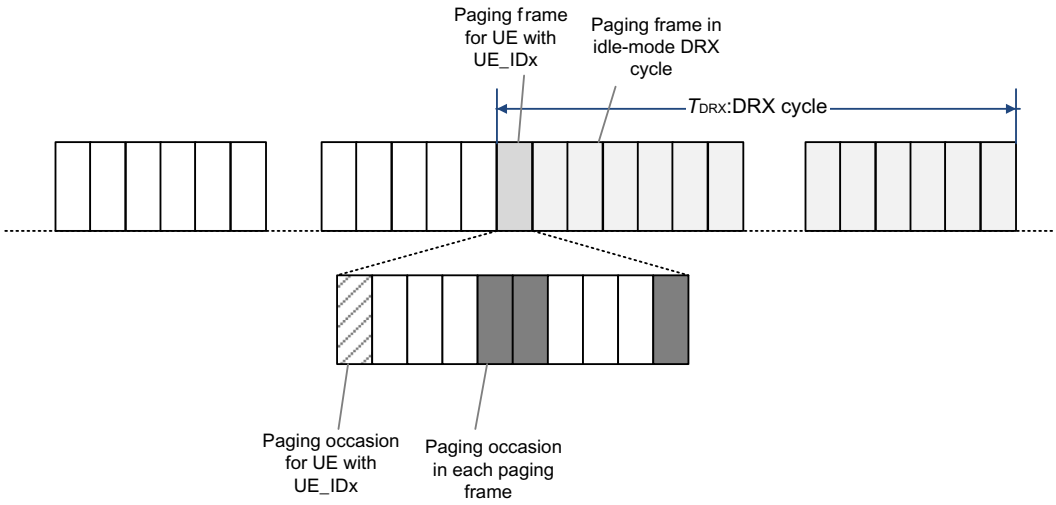


Figure 2.29
Example illustration of NR paging frames and paging occasions [12].

In CN-initiated and RAN-initiated paging, the POs of a UE are based on the same UE_ID, resulting in overlapping POs for both cases. The number of different POs in a DRX cycle is configurable via SI and a network may distribute UEs into those POs based on their UE_IDs. When in RRC_CONNECTED state, the UE monitors the PCHs in any PO signaled in the SI for SI change indication and public warning system notification. In the case of bandwidth adaptation, a UE in RRC_CONNECTED state only monitors PCHs on the active BWP with common search space configured (see Chapter 4).

To optimize the paging procedure for the UEs in CM-IDLE state, upon UE context release, the NG-RAN node may provide the AMF with a list of recommended cells and NG-RAN nodes as assistance information for subsequent paging of the UE. The AMF may further provide *paging attempt information* consisting of a *paging attempt count* and the *intended number of paging attempts* and possibly the *next paging area scope*. If *paging attempt information* is included in the paging message, each paged NG-RAN node receives the same information during a paging attempt. The *paging attempt count* is increased by one at each new paging attempt. The *next paging area scope*, when present, indicates whether the AMF intends to modify the paging area currently selected at the next paging attempt. If the UE has changed its state to CM-CONNECTED, the *paging attempt count* is reset [8].

To optimize the paging procedure for the UEs in RRC_INACTIVE state, upon RAN-initiated paging, the serving NG-RAN node provides the RAN paging area information. The serving NG-RAN node may also provide the RAN paging attempt information. Each paged NG-RAN node receives the same RAN paging attempt information during a paging attempt with the following content: *paging attempt count*, the intended number of paging attempts and the *next paging area scope*. The *paging attempt count* is increased by

one at each new paging attempt. The *next paging area scope*, when present, indicates whether the serving NG_RAN node plans to modify the RAN paging area currently selected for the next paging attempt. If the UE leaves RRC_INACTIVE state, the *paging attempt count* is reset [8].

When a UE is paged, the paging message is broadcast over a group of cells. In NR the basic principle of UE tracking is the same for idle and inactive modes, although the grouping is to some extent different in the two cases. The NR cells are grouped into RAN areas, where each RAN area is identified by a RAN area ID (RAI). The RAN areas are grouped into larger tracking areas, where each tracking area is identified by a tracking area ID (TAI). As a result, each cell belongs to one RAN area and one tracking area, the identities of which are provided as part of the cell SI. The tracking areas are the basis for device tracking at core-network level. Each device is assigned a UE registration area by the CN, consisting of a list of TAIs. When a device enters a cell that belongs to a tracking area not included in the assigned UE registration area, it accesses the network, including the CN, and performs a NAS registration update. The CN registers the device location and updates the device registration area and provides the device with a new TAI list that includes the new TAI. The reason that the device is assigned a set of TAIs is to avoid repeated NAS registration updates, every time that the device crosses the border of two neighbor tracking areas. By keeping the old TAI within the updated UE registration area, no new update is needed, if the device moves back to the old TAI. The RAN area is the basis for device tracking on RAN level. The UEs in the inactive mode can be assigned an RNA that consists of either a list of cell identities; a list of RAIs; or a list of TAIs [17].

The procedure for RNA update is similar to the update of the UE registration area. When a device enters a cell that is not included in its RNA, it accesses the network and performs an RNA update. The radio network registers the device location and updates the device RNA. Since the change of tracking area always implies the change of the device RAN area, an RNA update is implicitly performed every time a device performs a UE registration update. In order to track its movement within the network, the device searches for and measures SSBs similar to the initial cell search procedure. Once the device detects an SSB with a received power that exceeds the received power of its current SSB by a certain threshold, it detects the SIB1 of the new cell in order to acquire information about the tracking and RAN areas.

2.5.4 Measurements

The UE in RRC_CONNECTED state conducts measurements on multiple beams (at least one) of a cell and averages the measurement results mainly in the form of power values, in order to derive the cell quality. Therefore, the UE is configured to consider a subset of the detected beams. Filtering is applied at two different levels namely at the physical

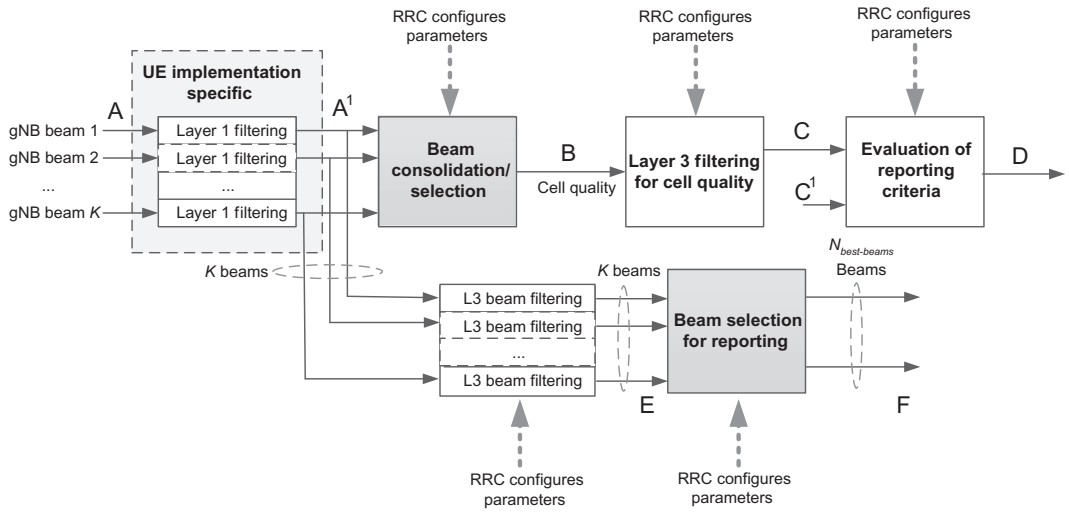


Figure 2.30
NR RRM measurement model [8].

layer to derive beam quality and then at RRC level to derive cell quality from multiple beams. The cell quality from beam measurements is derived in the same way for the serving cell(s) and for the non-serving cell(s). The measurement reports may contain the measurement results of the $N_{best-beams}$ best beams, if the UE is configured by the gNB. The K beams correspond to the measurements on SSB or CSI-RS resources configured for layer 3 mobility by gNB and detected by UE at layer 1.

The corresponding high-level measurement model is illustrated in Fig. 2.30 where the parameters can be described as follows [8]:

- A_i: The measurements (beam-specific samples) internal to the physical layer.
- *Layer 1 filtering*: Internal layer 1 filtering of the inputs measured at point A. The exact filtering function is implementation specific and the way that the measurements are conducted in the physical layer by an implementation (inputs A_i and layer 1 filtering) in not specified by the standard.
- A_i^1: The measurements (i.e., beam-specific measurements) reported by layer 1 to layer 3 after layer 1 filtering.
- *Beam consolidation/selection*: The beam-specific measurements are consolidated to derive cell quality. The behavior of the beam consolidation/selection is standardized, and the configuration of this module is provided via RRC signaling. The reporting period at B equals one measurement period at A^1.
- B: A measurement (i.e., cell quality) derived from beam-specific measurements reported to layer 3 after beam consolidation/selection.

- *Layer 3 filtering for cell quality*: The filtering performed on the measurements provided at point B. The behavior of the layer 3 filters is standardized and the configuration of the layer 3 filters is provided via RRC signaling. The filtering reporting period at C equals one measurement period at B.
- *C*: A measurement after processing in the layer 3 filter. The reporting rate is identical to the reporting rate at point B. This measurement is used as input for one or more evaluation of reporting criteria.
- *Evaluation of reporting criteria*: It checks whether actual measurement reporting is necessary at point D. The evaluation can be based on more than one flow of measurements at reference point C, for example, to compare different measurements. This is illustrated by inputs C and C^1. The UE evaluates the reporting criteria at least every time a new measurement result is reported at points C and C^1. The reporting criteria are standardized, and the configuration is provided via RRC signaling (UE measurements).
- *D*: The measurement report information (message) sent on the radio interface.
- *L3 beam filtering*: The filtering performed on the measurements (i.e., beam-specific measurements) provided at point A^1. The behavior of the beam filters is standardized, and the configuration of the beam filters is provided via RRC signaling. The filtering reporting period at E equals one measurement period at A^1.
- *E*: A measurement (i.e., beam-specific measurement) after processing in the beam filter. The reporting rate is identical to the reporting rate at point A^1. This measurement is used as input for selecting the $N_{best-beams}$ measurements to be reported.
- *Beam selection for beam reporting*: It selects the $N_{best-beams}$ measurements from the measurements provided at point E. The behavior of the beam selection is standardized, and the configuration of this module is provided via RRC signaling.
- *F*: The beam measurement information included in measurement report sent over the radio interface.

Layer 1 filtering introduces a certain level of measurement averaging and the manner through which the UE performs the required measurements is implementation specific to the point that the output at B fulfills the performance requirements. The layer 3 filtering function for cell quality and the related parameters do not introduce any delay in the sample availability at points B and C in Fig. 2.30. The measurements at points C and C^1 are the input used in the event evaluation. The L3 beam filtering and the related parameters do not cause any delay in the sample availability at points E and F. The measurement reports are characterized by the following criteria [8]:

- Measurement reports include the measurement identity of the associated measurement configuration that triggered the reporting.
- Cell and beam measurement quantities to be included in measurement reports are configured by the network.

- The number of non-serving cells to be reported can be limited through configuration by the network.
- Cells belonging to a blacklist configured by the network are not used in event evaluation and reporting, and conversely when a whitelist is configured by the network, only the cells belonging to the whitelist are used in event evaluation and reporting.
- Beam measurements to be included in measurement reports are configured by the network (beam ID only, measurement result and beam ID, or no beam reporting).

The intra-frequency neighbor (cell) measurements and inter-frequency neighbor (cell) measurements are defined as follows [8]:

- *SSB-based intra-frequency measurement*: A measurement is defined as an SSB-based intra-frequency measurement provided that the center frequency of the SSB of the serving cell and the center frequency of the SSB of the neighbor cell are the same, and the subcarrier spacing of the two SSBs are also the same.
- *SSB-based inter-frequency measurement*: A measurement is defined as an SSB-based inter-frequency measurement provided that the center frequency of the SSB of the serving cell and the center frequency of the SSB of the neighbor cell are different, or the subcarrier spacing of the two SSBs are different. It must be noted that for SSB-based measurements, one measurement object corresponds to one SSB and the UE considers different SSBs as different cells.
- *CSI-RS-based intra-frequency measurement*: A measurement is defined as a CSI-RS-based intra-frequency measurement, if the bandwidth of the CSI-RS resource on the neighbor cell configured for measurement is within the bandwidth of the CSI-RS resource on the serving cell configured for measurement, and the subcarrier spacing of the two CSI-RS resources is the same.
- *CSI-RS-based inter-frequency measurement*: A measurement is defined as a CSI-RS-based inter-frequency measurement, if the bandwidth of the CSI-RS resource on the neighbor cell configured for measurement is not within the bandwidth of the CSI-RS resource on the serving cell configured for measurement, or the subcarrier spacing of the two CSI-RS resources are different.

A measurement can be non-gap-assisted or gap-assisted depending on the capability of the UE, the active BWP of the UE and the current operating frequency, described as follows [8]:

- For an SSB-based inter-frequency measurement, a measurement gap configuration is always provided if the UE only supports per-UE measurement gaps or if the UE supports per-FR measurement gaps and any of the configured BWP frequencies of any of the serving cells are in the same FR of the measurement object.

- For an SSB-based intra-frequency measurement, a measurement gap configuration is always provided in the case where, other than the initial BWP, if any of the UE configured BWPs do not contain the frequency domain resources of the SSB associated with the initial DL BWP.

In non-gap-assisted scenarios, the UE must be able to conduct measurements without measurement gaps. In gap-assisted scenarios, the UE cannot be assumed to be able to conduct measurements without measurement gaps.

2.6 UE and Network Identifiers

An NR UE in the connected mode uses a number of network-assigned temporary identifiers in order to communicate to gNB and 5GC. Those identifiers, their descriptions, and their usage are summarized as follows [8]:

- C-RNTI: A unique UE identification used as an identifier of the RRC connection and for scheduling purposes.
- CS-RNTI: A unique UE identification used for SPS in the downlink or configured grant in the uplink.
- INT-RNTI: An identification of preemption in the downlink.
- P-RNTI: An identification of paging and SI change notification in the downlink.
- SI-RNTI: An identification of broadcast and SI in the downlink.
- SP-CSI-RNTI: Unique UE identification used for semi-persistent CSI reporting on PUSCH.

The following identifiers are used for power and slot format control [8]:

- SFI-RNTI: An identification of slot format.
- TPC-PUCCH-RNTI: A unique UE identification to control the power of PUCCH.
- TPC-PUSCH-RNTI: A unique UE identification to control the power of PUSCH.
- TPC-SRS-RNTI: A unique UE identification to control the power of SRS.

The following identities are used during random-access procedure [8]:

- RA-RNTI: An identification of the RAR message in the downlink.
- Temporary C-RNTI: A UE identification temporarily used for scheduling during the random-access procedure.
- Random value for contention resolution: A UE identification temporarily used for contention resolution purposes during the random-access procedure.

The following identities are used at NG-RAN level by an NR UE connected to 5GC:

- I-RNTI: An ID used to identify the UE context in RRC_INACTIVE state.

Table 2.5: Radio network temporary identifiers (RNTIs) in NR and their usage [9].

RNTI	Usage	Transport Channel	Logical Channel
P-RNTI	Paging and system information change notification	PCH	PCCH
SI-RNTI	Broadcast of system information	DL-SCH	BCCH
RA-RNTI	Random-access response	DL-SCH	N/A
Temporary C-RNTI	Contention resolution (when no valid C-RNTI is available)	DL-SCH	CCCH
Temporary C-RNTI	Msg3 transmission	UL-SCH	CCCH, DCCH, DTCH
C-RNTI, MCS-C-RNTI	Dynamically scheduled unicast transmission	UL-SCH	DCCH, DTCH
C-RNTI	Dynamically scheduled unicast transmission	DL-SCH	CCCH, DCCH, DTCH
MCS-C-RNTI	Dynamically scheduled unicast transmission	DL-SCH	DCCH, DTCH
C-RNTI	Triggering of PDCCH ordered random access	N/A	N/A
CS-RNTI	Configured scheduled unicast transmission (activation, reactivation, and retransmission)	DL-SCH, UL-SCH	DCCH, DTCH
CS-RNTI	Configured scheduled unicast transmission (deactivation)	N/A	N/A
TPC-PUCCH-RNTI	PUCCH power control	N/A	N/A
TPC-PUSCH-RNTI	PUSCH power control	N/A	N/A
TPC-SRS-RNTI	SRS trigger and power control	N/A	N/A
INT-RNTI	Indication of preemption in the downlink	N/A	N/A
SFI-RNTI	Slot format indication in a given cell	N/A	N/A
SP-CSI-RNTI	Activation of semi-persistent CSI reporting on PUSCH	N/A	N/A

A complete list of UE RNTIs is provided in Table 2.5.

The following identities are used in NG-RAN for identifying a specific network entity [8]:

- AMF name: This is used to identify an AMF.
- NCGI: This is used to globally identify the NR cells. The NCGI is constructed from the PLMN identity to which the cell belongs to and the NR cell identity (NCI) of the cell.
- gNB ID: This is used to identify the gNBs within a PLMN. The gNB ID is contained within the NCI of its cells.
- Global gNB ID: This is used to globally identify the gNBs. The global gNB ID is constructed from the PLMN identity to which the gNB belongs and the gNB ID. The MCC and MNC are the same as included in the NCGI.
- Tracking area identity (TAI): This is used to identify tracking areas. The TAI is constructed from the PLMN identity the tracking area belongs to and the tracking area code of the tracking area.

- Single network slice selection assistance information (S-NSSAI): This is used to identify a network slice.

2.7 Random-Access Procedure (L2/L3 Aspects)

The random-access procedure is triggered by a number of events including initial access from RRC_IDLE state; RRC connection reestablishment procedure, handover, downlink/uplink data arrival when in RRC_CONNECTED state and if uplink synchronization status is *non-synchronized*; uplink data arrival while in RRC_CONNECTED state and when there are no available PUCCH resources for SR; SR failure; request by RRC upon synchronous reconfiguration; transition from RRC_INACTIVE state; establishing timing alignment upon SCell addition; request for other SI; and beam failure recovery. Furthermore, the random-access procedure takes two distinct forms: contention-based random access and contention-free random access as shown in Fig. 2.31. For random access in a cell configured with SUL, the network can explicitly signal which carrier to use (uplink or SUL); otherwise, the UE selects the SUL carrier, if the measured quality of the downlink is lower than a broadcast threshold. Once started, all uplink transmissions of the random-access procedure remain on the selected carrier [8]. The complete description of the random-access procedure including the physical layer aspects can be found in Chapter 3.

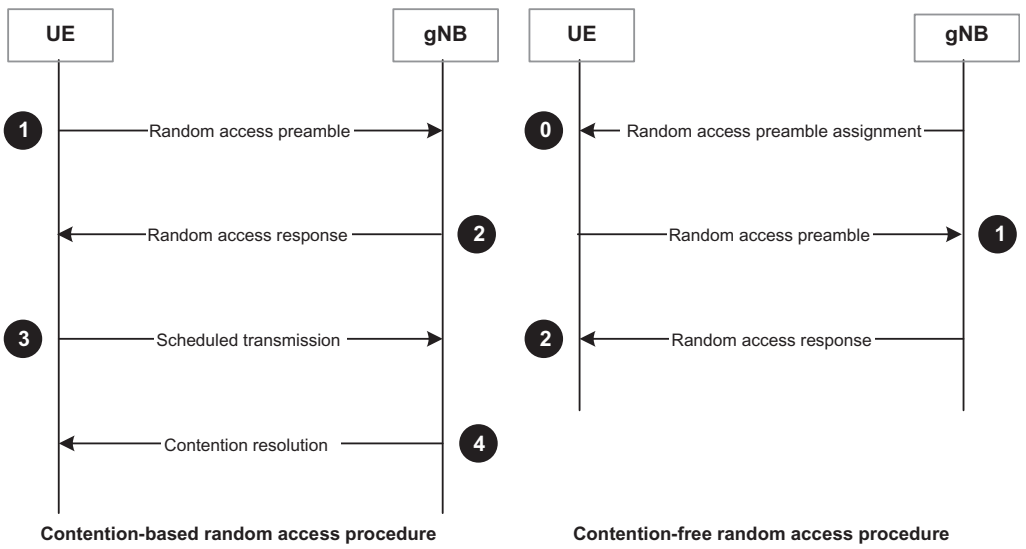


Figure 2.31
Contention-based and contention-free random-access procedures [8].

2.8 Multi-radio Dual Connectivity (L2/L3 Aspects)

MR-DC is a generalization of the intra-E-UTRA dual connectivity. The NG-RAN supports MR-DC operation wherein a UE in RRC_CONNECTED state is configured to utilize radio resources provided by (at least) two distinct schedulers, located in two different NG-RAN nodes connected via a non-ideal backhaul, one providing the NR access and the other providing either E-UTRA or NR access. In MR-DC, one network node acts as the master node (MN) and the other as the secondary node (SN). The MN and SN entities are connected via a network interface and the MN is connected to the CN. The NR MR-DC scheme is designed based on the assumption of non-ideal backhaul between different nodes but can also be used in the case of ideal backhaul. The LTE network supports MR-DC via EN-DC, in which a UE is connected to one eNB that acts as the MN and one en-gNB[11] that acts as a SN. The eNB is connected to the EPC via the S1 interface and to the en-gNB via the X2 interface. The en-gNB may also be connected to the EPC via the S1-U interface and other en-gNBs via the X2-U interface [4].

The NG-RAN supports NG-RAN EN-DC (NGEN-DC), in which a UE is connected to one ng-eNB that acts as the MN and one gNB that acts as an SN. The ng-eNB is connected to the 5GC and the gNB is connected to the ng-eNB via the Xn interface. The NG-RAN further supports NR-E-UTRA DC (NE-DC), in which a UE is connected to one gNB that acts as the MN and one ng-eNB that acts as an SN. The gNB is connected to 5GC and the ng-eNB is connected to the gNB via the Xn interface. Another important scenario is NR−NR dual connectivity (NR-DC), in which a UE is connected to one gNB that acts as the MN and another gNB that acts as the SN. The master gNB is connected to the 5GC via the NG interface and to the secondary gNB via the Xn interface. The secondary gNB may also be connected to the 5GC via the NG-U interface. In addition, NR-DC can also be used when a UE is connected to two gNB-DUs, one serving the MCG and the other serving the SCG, connected to the same gNB-CU, acting both as the MN and the SN. When the UE is configured with SCG, it is configured with two MAC entities: one MAC entity for the MCG and one MAC entity for the SCG [4].

In MR-DC, the UE has a single RRC state, based on the MN RRC and a single control-plane connection toward the CN. Each radio node has its own RRC entity (LTE version, if the node is an eNB or NR version if the node is a gNB) which can generate RRC PDUs to be sent to the UE. The RRC PDUs generated by the SN can be transported via the MN to the UE. The MN always sends the initial SN RRC configuration via MCG SRB (SRB1); however, subsequent reconfigurations may be transported via MN or SN. When

[11] en-gNB is a node providing NR user-plane and control-plane protocol terminations toward the UE and acts as a secondary node in EN-DC. A secondary node in MR-DC is a radio access node, with no control-plane connection to the core network, providing only additional radio resources to the UE. It may be an en-gNB (in EN-DC), a secondary ng-eNB (in NE-DC), or a secondary gNB (in NR-DC and NGEN-DC) [4].

transporting RRC PDU from the SN, the MN does not modify the UE configuration provided by the SN [4].

When an LTE node is connected to the EPC, upon initial connection establishment, SRB1 uses LTE PDCP. If the UE supports EN-DC, regardless of whether EN-DC is configured, after initial connection establishment, the MCG SRBs (SRB1 and SRB2) can be configured by the network to use either LTE PDCP or NR PDCP (either SRB1 and SRB2 are both configured with LTE PDCP, or they are both configured with NR PDCP). A change from LTE PDCP to NR PDCP (or vice versa) is supported via a handover procedure (reconfiguration with mobility) or, for the initial change of SRB1 from LTE PDCP to NR PDCP, with a reconfiguration without mobility before the initial security activation.

If the SN is a gNB (i.e., the case for EN-DC, NGEN-DC, and NR-DC scenario), the UE can be configured to establish an SRB with the SN (SRB3) to enable RRC PDUs for the SN to be directly transferred between the UE and the SN. The RRC PDUs for the SN can only be transported directly to the UE for SN RRC reconfiguration without any coordination with the MN. Measurement reporting for mobility within the SN can be sent directly from the UE to the SN, if SRB3 is configured. The split SRB is supported for all MR-DC options, allowing duplication of RRC PDUs generated by the MN via the direct path and via the SN. The split SRB utilizes the NR PDCP. The NR Rel-15 specifications do not support duplication of RRC PDUs generated by the SN via the MN and SN paths. In EN-DC, the SCG configuration is maintained in the UE during suspension. The UE releases the SCG configuration (but not the radio bearer configuration) during resumption initiation (see Fig. 2.24). In MR-DC with 5GC, the UE stores the PDCP/SDAP configuration when moving to RRC_INACTIVE state, but it releases the SCG configuration [4].

There are three bearer types in MR-DC from a UE perspective: MCG bearer, SCG bearer, and split bearer. For EN-DC, the network can configure either LTE PDCP or NR PDCP for the MN-terminated MCG bearers, while NR PDCP is always used for all other bearers. In MR-DC with 5GC, NR PDCP is always used for all bearer types. In NGEN-DC, LTE RLC/MAC is used in the MN, while NR RLC/MAC is used in the SN. In EN-DC, NR RLC/MAC is used in the MN while LTE RLC/MAC is used in the SN. In NR-DC, NR RLC/MAC is used in both MN and SN. From the network perspective, each bearer (MCG, SCG, and split bearer) can be terminated either in MN or in SN. If only SCG bearers are configured for a UE, for SRB1 and SRB2, the logical channels are always configured at least in the MCG, that is, this is still an MR-DC configuration and a PCell always exists. If only MCG bearers are configured for a UE, that is, there is no SCG, this is still considered an MR-DC configuration, if at least one of the bearers is terminated in the SN [4].

In MR-DC, two or more component carriers may be aggregated over two cell groups. A UE may simultaneously receive or transmit on multiple component carriers depending on its capabilities. The maximum number of configured component carriers for a UE is 32 for

downlink and uplink. Depending on UE's capabilities, up to 31 component carriers can be configured for an LTE cell group when the NR cell group is configured. For the NR cell group, the maximum number of configured component carriers for a UE is 16 for downlink and 16 for uplink. A gNB may configure the same physical cell IDs (PCIs) for several NR cells that it serves. To avoid PCI confusion for MR-DC, the NR PCIs may be allocated in a way that an NR cell is uniquely identifiable by a PCell ID. This PCell is in the coverage area of an NR cell included in the MR-DC operation. In addition, the NR PCIs may only be reused in NR cells on the same SSB frequency sufficiently apart from each other. An X2-C/Xn-C signaling can be used to help identify NR PCIs by including the cell global identifier (CGI) of the PCell in the respective X2AP/XnAP messages and by providing neighbor cell relationship via non-UE-associated signaling [4].

In MR-DC, the UE is configured with two MAC entities: one MAC entity for the MCG and one MAC entity for the SCG. In MR-DC, SPS resources can be configured on both PCell and PSCell. In MR-DC, the BSR configuration, triggering, and reporting are independently performed per cell group. For split bearers, the PDCP data is considered in BSR in the cell group(s) configured by RRC signaling. In EN-DC, separate DRX configurations are provided for MCG and SCG. Both RLC-AM and RLC-UM can be configured in MR-DC for all bearer types (i.e., MCG, SCG, and split bearers). In EN-DC, packet duplication can be applied to carrier-aggregation in the MN and in the SN; however, MCG bearer carrier-aggregation packet duplication can be configured only in combination with LTE PDCP; and MCG DRB carrier-aggregation duplication can be configured only if dual-connectivity packet duplication is not configured for any split DRB. In NGEN-DC, carrier-aggregation packet duplication can only be configured for SCG bearer. In NE-DC, carrier-aggregation packet duplication can only be configured for MCG bearer. In NR-DC, carrier-aggregation packet duplication can be configured for both MCG and SCG bearers. In EN-DC, ROHC can be configured for all bearer types. In MR-DC with 5GC, the network may host up to two SDAP protocol entities for each PDU session, one for MN and the other one for SN. The UE is configured with one SDAP protocol entity per PDU session [4].

In MR-DC, the SN is not required to broadcast SI other than for radio frame timing and SFN. The SI for initial configuration is provided to the UE by dedicated RRC signaling via the MN. The UE acquires radio frame timing and SFN of SCG from the LTE primary and secondary synchronization signals and MIB (if the SN is an eNB) and from the NR primary and secondary synchronization signals and MIB (if the SN is a gNB) of the PSCell. Moreover, upon change of the relevant SI of a configured SCell, the network releases and subsequently adds the corresponding SCell (with updated SI), via one or more RRC reconfiguration messages sent on SRB1 or SRB3. If the measurement is configured for the UE in preparation for the secondary node addition procedure, the MN may configure the measurement for the UE. In the case of the intra-secondary node mobility, the SN may configure the measurement for the UE in coordination with the MN [4].

The secondary node change procedure can be triggered by both MN (only for inter-frequency secondary node change) and SN. For secondary node changes triggered by the SN, the RRM measurement configuration is maintained by the SN which also processes the measurement reporting, without providing the measurement results to the MN. Measurements can be configured independently by the MN and by the SN (intra-RAT measurements on serving and non-serving frequencies). The MN indicates the maximum number of frequency layers and measurement identities that can be used in the SN to ensure that UE capabilities are not exceeded. If MN and SN both configure measurements on the same carrier frequency then those configurations must be consistent. Each node (MN or SN) can independently configure a threshold for the SpCell quality. When the PCell quality is above the threshold configured by the MN, the UE is still required to perform inter-RAT measurements configured by the MN on the SN RAT. When SpCell quality is above the threshold configured by the SN, the UE is not required to perform measurements configured by the SN [4].

The measurement reports, configured by the SN, are sent on SRB1 when SRB3 is not configured; otherwise, the measurement reports are sent over SRB3. The measurement results related to the target SN can be provided by MN to target SN at MN-initiated SN change procedure. The measurement results of target SN can be forwarded from the source SN to the target SN via MN at SN-initiated SN change procedure. The measurement results corresponding to the target SN can be provided by the source MN to the target MN at inter-MN handover with/without SN change procedure [4].

Per-UE or per-FR measurement gaps can be configured, depending on UE capability to support independent FR measurement and network preference. Per-UE gap applies to both FR1 (LTE and NR) and FR2 (NR) bands. For per-FR gap, two independent gap patterns (i.e., FR1 gap and FR2 gap) are configured for FR1 and FR2. The UE may also be configured with a per-UE gap sharing configuration (applying to per-UE gap) or with two separate gap sharing configurations (applying to FR1 and FR2 measurement gaps, respectively). A measurement gap configuration is always provided in the following scenarios: for UEs configured with LTE inter-frequency measurements; and for UEs that support either per-UE or per-FR gaps, when the conditions to measure SSB-based inter-frequency measurement or SSB-based intra-frequency measurement are
satisfied [4].

2.9 Carrier Aggregation (L2/L3 Aspects)

Multiple NR component carriers can be aggregated and simultaneously transmitted to a UE in the downlink or from a UE in the uplink, allowing an increased operating bandwidth and correspondingly higher link data rates. The component carriers do not need to be contiguous in

the frequency domain and can be in the same frequency band or different frequency bands, resulting in three scenarios: intra-band carrier aggregation with frequency-contiguous component carriers, intra-band carrier aggregation with non-contiguous component carriers, and inter-band carrier aggregation with non-contiguous component carriers. While the system-level operation for the three scenarios is the same, the architecture and complexity of RF transceivers can be very different. The NR supports up to 16 downlink/uplink carriers of different bandwidths and different duplex schemes, with the minimum and maximum contiguous bandwidth of 5 and 400 MHz per component carrier, respectively [5,21].

A UE capable of carrier aggregation may receive or transmit simultaneously on multiple component carriers, while a device not capable of carrier aggregation can access one of the component carriers at any given time. In the case of inter-band carrier aggregation of multiple half-duplex TDD carriers (supplemental uplink or downlink), the transmission direction of different carriers does not necessarily have to be the same, which implies that a carrier-aggregation-capable TDD device may need a front-end duplexer, unlike a typical carrier-aggregation-incapable TDD device that does not include a duplexer. In the LTE and NR specifications, the carrier aggregation is treated as a cell, that is, a carrier-aggregation-capable UE is said to able transmit/receive to/from multiple cells. One of these cells is referred to as the PCell that is the cell which the device initially selects and connects to. Once connected to the gNB, one or more SCells can be configured. The SCells can be activated or deceived to meet various application requirements. Different UEs may have different designated cells as their PCell, meaning that the configuration of the PCell is UE-specific. Furthermore, the number of carriers (or cells) does not have to be the same in uplink and downlink. In a typical scenario, there are more downlink carriers than uplink carriers, since there is often more traffic in the downlink than in the uplink. Furthermore, the RF implementation complexity and cost of operating multiple simultaneously active uplink carriers are often higher than the corresponding complexity/cost of the downlink. The scheduling grants and radio resource assignments can be transmitted on either the same cell as the corresponding data, referred to as self-scheduling, or on a different cell than the corresponding data, referred to as cross-carrier scheduling.

The NR carrier aggregation uses L1/L2 control signaling for scheduling the UE in the downlink, and uplink control signaling to transmit HARQ-ACKs. The uplink feedback is typically transmitted on the PCell to allow asymmetric carrier aggregation. In certain use cases where there are a large number of downlink component carriers and a single uplink component carrier, the uplink carrier would be overloaded with a large number of feedback information. To avoid overloading a single carrier, it is possible to configure two PUCCH groups where the feedback corresponding to the first group is transmitted in the uplink of the PCell and the feedback corresponding to the other group of carriers is transmitted on the PSCell. If carrier aggregation is enabled, the UE may receive and transmit on multiple

carriers, but operating multiple carriers is only needed for high data rates, thus is advantageous to deactivate unused carriers. Activation and deactivation of component carriers can be done through MAC CEs (see Section 2.2.1), where a bitmap is used to indicate whether a configured SCell should be activated or deactivated.

As we mentioned earlier, to ensure reasonable UE power consumption when carrier aggregation is configured, an activation/deactivation mechanism of cells is supported. When an SCell is deactivated, the UE no longer needs to receive the corresponding PDCCH or PDSCH, it cannot transmit in the corresponding uplink, it is not required to perform CQI measurements on that cell. On the other hand, when an SCell is activated, the UE receives PDSCH and PDCCH (if the UE is configured to monitor PDCCH on this SCell) and is expected to be able to perform CQI measurements on that cell. The NG-RAN ensures that while PUCCH SCell (i.e., an SCell configured with PUCCH) is deactivated, SCells of the secondary PUCCH group (i.e., a group of SCells whose PUCCH signaling is associated with the PUCCH on the PUCCH SCell) are activated. The NG-RAN further ensures that SCells mapped to PUCCH SCell are deactivated before the PUCCH SCell is changed or removed. When reconfiguring the set of serving cells, SCells added to the set are initially deactivated and SCells which remain in the set (either unchanged or reconfigured) do not change their activation status. During handover, the SCells are deactivated. When bandwidth adaptation is configured, only one UL BWP for each uplink carrier and one DL BWP or only one DL/UL BWP pair can be active at any given time in an active serving cell, all other BWPs that the UE is configured with will be deactivated. The UE does not monitor the PDCCH and does not transmit on PUCCH, PRACH, and UL-SCH of the deactivated BWPs [8].

The SCell activation/deactivation is an efficient mechanism to reduce UE power consumption in addition to DRX. On a deactivated SCell, the UE neither receives downlink signals nor transmits any uplink signal. The UE is also not required to perform measurements on a deactivated SCell. Deactivated SCells can be used as pathloss reference for measurements in uplink power control. It is assumed that these measurements would be less frequent while the SCell is deactivated in order to conserve the UE power. On the other hand, for an activated SCell, the UE performs normal activities for downlink reception and uplink transmission. Activation and deactivation of SCells is controlled by the gNB. As shown in Fig. 2.32, the SCell activation/deactivation is performed when the gNB sends an activation/deactivation command in the form of a MAC CE. A timer may also be used for automatic deactivation, if no data or PDCCH messages are received on a SCell for a certain period of time. This is the only case in which deactivation can be executed autonomously by the UE. Serving cell activation/deactivation is performed independently for each SCell, allowing the UE to be activated only on a particular set of SCells. Activation/deactivation is not applicable to the PCell because it is required to always remain activated when the UE has an RRC connection to the network [8].

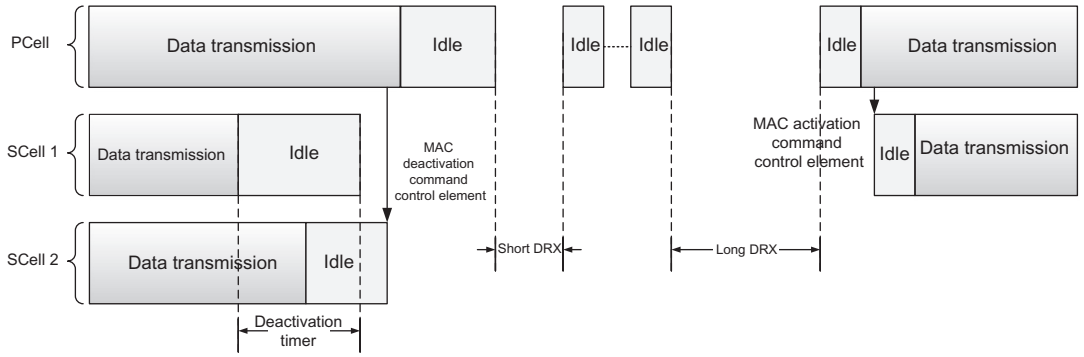


Figure 2.32
Illustration of SCell activation/deactivation procedure [9].

As already mentioned, if the UE is configured with one or more SCells, the network may activate and deactivate the configured SCells. The PCell is always activated. The network activates and deactivates the SCell(s) by sending an activation/deactivation MAC CE described in Section 2.2.1. Furthermore, the UE maintains a *sCellDeactivationTimer* timer per configured SCell (except the SCell configured with PUCCH) and deactivates the associated SCell upon its expiration. The same initial timer value is applied to each instance of the *sCellDeactivationTimer* and it is configured by RRC signaling. The configured SCells are initially deactivated upon addition and after a handover. The HARQ feedback for the MAC PDU containing SCell activation/deactivation MAC CE is not impacted by PCell, PSCell, and PUCCH SCell interruptions due to SCell activation/deactivation [9].

References

3GPP Specifications[12]

[1] 3GPP TS 23.501, System Architecture for the 5G System (Release 15), December 2018.
[2] 3GPP TS 33.501, Security Architecture and Procedures for 5G System (Release 15), December 2018.
[3] 3GPP TS 37.324, NR, Service Data Adaptation Protocol (SDAP) Specification (Release 15), September 2018.
[4] 3GPP TS 37.340, Multi-connectivity, Stage 2 (Release 15), December 2018.
[5] 3GPP TS 38.104, NR, Base Station (BS) Radio Transmission and Reception (Release 15), December 2018.
[6] 3GPP TS 38.133, NR, Requirements for Support of Radio Resource Management (Release 15), December 2018.
[7] 3GPP TS 38.215, NR, Physical Layer Measurements (Release 15), March 2018.
[8] 3GPP TS 38.300, NR, Overall Description, Stage-2 (Release 15), December 2018.
[9] 3GPP TS 38.321, NR, Medium Access Control (MAC) Protocol Specification (Release 15), December 2018.
[10] 3GPP TS 38.322, NR, Radio Link Control (RLC) Protocol Specification (Release 15), December 2018.

[12] 3GPP specifications can be accessed at http://www.3gpp.org/ftp/Specs/archive/.

[11] 3GPP TS 38.323, NR, Packet Data Convergence Protocol (PDCP) Specification (Release 15), December 2018.
[12] 3GPP TS 38.304, NR, User Equipment (UE) Procedures in Idle Mode and RRC Inactive State (Release 15), December 2018.
[13] 3GPP TS 38.306, NR, User Equipment (UE) Radio Access Capabilities (Release 15), December 2018.
[14] 3GPP TS 38.331, NR, Radio Resource Control (RRC); Protocol Specification (Release 15), December 2018.
[15] 3GPP TS 38.401, NG-RAN, Architecture Description (Release 15), December 2018.

IETF Specifications[13]

[16] IETF RFC 5795, The RObust Header Compression (ROHC) Framework, March 2010.

Articles, Books, White Papers, and Application Notes

[17] E. Dahlman, S. Parkvall, 5G NR: The Next Generation Wireless Access Technology, Academic Press, August 2018.
[18] S. Ahmadi, LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies, Academic Press, November 2013.
[19] 3GPP RWS-180010, NR Radio Interface Protocols, Workshop on 3GPP Submission Towards IMT-2020, Brussels, Belgium, October 2018.
[20] 5G New Radio, ShareTechNote. <http://www.sharetechnote.com>.
[21] MediaTek, A New Era for Enhanced Mobile Broadband, White Paper, March 2018.

[13] IETF specifications can be accessed at https://datatracker.ietf.org/.