CHAPTER

SCHEDULING AND RATE ADAPTATION

9

Scheduling is a central part of an LTE system. For each time instant, the scheduler determines to which user(s) the shared time−frequency resource should be assigned and determines the data rate to use for the transmission. The scheduler is a key element and to a large degree determines the overall behavior of the system. Both uplink and downlink transmissions are scheduled and, consequently, there is a downlink and an uplink scheduler in the eNodeB.

The *downlink scheduler* is responsible for dynamically controlling the device(s) to transmit to. Each of the scheduled devices is provided with a *scheduling assignment* consisting of the set of resource blocks upon which the device's DL-SCH[1] is transmitted, and the associated transport-format. The basic mode of operation is the so-called *dynamic* scheduling, where the eNodeB for each 1 ms TTI conveys scheduling assignments to the selected devices using the (E)PDCCHs as described in Chapter 6, but there is also a possibility for *semi-persistent* scheduling to reduce the control-signaling overhead. Downlink scheduling assignments and logical channel multiplexing are controlled by the eNodeB as illustrated in the left part of Figure 9.1.

The *uplink scheduler* serves a similar purpose, namely to dynamically control which devices are to transmit on their UL-SCH. Similarly to the downlink case, each scheduled device is provided with a *scheduling grant* consisting of the set of resource blocks upon which the device should transmit its UL-SCH and the associated transport-format. Also in this case, either dynamic or semi-persistent scheduling can be used. The uplink scheduler is in complete control of the transport format the device shall use but, unlike the downlink case, not the logical-channel multiplexing. Instead, the logical-channel multiplexing is controlled by the device according to a set of rules. Thus, uplink scheduling is *per device* and not per radio bearer. This is illustrated in the right part of Figure 9.1, where the scheduler controls the transport format and the device controls the logical channel multiplexing.

The following sections describe the details in the LTE scheduling framework after a brief review of basic scheduling principles.

[1] In case of carrier aggregation there is one DL-SCH (or UL-SCH) per component carrier.

4G, LTE-Advanced Pro and The Road to 5G. http://dx.doi.org/10.1016/B978-0-12-804575-6.00009-1

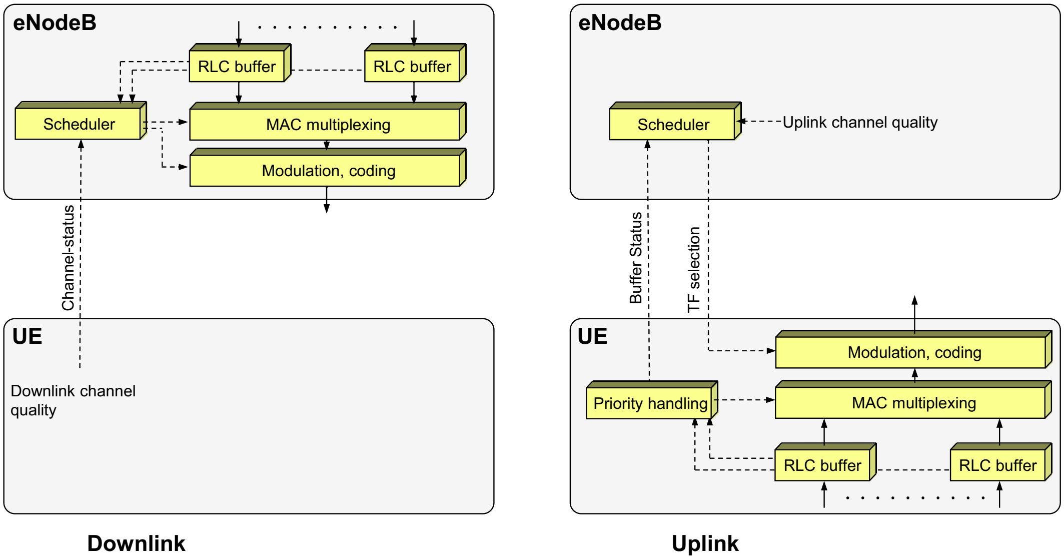

FIGURE 9.1

Transport format selection in downlink (left) and uplink (right).

9.1 SCHEDULING STRATEGIES

The scheduling strategy in LTE is not standardized but is a base-station-implementation issue—and an important one as the scheduler is a key element in LTE and to a large extent defines the overall behavior. Different vendors may choose different strategies in various scenarios to match the user needs. What is standardized is the supporting functions for scheduling such as transmission of scheduling grants, quality-of-service mechanisms, and various feedback information, for example channel-state reports and buffer-status reports. However, there are some basic scheduling strategies in the literature, useful to illustrate the principles.

For the purpose of illustrating the principles, consider time-domain-only scheduling with a single user being scheduled at a time and all users having an infinite amount of data to transmit. In this case, the utilization of the radio resources is maximized if, at each time instant, all resources are assigned to the user with the best instantaneous channel condition. Together with rate control, this implies that the highest data rate is achieved for a given transmit power or, in other words, for a given interference to other cells, the highest link utilization is achieved. Rate control is more efficient compared to power control, which adjusts the transmission power to follow the channel variations while keeping the data rate constant [11,12]. This scheduling strategy is an example of channel-dependent scheduling known as *max*-C/I (or *maximum rate*) scheduling. Since the radio conditions for the different radio links within a cell typically vary independently, at each point in time there is almost always a radio link whose channel quality is near its peak and supporting a correspondingly high data rate. This translates into a high system capacity. The gain obtained by transmitting to users with favorable radio-link conditions is commonly known as multi-user diversity; the multi-user diversity gains are larger, the larger the channel variations and the larger the number of users in a cell. Hence, in contrast to the traditional view that rapid variations in the radio-link quality is an undesirable effect that has to be combated, the possibility of channel-dependent scheduling implies that *rapid variations are in fact potentially beneficial and should be exploited*.

Mathematically, the max-C/I scheduler can be expressed as scheduling user k given by

$$k = \arg\max_i R_i$$

where R_i is the instantaneous data rate for user i. Although, from a system capacity perspective, max-C/I scheduling is beneficial, this scheduling principle will not be fair in all situations. If all devices are, on average, experiencing similar channel conditions and the variations in the instantaneous channel conditions are only due to, for example, fast multi-path fading, all users will experience the same average data rate. Any variations in the instantaneous data rate are rapid and often not even noticeable by the user. However, in practice different devices will also experience differences in the (short-term) average channel conditions—for example, due to differences in the distance between the base station and the device. In this case, the channel conditions experienced by one device may, for a relatively long time, be worse than the channel conditions experienced by other devices. A pure max-C/I-scheduling strategy would then "starve" the device with the bad channel conditions, and the device with bad channel conditions will never be scheduled. This is illustrated in Figure 9.2(a), where a max-C/I

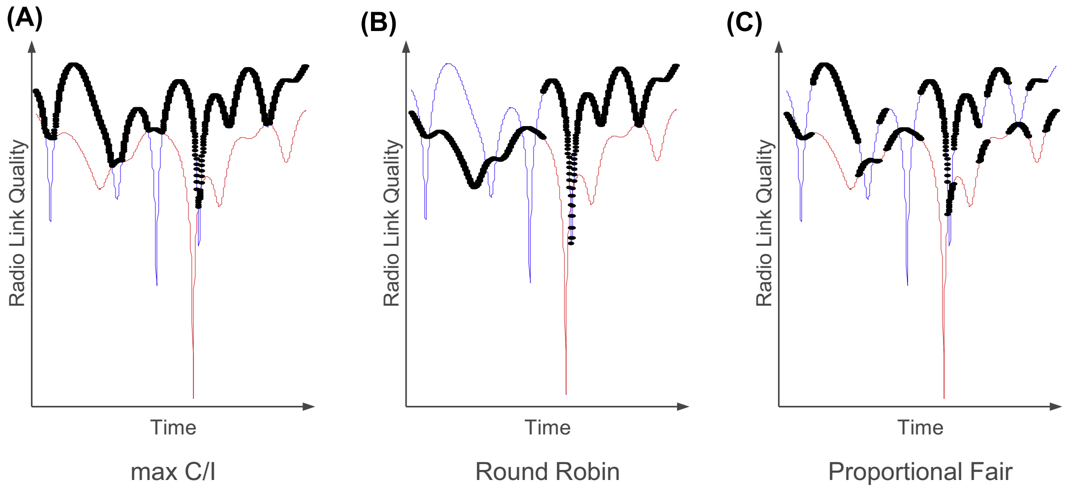

FIGURE 9.2

Example of three different scheduling behaviors for two users with different average channel quality: (a) max C/I, (b) round robin, and (c) proportional fair. The selected user is shown with bold lines.

scheduler is used to schedule between two different users with different average channel quality. Although resulting in the highest system capacity, this situation is often not acceptable from a quality-of-service point of view.

An alternative to the max-C/I scheduling strategy is so-called *round-robin* scheduling, illustrated in Figure 9.2(b). This scheduling strategy lets users take turns in using the shared resources, without taking the instantaneous channel conditions into account. Round-robin scheduling can be seen as fair scheduling in the sense that the same amount of radio resources (the same amount of time) is given to each communication link. However, round-robin scheduling is not fair in the sense of providing the same service quality to all communication links. In that case more radio resources (more time) must be given to communication links with bad channel conditions. Furthermore, as round-robin scheduling does not take the instantaneous channel conditions into account in the scheduling process, it will lead to lower overall system performance but more equal service quality between different communication links, compared to max-C/I scheduling.

A third possibility is the so-called *proportional fair* scheduler, see Figure 9.2(c), which tries to exploit rapid channel variations while suppressing the effect of differences in the average channel gain. In this strategy, the shared resources are assigned to the user with the *relatively* best radio-link conditions—that is, at each time instant, user k is selected for transmission according to

$$k = \arg\max_i \frac{R_i}{\overline{R_i}}$$

where R_i is the instantaneous data rate for user i and $\overline{R_i}$ is the average data rate for user i. The average is calculated over a certain averaging period long enough to average out differences in the fast channel-quality variations and at the same time short enough so that quality variations within the interval are not strongly noticed by a user.

From the discussion above it is seen that there is a fundamental trade-off between fairness and system capacity. The more unfair the scheduler is, the higher the system throughput under the assumption of an infinite amount of data to transmit for each user. However, in real situations there is not an infinite amount of data and the properties of the (bursty) traffic plays large role. At low system load—that is, when only one or, in some cases, a few users have data waiting for transmission at the base station at each scheduling instant—the differences between different scheduling strategies above are fairly small while they are more pronounced at higher loads and can be quite different compared to the full-buffer scenario above. This is illustrated in Figure 9.3 for web-browsing scenario. Each web page has a certain size and, after transmitting a page, there is no more data to be transmitted to the device in question until the user requests a new page by clicking on a link. In this case, a max-C/I scheduler can still provide a certain degree of fairness. Once the buffer for the user with the highest C/I has been emptied, another user with non-empty buffers will have the highest C/I and be scheduled and so on. The proportional fair scheduler has similar performance in both scenarios.

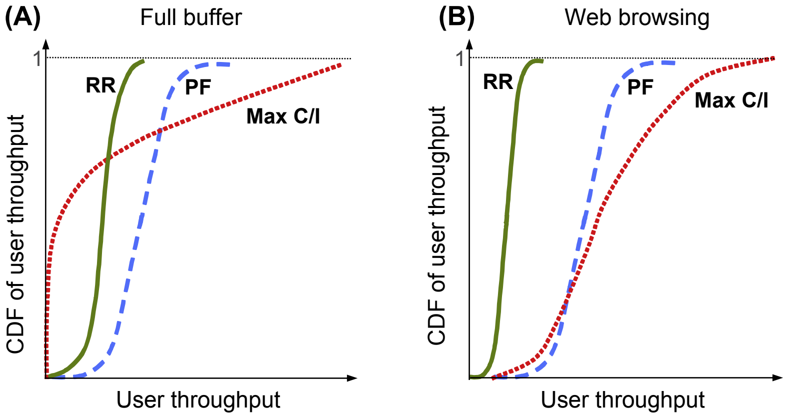

FIGURE 9.3

Illustration of the principle behavior of different scheduling strategies: (a) for full buffers and (b) for web browsing traffic model.

Clearly, the degree of fairness introduced by the traffic properties depends heavily on the actual traffic; a design made with certain assumptions may be less desirable in an actual network where the traffic pattern may be different from the assumptions made during the design. Therefore, relying solely on the traffic properties to achieve fairness is not a good strategy, but the discussion above also emphasizes the need to design the scheduler not only for the full buffer case. Traffic priorities, for example prioritizing a latency-critical services over a delay-tolerant service despite the channel quality for the latter being superior, is another example where the full-buffer discussion above was simplified to illustrate the basic principles. Other examples of scheduling input are DRX cycles, retransmissions, device capabilities, and device power consumption, all impacting the overall scheduling behavior.

The general discussion above is applicable to both downlink and uplink transmissions. However, there are some differences between the two. Fundamentally, the uplink power resource is *distributed* among the users, while in the downlink the power resource is *centralized* within the base station. Furthermore, the maximum uplink transmission power of a single device is typically significantly lower than the output power of a base station. This has a significant impact on the scheduling strategy. Unlike the downlink, where pure time-domain scheduling often can be used [15,16] and can be shown to be preferable from a theoretical point of view [13,14], uplink scheduling typically has to rely on sharing in the frequency domain in addition, as a single device may not have sufficient power for efficiently utilizing the link capacity. There are also other reasons to complement the time domain with the frequency domain, both in uplink and downlink, for example,

- in the case of insufficient payload—that is, the amount of data to transfer to a user is not sufficiently large to utilize the full channel bandwidth, or
- in the case where not only time-domain variations but also frequency-domain variations are to be exploited.

The scheduling strategies in these cases can be seen as generalizations of the schemes discussed for the time-domain-only cases in the preceding paragraphs. For example, to handle small payloads, a greedy filling approach can be used, where the scheduled user is selected according to max-C/I (or any other scheduling scheme). Once this user has been assigned resources matching the amount of data awaiting transmission, the second best user according to the scheduling strategy is selected and assigned (a fraction of) the residual resources and so on.

In the following sections, dynamic downlink and uplink scheduling, as well as related functionality such as uplink priority handling, scheduling requests, buffer status and power headroom reporting, semi-persistent scheduling, half-duplex FDD operation, and DRX functionality are described. Channel-state reporting, an important input to any channel-dependent scheduling, is discussed in Chapter 10. Remember that in LTE, only the general framework, and not the scheduling strategy, is standardized.

9.2 DOWNLINK SCHEDULING

The task of the downlink scheduler is to dynamically determine the device(s) to transmit to and, for each of these devices, the set of resource blocks upon which the device's DL-SCH should be transmitted. As discussed in the previous section, the amount of data in the transmission buffers as well as the desire to exploit channel-variations in the frequency domain implies that transmissions to multiple users in different parts of the spectrum is needed. Therefore, multiple devices can be scheduled in parallel in a subframe, in which case there is one DL-SCH per scheduled device (and component carrier), each dynamically mapped to a (unique) set of frequency resources.

The scheduler is in control of the instantaneous data rate used, and the RLC segmentation and MAC multiplexing will therefore be affected by the scheduling decision. Although formally part of the MAC layer but to some extent better viewed as a separate entity, the scheduler is thus controlling most of the functions in the eNodeB associated with downlink data transmission:

- *RLC*: Segmentation/concatenation of RLC SDUs is directly related to the instantaneous data rate. For low data rates, it may only be possible to deliver a part of an RLC SDU in a TTI, in which case segmentation is needed. Similarly, for high data rates, multiple RLC SDUs may need to be concatenated to form a sufficiently large transport block.
- *MAC*: Multiplexing of logical channels depends on the priorities between different streams. For example, radio resource control signaling, such as handover commands, typically has a higher priority than streaming data, which in turn has higher priority than a background file transfer. Thus, depending on the data rate and the amount of traffic of

different priorities, the multiplexing of different logical channels is affected. Hybrid-ARQ retransmissions also need to be accounted for.

- *L1*: Coding, modulation, and, if applicable, the number of transmission layers and the associated precoding matrix are affected by the scheduling decision. The choices of these parameters are mainly determined by the radio conditions and the selected data rate, that is, the transport block size.

The scheduling strategy is implementation-specific and not a part of the 3GPP specifications; in principle any strategy can be applied as discussed in Section 9.1. However, the overall goal of most schedulers is to take advantage of the channel variations between devices and preferably to schedule transmissions to a device when the channel conditions are advantageous. Most scheduling strategies therefore at least need information about:

- channel conditions at the device;
- buffer status and priorities of the different data flows;
- the interference situation in neighboring cells (if some form of interference coordination is implemented).

Information about the channel conditions at the device can be obtained in several ways. In principle, the eNodeB can use any information available, but typically the CSI reports from the device are used. The details of the transmission-mode-dependent CSI reports are found in Chapter 10. Other sources of channel knowledge, for example, exploiting channel reciprocity to estimate the downlink quality from uplink channel estimates in the case of TDD, can also be exploited by a particular scheduler implementation, either alone or in combination with CSI reports.

In addition to the channel-state information, the scheduler should take buffer status and priority levels into account. For example, it does not make sense to schedule a device with empty transmission buffers. Priorities of the different types of traffic may also vary; RRC signaling may be prioritized over user data. Furthermore, RLC and hybrid-ARQ retransmissions, which are in no way different from other types of data from a scheduler perspective, are typically also given priority over initial transmissions.

Downlink inter-cell interference coordination is also part of the implementation-specific scheduler strategy. A cell may signal to its neighboring cells the intention to transmit with a lower transmission power in the downlink on a set of resource blocks. This information can then be exploited by neighboring cells as a region of low interference where it is advantageous to schedule devices at the cell edge, devices that otherwise could not attain high data rates due to the interference level. Inter-cell interference handling is further discussed in Chapter 13.

9.3 UPLINK SCHEDULING

The basic function of the *uplink scheduler* is similar to its downlink counterpart, namely to dynamically determine, for each 1 ms interval, which devices are to transmit and on which uplink resources. As discussed before, the LTE uplink is primarily based on maintaining orthogonality between different uplink transmissions and the shared resource controlled by the eNodeB scheduler is time–frequency resource units. In addition to assigning the time–frequency resources to the device, the eNodeB scheduler is also responsible for controlling the transport format the device will use for each of the uplink component carriers. This allows the scheduler to tightly control the uplink activity to maximize the resource usage compared to schemes where the device autonomously selects the data rate, as autonomous schemes typically require some margin in the scheduling decisions. A consequence of the scheduler being responsible for selection of the transport format is that accurate and detailed knowledge in the eNodeB about the device situation with respect to buffer status and power availability is more accentuated in LTE compared to systems where the device autonomously controls the transmission parameters.

The basis for uplink scheduling is *scheduling grants*, containing the scheduling decision and providing the device information about the resources and the associated transport format to use for transmission of the UL-SCH on one component carrier. Only if the device has a valid grant is it allowed to transmit on the corresponding UL-SCH; autonomous transmissions are not possible. Dynamic grants are valid for one subframe, that is, for each subframe in which the device is to transmit on the UL-SCH, the scheduler issues a new grant.

Similarly to the downlink case, the uplink scheduler can exploit information about channel conditions, and, if some form of interference coordination is employed, the interference situation in neighboring cells. Information about the buffer status in the device, and its available transmission power, is also beneficial to the scheduler. This calls for the reporting mechanisms described in the following, unlike the downlink case where the scheduler, power amplifier, and transmission buffers all are in the same node. Uplink priority handling is, as already touched upon, another area where uplink and downlink scheduling differs.

Channel-dependent scheduling, which typically is used for the downlink, can be used for the uplink as well. In the uplink, estimates of the channel quality can be obtained from the use of uplink channel sounding, as described in Chapter 7. For scenarios where the overhead from channel sounding is too costly, or when the variations in the channel are too rapid to be tracked, for example at high device speeds, uplink diversity can be used instead. The use of frequency hopping as discussed in Chapter 7 is one example of obtaining diversity in the uplink.

Finally, inter-cell interference coordination can be used in the uplink for similar reasons as in the downlink by exchanging information between neighboring cells, as discussed in Chapter 13.

9.3.1 UPLINK PRIORITY HANDLING

Multiple logical channels of different priorities can be multiplexed into the same transport block using the similar MAC multiplexing functionality as in the downlink (described in Chapter 4). However, unlike the downlink case, where the prioritization is under control of the scheduler and up to the implementation, the uplink multiplexing is done according to a set of well-defined rules in the device with parameters set by the network as a scheduling grant applies to a specific uplink carrier of a device, not to a specific radio bearer within the device. Using radio-bearer-specific scheduling grants would increase the control signaling overhead in the downlink and hence per-device scheduling is used in LTE.

The simplest multiplexing rule would be to serve logical channels in strict priority order. However, this may result in starvation of lower-priority channels; all resources would be given to the high-priority channel until its transmission buffer is empty. Typically, an operator would instead like to provide at least some throughput for low-priority services as well. Therefore, for each logical channel in an LTE device, a *prioritized data rate* is configured in addition to the priority value. The logical channels are then served in decreasing priority order up to their prioritized data rate, which avoids starvation as long as the scheduled data rate is at least as large as the sum of the prioritized data rates. Beyond the prioritized data rates, channels are served in strict priority order until the grant is fully exploited or the buffer is empty. This is illustrated in Figure 9.4.

9.3.2 SCHEDULING REQUESTS

The scheduler needs knowledge about devices having data to transmit and therefore need to be scheduled uplink resources. There is no need to provide uplink resources to a device with no data to transmit as this would only result in the device performing padding to fill up the granted resources. Hence, as a minimum, the scheduler needs to know whether the device has data to transmit and should be given a grant. This is known as a *scheduling request*. Scheduling requests are used for devices not having a valid scheduling grant; devices that have a valid grant and are transmitting in the uplink provide more detailed information to the eNodeB as discussed in the next section.

A scheduling request is a simple flag, raised by the device to request uplink resources from the uplink scheduler. Since the device requesting resources by definition has no PUSCH resource, the scheduling request is transmitted on the PUCCH. Each device can be assigned a dedicated PUCCH scheduling request resource, occurring every *n*th subframe, as described in Chapter 7. With a dedicated scheduling-request mechanism, there is no need to provide the identity of the device requesting to be scheduled as the identity of the device is implicitly known from the resources upon which the request is transmitted. When data with higher

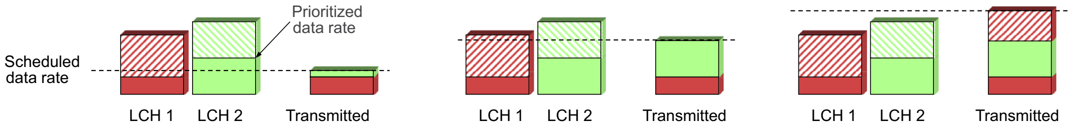

FIGURE 9.4

Prioritization of two logical channels for three different uplink grants.

priority than already existing in the transmit buffers arrives at the device and the device has no grant and hence cannot transmit the data, the device transmits a scheduling request at the next possible instant, as illustrated in Figure 9.5. Upon reception of the request, the scheduler can assign a grant to the device. If the device does not receive a scheduling grant until the next possible scheduling-request instant, the scheduling request is repeated up to a configurable limit after which the device resorts to random access to request resources form the eNodeB.

The use of a single bit for the scheduling request is motivated by the desire to keep the uplink overhead small, as a multi-bit scheduling request would come at a higher cost. A consequence of the single-bit scheduling request is the limited knowledge at the eNodeB about the buffer situation at the device when receiving such a request. Different scheduler implementations handle this differently. One possibility is to assign a small amount of resources to ensure that the device can exploit them efficiently without becoming power limited. Once the device has started to transmit on the UL-SCH, more detailed information about the buffer status and power headroom can be provided through the inband MAC control message, as discussed in the following text. Knowledge of the service type may also be used—for example, in the case of voice the uplink resource to grant is preferably the size of a typical voice-over-IP package. The scheduler may also exploit, for example, path-loss measurements used for mobility and handover decisions to estimate the amount of resources the device may efficiently utilize.

An alternative to a dedicated scheduling-request mechanism would be a contention-based design. In such a design, multiple devices share a common resource and provide their identity as part of the request. This is similar to the design of the random access. The number of bits transmitted from a device as part of a request would, in this case, be larger with the correspondingly larger need for resources. In contrast, the resources are shared by multiple users. Basically, contention-based designs are suitable for a situation where there are a large number of devices in the cell and the traffic intensity, and hence the scheduling intensity, is low. In situations with higher intensities, the collision rate between different devices simultaneously requesting resources would be too high and lead to an inefficient design.

Since the scheduling-request design for LTE relies on dedicated resources, a device that has not been allocated such resources cannot transmit a scheduling request. Instead, devices

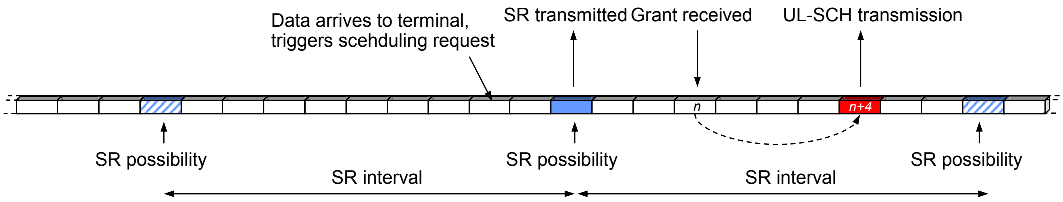

FIGURE 9.5

Scheduling request transmission.

without scheduling-request resources configured rely on the random-access mechanism described in Chapter 11. In principle, an LTE device can therefore be configured to rely on a contention-based mechanism if this is advantageous in a specific deployment.

9.3.3 BUFFER STATUS REPORTS

Devices that already have a valid grant do not need to request uplink resources. However, to allow the scheduler to determine the amount of resources to grant to each device in future subframes, information about the buffer situation, discussed in this section, and the power availability, discussed in the next section, is useful. This information is provided to the scheduler as part of the uplink transmission through MAC control elements (see Chapter 4 for a discussion on MAC control elements and the general structure of a MAC header). The LCID field in one of the MAC subheaders is set to a reserved value indicating the presence of a buffer status report, as illustrated in Figure 9.6.

From a scheduling perspective, buffer information for each logical channel is beneficial, although this could result in a significant overhead. Logical channels are therefore grouped into logical-channel groups and the reporting is done per group. The buffer-size field in a buffer-status report indicates the amount of data awaiting transmission across all logical channels in a logical-channel group. A buffer-status report can be triggered for the following reasons:

- Arrival of data with higher priority than currently in the transmission buffer—that is, data in a logical-channel group with higher priority than the one currently being transmitted—as this may impact the scheduling decision.
- Change of serving cell, in which case a buffer-status report is useful to provide the new serving cell with information about the situation in the device.
- Periodic reporting, as controlled by a timer.

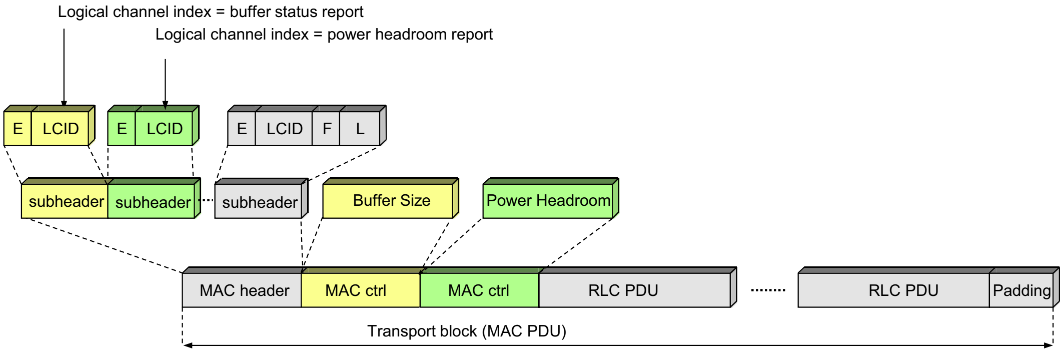

FIGURE 9.6

Signaling of buffer status and power headroom reports.

- To reduce padding. If the amount of padding required to match the scheduled transport block size is larger than a buffer-status report, a buffer-status report is inserted as it is better to exploit the available payload for useful scheduling information instead of padding, if possible.

9.3.4 POWER HEADROOM REPORTS

In addition to buffer status, the amount of transmission power available in each device is also relevant for the uplink scheduler. There is little reason to schedule a higher data rate than the available transmission power can support. In the downlink, the available power is immediately known to the scheduler as the power amplifier is located in the same node as the scheduler. For the uplink, the power availability, or *power headroom*, needs to be provided to the eNodeB. Power headroom reports are therefore transmitted from the device to the eNodeB in a similar way as the buffer-status reports, that is, only when the device is scheduled to transmit on the UL-SCH. A power headroom report can be triggered for the following reasons:

- Periodic reporting as controlled by a timer.
- Change in path loss (when the difference between the current power headroom and the last report is larger than a configurable threshold).
- To reduce padding (for the same reason as buffer-status reports).

It is also possible to configure a prohibit timer to control the minimum time between two power-headroom reports and thereby the signaling load on the uplink.

There are two different types of power-headroom reports defined in LTE, *Type 1* and *Type 2*. Type 1 reporting reflects the power headroom assuming PUSCH-only transmission on the carrier, while the Type 2 report, introduced in release 10, assumes combined PUSCH and PUCCH transmission.

The Type 1 power headroom, valid for a certain subframe (and a certain component carrier) and assuming that the device was scheduled for PUSCH transmission in that subframe, is given (in dB) by

$$\text{Power Headroom} = P_{\mathrm{CMAX,c}} - \left(P_{0,\mathrm{PUSCH}} + \alpha \cdot PL_{\mathrm{DL}} + 10 \cdot \log_{10}(M) + \Delta_{\mathrm{MCS}} + \delta\right) \tag{9.1}$$

where the values for M and Δ_{MCS} correspond to the resource assignment and modulation-and-coding scheme used in the subframe to which the power-headroom report corresponds.[2] The quantity δ captures the change in transmission power due to the closed-loop power control as described in Chapter 7. The explicitly configured *maximum per-carrier transmit power* for component carrier c is denoted, $P_{\mathrm{CMAX,c}}$. It can be noted that the power headroom is not a measure of the difference between the maximum per-carrier transmit power and the actual carrier transmit power. Rather, it can be seen that the power headroom is a measure of the difference between $P_{\mathrm{CMAX,c}}$ and the transmit power that would have been used *assuming that there would have been no upper limit on the transmit power.* Thus, the

[2] In case of carrier aggregation, type 1 reports are supported for each of the component carriers.

power headroom can very well be negative, indicating that the per-carrier transmit power was limited by $P_{\text{CMAX,c}}$ at the time of the power headroom reporting, that is, the network has scheduled a higher data rate than the device can support given the available transmission power. As the network knows what modulation-and-coding scheme and resource size the device used for transmission in the subframe to which the power-headroom report corresponds, it can determine the valid combinations of modulation-and-coding scheme and resource size M, assuming that the downlink path loss PL_{DL} and the term δ have not changed substantially.

Type 1 power headroom can also be reported for subframes where there is no actual PUSCH transmission. In such cases, $10 \cdot \log_{10}(M)$ and Δ_{MCS} in the expression above are set to zero. This can be seen as the power headroom assuming a default transmission configuration corresponding to the minimum possible resource assignment ($M = 1$) and the modulation-and-coding scheme associated with $\Delta_{\text{MCS}} = 0$ dB.

Similarly, Type 2 power headroom reporting is defined as the difference between the maximum per-carrier transmit power and the sum of the PUSCH and PUCCH transmit power (Eqs. (7.4) and (7.3), respectively), once again not taking into account any maximum per-carrier power when calculating the PUSCH and PUCCH transmit power.[3]

Along the lines of Type 1 power headroom reporting, the Type 2 power headroom can also be reported for subframes in which no PUSCH and/or PUCCH is transmitted. In that case a virtual PUSCH and PUCCH transmit power is calculated, assuming the smallest possible resource assignment ($M = 1$) and $\Delta_{\text{MCS}} = 0$ dB for PUSCH and $\Delta_{\text{Format}} = 0$ for PUCCH.

9.4 TIMING OF SCHEDULING ASSIGNMENTS/GRANTS

The scheduling decisions, downlink scheduling assignments and uplink scheduling grants, are communicated to each of the scheduled devices through the downlink L1/L2 control signaling as described in Chapter 6, using one (E)PDCCH per downlink assignment. Each device monitors a set of (E)PDCCHs for valid scheduling assignment or grants and, upon detection of a valid assignment or grant, receives PDSCH or transmits PUSCH, respectively. The device needs to know which subframe the scheduling command relates to.

9.4.1 DOWNLINK SCHEDULING TIMING

For downlink data transmission, the scheduling assignment is transmitted in the same subframe as the data. Having the scheduling assignment in the same subframe as the corresponding data minimizes the latency in the scheduling process. Also, note that there is no possibility for dynamically scheduling future subframes—data and the scheduling assignment are always in the same subframe. This holds for FDD as well as TDD.

[3] In case of carrier aggregation, type 2 reports are supported for the primary component carrier only, as PUCCH cannot be transmitted on a secondary component carrier (prior to release 13).

9.4.2 UPLINK SCHEDULING TIMING

The timing for uplink scheduling grants is more intricate than the corresponding downlink scheduling assignments, especially for TDD. The grant cannot relate to the same subframe it was received in as the uplink subframe has already started when the device has decoded the grant. The device also needs some time to prepare the data to transmit. Therefore, a grant received in subframe *n* affects the uplink transmission in a later subframe.

For FDD, the grant timing is straightforward. An uplink grant received in a subframe *n* triggers an uplink transmission in subframe $n+4$, as illustrated in Figure 9.7. This is the same timing relation as used for uplink retransmission triggered by the PHICH, motivated by the possibility to override the PHICH by a dynamic scheduling grant, as described in Chapter 8.

For TDD, the situation is more complicated. A grant received in subframe *n* in TDD may not necessarily trigger an uplink transmission in subframe $n+4$—the timing relation used in FDD—as subframe $n+4$ may not be an uplink subframe. Hence, for TDD configurations 1–6 the timing relation is modified such that the uplink transmission occurs in subframe $n+k$, where k is the smallest value larger than or equal to 4 such that subframe $n+k$ is an uplink subframe. This provides at least the same processing time for the device as in the FDD

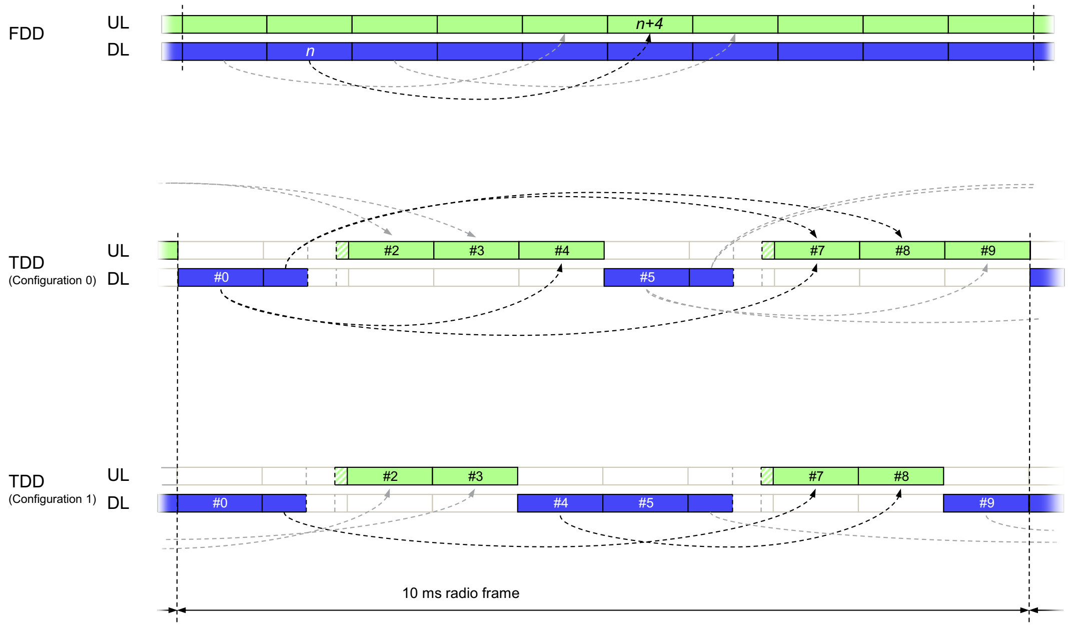

FIGURE 9.7

Timing relation for uplink grants in FDD and TDD configurations 0 and 1.

case while minimizing the delay from receipt of the uplink grant to the actual transmission. Note that this implies that the time between grant reception and uplink transmission may differ between different subframes. Furthermore, for the downlink-heavy configurations 1—5, another property is that uplink scheduling grants can only be received in some of the downlink subframes.

For TDD configuration 0 there are more uplink subframes than downlink subframes, which calls for the possibility to schedule transmissions in multiple uplink subframes from a single downlink subframe. The same timing relation as for the other TDD configurations is used but with slight modifications. Recall from Section 6.4.7 that the grant transmitted in the downlink contains an uplink index consisting of two bits. For uplink-downlink configuration 0, the index field specifies which uplink subframe(s) a grant received in a downlink subframe applies to. For example, as illustrated in Figure 9.7, an uplink scheduling grant received in downlink subframe 0 applies to one or both of the uplink subframes 4 and 7, depending on which of the bits in the uplink index are set.

9.5 SEMI-PERSISTENT SCHEDULING

The basis for uplink and downlink scheduling is dynamic scheduling, as described in Sections 9.2 and 9.3. Dynamic scheduling with a new scheduling decision taken in each subframe allows for full flexibility in terms of the resources used and can handle large variations in the amount of data to transmit at the cost of the scheduling decision being sent on an (E)PDCCH in each subframe. In many situations, the overhead in terms of control signaling on the (E) PDCCH is well motivated and relatively small compared to the payload on DL-SCH/UL-SCH. However, some services, most notably voice-over IP, are characterized by regularly occurring transmission of relatively small payloads. To reduce the control signaling overhead for those services, LTE provides semi-persistent scheduling in addition to dynamic scheduling.

With semi-persistent scheduling, the device is provided with the scheduling decision on the (E)PDCCH, together with an indication that this applies to every *n*th subframe until further notice. Hence, control signaling is only used once and the overhead is reduced, as illustrated in Figure 9.8. The periodicity of semi-persistently scheduled transmissions, that is, the value of *n*, is configured by RRC signaling in advance, while activation (and deactivation)

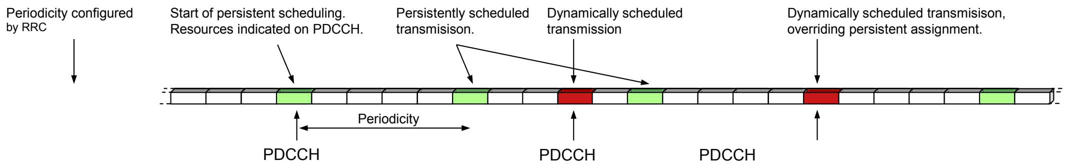
Periodicity configured by RRC
Start of persistent scheduling. Resources indicated on PDCCH.
Persistently scheduled transmisison.
Dynamically scheduled transmission
Dynamically scheduled transmisison, overriding persistent assignment.
Periodicity
PDCCH
PDCCH
PDCCH

FIGURE 9.8

Example of semi-persistent scheduling.

is done using the (E)PDCCH with the semi-persistent C-RNTI.[4] For example, for voice-over IP the scheduler can configure a periodicity of 20 ms for semi-persistent scheduling and, once a talk spurt starts, the semi-persistent pattern is triggered by the (E)PDCCH.

After enabling semi-persistent scheduling, the device continues to monitor the set of candidate (E)PDCCHs for uplink and downlink scheduling commands. When a dynamic scheduling command is detected, it takes precedence over the semi-persistent scheduling in that particular subframe, which is useful if the semi-persistently allocated resources occasionally need to be increased. For example, for voice-over IP in parallel with web browsing it may be useful to override the semi-persistent resource allocation with a larger transport block when downloading the web page.

For the downlink, only initial transmissions use semi-persistent scheduling. Retransmissions are explicitly scheduled using an (E)PDCCH assignment. This follows directly from the use of an asynchronous hybrid-ARQ protocol in the downlink. Uplink retransmissions, in contrast, can either follow the semi-persistently allocated subframes or be dynamically scheduled.

9.6 SCHEDULING FOR HALF-DUPLEX FDD

Half-duplex FDD implies that a single device cannot receive and transmit at the same time while the eNodeB still operates in full duplex. In LTE, half-duplex FDD is implemented as a scheduler constraint, implying it is up to the scheduler to ensure that a single device is not simultaneously scheduled in uplink and downlink. Hence, from a device perspective, subframes are dynamically used for uplink or downlink. Briefly, the basic principle for half-duplex FDD is that a device is receiving in the downlink unless it has been explicitly instructed to transmit in the uplink (either UL-SCH transmission or hybrid-ARQ acknowledgements triggered by a downlink transmission). The timing and structure for control signaling are identical between half- and full-duplex FDD devices. Note that, as the eNodeB is operating in full duplex, regardless of the duplex capability of the devices, the cell capacity is hardly affected by the presence of half-duplex devices as, given a sufficient number of devices with data to transmit/receive, the scheduler can with a high likelihood find a set of devices to schedule in the uplink and another set to schedule in the downlink in a given subframe.

An alternative to a dynamic half-duplex FDD based on scheduling restrictions would be to base half-duplex FDD on the TDD control signaling structure and timing, with a semi-static configuration of subframes to either downlink or uplink. However, this would complicate supporting a mixture of half- and-full duplex devices in the same cell as the timing of the control signaling would differ. It would also imply a waste of uplink spectrum resources. All FDD devices need to be able to receive subframes 0 and 5 in some situations as those subframes are used for system information and synchronization signals. Hence, if a fixed

[4] Each device has two identities, the "normal" C-RNTI for dynamic scheduling and the semi-persistent C-RNTI for semi-persistent scheduling.

uplink—downlink allocation were to be used, no uplink transmissions could take place in those two subframes, resulting in a loss in uplink spectral efficiency of 20%. This is not attractive and led to the choice of implementing half-duplex FDD as a scheduling strategy instead.

Support for half-duplex FDD has been part of LTE since the beginning but has so far seen limited use in practice. However, with the increased interest in massive machine-type communication in LTE release 12 and later, there is a renewed interest in half-duplex FDD as part of reducing the device cost for these devices. Hence, there are two ways of handling half-duplex FDD in LTE, differing in the optimization criterion and how the necessary guard time between reception and transmission is created:

- type A, part of LTE from the beginning and aiming at high performance by minimizing the guard time between reception and transmission, and
- type B, introduced in LTE release 12 and providing a long guard time to facilitate simple low-cost implementation for massive MT devices.

As stated, half-duplex type A has been part of LTE from the beginning, focusing on minimizing the guard time between reception and transmission. In this mode-of-operation, guard time for the downlink-to-uplink switch is created by allowing the device to skip reception of the last OFDM symbols in a downlink subframe immediately preceding an uplink subframe, as described in Chapter 5. Note that skipping reception of the last symbol in the downlink is only required if there is an uplink transmission immediately after the downlink subframe, otherwise the full downlink subframe is received. Guard time for the uplink-to-downlink switch is handled by setting the appropriate amount of timing advance in the devices. Compared to type B, the guard time is fairly short, resulting in high performance.

An example of half-duplex type A operation as seen from a device perspective is shown in Figure 9.9. In the leftmost part of the figure, the device is explicitly scheduled in the uplink and, consequently, cannot receive data in the downlink in the same subframe. The uplink transmission implies the receipt of an acknowledgement on the PHICH four subframes later, as mentioned in Chapter 8, and therefore the device cannot be scheduled in the uplink in this subframe. Similarly, when the device is scheduled to receive data in the downlink in subframe *n*, the corresponding hybrid-ARQ acknowledgement needs to be transmitted in the uplink

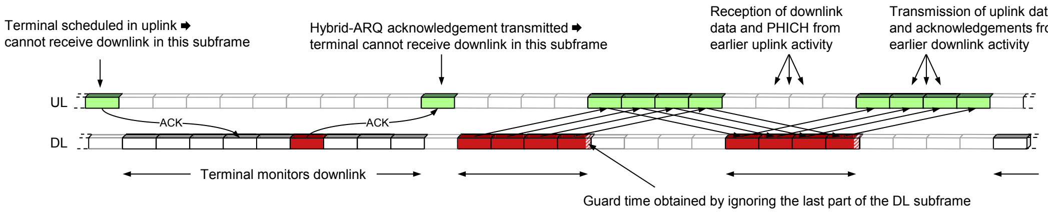
Terminal scheduled in uplink ➡ cannot receive downlink in this subframe
Hybrid-ARQ acknowledgement transmitted ➡ terminal cannot receive downlink in this subframe
Reception of downlink data and PHICH from earlier uplink activity
Transmission of uplink data and acknowledgements from earlier downlink activity
UL
DL
ACK
ACK
Terminal monitors downlink
Guard time obtained by ignoring the last part of the DL subframe

FIGURE 9.9

Example of type A half-duplex FDD device operation.

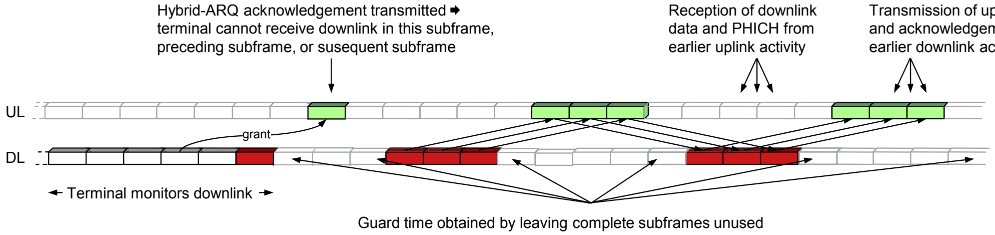
Hybrid-ARQ acknowledgement transmitted ➡ terminal cannot receive downlink in this subframe, preceding subframe, or susequent subframe
Reception of downlink data and PHICH from earlier uplink activity
Transmission of uplink data and acknowledgements from earlier downlink activity
UL
DL
grant
Terminal monitors downlink
Guard time obtained by leaving complete subframes unused

FIGURE 9.10

Example of type B half-duplex FDD device operation.

subframe $n + 4$, preventing downlink reception in subframe $n + 4$. The scheduler can exploit this by scheduling downlink data in four consecutive subframes and uplink transmission in the four next subframes when the device needs to transmit hybrid-ARQ acknowledgements in the uplink anyway, and so on. Hence, at most half of the time can be used in the downlink and half in the uplink or, in other words, the asymmetry in half-duplex FDD type A is 4:4. Efficient support of half-duplex FDD is one of the reasons why the same number of hybrid-ARQ processes was selected in uplink and downlink.

Half-duplex FDD type B was introduced in LTE release 12 as part of the overall work on enhancements for massive machine-type communication. In type B, a complete subframe is used as guard time between reception and transmission as well as transmission and reception as illustrated in Figure 9.10. The motivation behind half-duplex type B, as well as a thorough description of the enhancements for massive machine-type communication can be found in Chapter 20.

9.7 DISCONTINUOUS RECEPTION

Packet-data traffic is often highly bursty, with occasional periods of transmission activity followed by longer periods of silence. From a delay perspective, it is beneficial to monitor the downlink control signaling in each subframe to receive uplink grants or downlink data transmissions and instantaneously react on changes in the traffic behavior. At the same time this comes at a cost in terms of power consumption at the device; the receiver circuitry in a typical device represents a non-negligible amount of power consumption. To reduce the device power consumption, LTE includes mechanisms for *discontinuous reception* (DRX).

The basic mechanism for DRX is a configurable DRX cycle in the device. With a DRX cycle configured, the device monitors the downlink control signaling only in one subframe per DRX cycle, sleeping with the receiver circuitry switched off in the remaining subframes. This allows for a significant reduction in power consumption: the longer the cycle, the lower the power consumption. Naturally, this implies restrictions to the scheduler as the device can be addressed only in the active subframes.

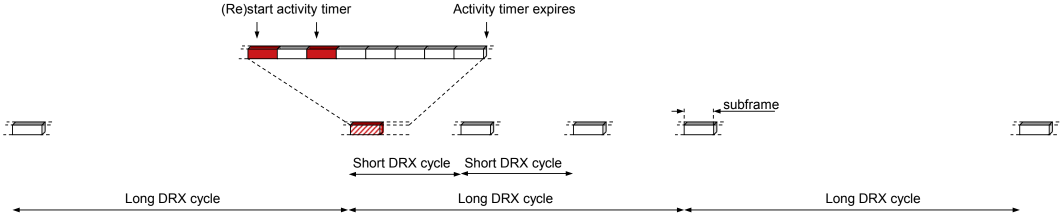

FIGURE 9.11

Illustration of DRX operation.

In many situations, if the device has been scheduled and active with receiving or transmitting data in one subframe, it is highly likely it will be scheduled again in the near future. One reason could be that it was not possible to transmit all the data in the transmission buffer in one subframe and additional subframes are required. Waiting until the next active subframe according to the DRX cycle, although possible, would result in additional delays. Hence, to reduce the delays, the device remains in the active state for a certain configurable time after being scheduled. This is implemented by the device (re)starting an inactivity timer every time it is scheduled and remaining awake until the time expires, as illustrated at the top of Figure 9.11.

Retransmissions take place regardless of the DRX cycle. Thus, the device receives and transmits hybrid-ARQ acknowledgements as normal in response to data transmission. In the uplink, this also includes retransmissions in the subframes given by the synchronous hybrid-ARQ timing relation. In the downlink, where asynchronous hybrid ARQ is used, the retransmission time is not fixed in the specifications. To handle this, the device monitors the downlink for retransmissions in a configurable time window after the previous transmission.

The above mechanism, a (long) DRX cycle in combination with the device remaining awake for some period after being scheduled, is sufficient for most scenarios. However, some services, most notably voice-over IP, are characterized by periods of regular transmission, followed by periods of no or very little activity. To handle these services, a second short DRX cycle can optionally be used in addition to the long cycle described above. Normally, the device follows the long DRX cycle, but if it has recently been scheduled, it follows a shorter DRX cycle for some time. Handling voice-over IP in this scenario can be done by setting the short DRX cycle to 20 ms, as the voice codec typically delivers a voice-over-IP packet per 20 ms. The long DRX cycle is then used to handle longer periods of silence between talk spurts.