CHAPTER

DUAL CONNECTIVITY

16

For a device to communicate at least one connection between the device and the network is required. As a baseline, the device is connected to one cell handling all the uplink as well as downlink transmissions. All data flows, user data as well as RRC signaling, is handled by this cell. This is a simple and robust approach, suitable for a wide range of deployments and the basis for LTE.

However, allowing the device to connect to the network at multiple cells can be beneficial in some scenarios (Figure 16.1):

- User-plane aggregation, where the device is transmitting and receiving data to/from multiple sites, in order to increase the overall data rate. Note that there are *different* streams transmitted/received to/from the different nodes unlike, for example, uplink CoMP where the *same* stream is received at multiple antenna sites.
- Control-plane/user-plane separation, where control plane communication is handled by one node and user plane by another. This can, for example, be used to maintain a robust control-plane connection to the macro cell while offloading the user-plane data to the pico cell.

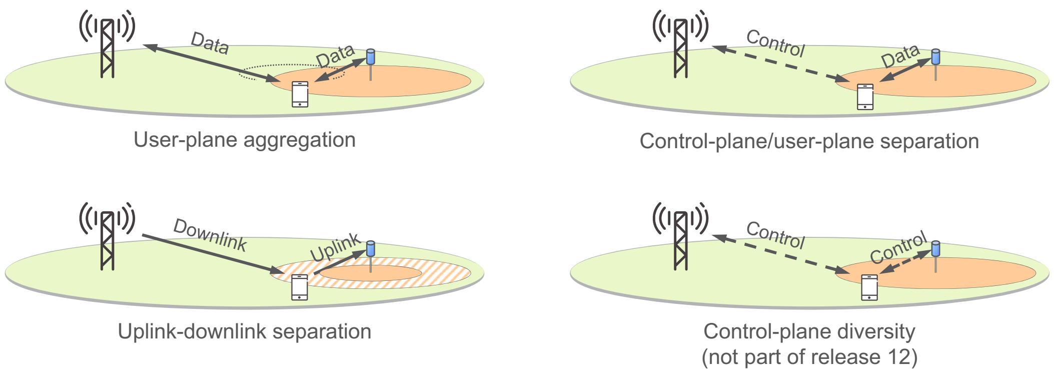

FIGURE 16.1

Usage scenarios for dual connectivity.

4G, LTE-Advanced Pro and The Road to 5G. http://dx.doi.org/10.1016/B978-0-12-804575-6.00016-9

- Uplink–downlink separation where downlink and uplink are handled by separate sites. In heterogeneous deployments, the best downlink is not necessarily the best uplink which motivates handling the two separately as discussed in Chapter 14.
- Control-plane diversity where RRC commands are sent from two nodes. One example of when this can be beneficial is mobility in a heterogeneous deployment. A device connected to a pico cell but moving into the macro cell will receive a handover command from the network. To reduce the likelihood of not receiving this command due to rapid degradation of the pico-link quality, the handover command can be transmitted from two nodes. Although this scenario was discussed as part of release 12, it is not supported by the final specifications due to time constraints in the development of release 12.

Note that control plane in this context refers to the control signaling transmitted from higher layers, for example, RRC signaling, and not to the L1/L2 control signaling. The device is connected to two cells and each of these cells handles their own L1/L2 control signaling.

Some of these scenarios can be supported partially already in LTE release 8. For example, uplink–downlink separation can be achieved by using remote radio units connected to a central baseband unit as discussed in Chapter 14. Carrier aggregation can also be used for support of these scenarios, for example, by transmitting different carriers from different antenna points connected to a common baseband unit. However, support for these scenarios is very rudimentary, if at all existing, and requires centralized baseband processing with correspondingly high backhaul requirements in terms of low latency and high capacity.

To address these shortcomings, *dual connectivity* was introduced in release 12 and further refined in release 13. Dual connectivity implies a device is simultaneously connected to two eNodeBs, the *master eNodeB* and the *secondary eNodeB*, each with its own scheduler and interconnected using the X2 interface, see Figure 16.2. Note that this is as seen from a device perspective or, expressed differently, a master eNodeB for one device could act as a secondary eNodeB for another device. Since both the master and secondary eNodeBs handle their own scheduling and has their own timing relations, the normal X2 interface can be used to connect the two sites with relaxed latency requirements. This is in contrast to two other techniques where a device can be connected to multiple cells, carrier aggregation and CoMP, which are

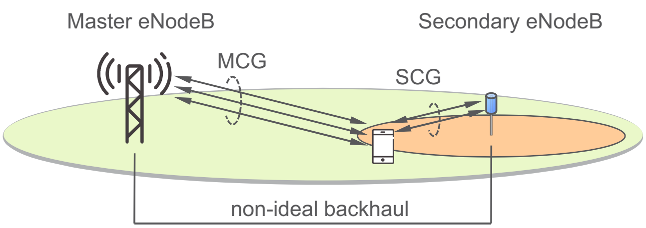

FIGURE 16.2

Dual connectivity.

more demanding on the interaction between the involved cells and typically requires low-latency, high-capacity backhaul and have stringent requirements on synchronization.

Carrier aggregation can be used in each of the two eNodeBs—that is, a device may be connected to multiple cells in each of the two nodes. The master eNodeB is responsible for scheduling transmissions in the *master cell group* (MCG) while the secondary eNodeB handles the *secondary cell group* (SCG). Within each cell group there is one primary component carrier[1] and, if carrier aggregation is used, one or more secondary component carriers. Similarly to carrier aggregation, dual connectivity is supported only when the device is active—that is, in the RRC_CONNECTED state, and not in idle mode.

It is assumed that the device is capable of simultaneous transmission (and reception) toward both the eNodeBs involved in dual connectivity. This is more or less a consequence of the fundamental assumption that dual connectivity should not require any dynamic coordination between the two eNodeBs. Furthermore, it is assumed that the two cell groups use different, nonoverlapping frequency bands—that is, only interband dual connectivity is supported. Although the dual-connectivity framework as such would allow same-frequency deployment, the resulting interference scenarios would require additional mechanisms not part of release 13. Finally, although the specifications are eNodeB-type agnostic, a heterogeneous scenario with the master cell group handled by a macro base station and the secondary cell group handled by a pico base station has often been assumed in the development of the overall solution. This is however not required. The first not the device connects to take the role of the master eNodeB and the secondary eNodeB is added subsequently; in case the device connected to the pico node first the pico node takes the role of the master eNodeB and the macro node may be added as a secondary eNodeB.

The dual-connectivity framework has turned out to be applicable also for inter-RAT aggregation. One example hereof is aggregation of LTE and WLAN, specified in release 13, where a primary cell group using LTE is aggregated with one or more WLAN carriers on a secondary basis. It is also expected that an evolution of the dual connectivity framework will play an important role for the tight interworking between LTE and a new 5G radio-access technology as discussed in Chapter 23.

16.1 ARCHITECTURE

The overall architecture described in Chapter 4 is used also for dual connectivity—that is, an X2 interface between the eNodeBs and an S1-c interface between the master eNodeB and the core network, see Figure 16.3. The master eNodeB always has a direct S1-u interface to S-GW, while the secondary eNodeB may or may not have a direct S1-u interface to S-GW depending on the architecture. Note that the figure illustrates the interfaces relevant for

[1] In the specifications, the primary component carrier is denoted PCell in the MCG and PSCell in the SCG, and the secondary component carriers are denoted SCell in both the cell groups.

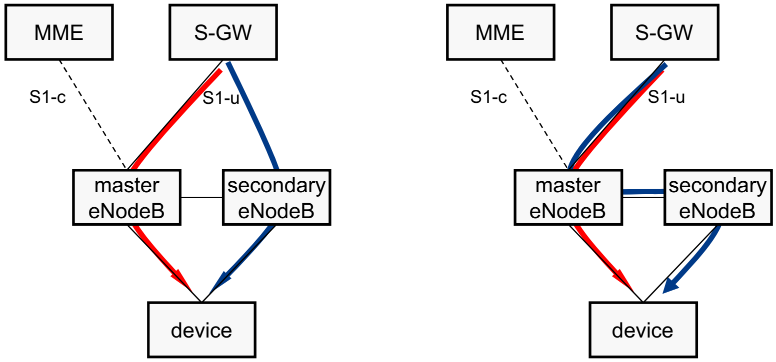

FIGURE 16.3

Architecture.

handling a certain device only. The secondary eNodeB is usually connected to the core network with S1-u and S1-c as it may act as a master eNodeB for another device.

For the user plane part, there are two possibilities of routing data to a device via the secondary eNodeB. In the first alternative, data to be transmitted via the secondary eNodeB is sent on the S1 interface to the secondary eNodeB directly (left part of Figure 16.3); in the second alternative the data to be transmitted by the secondary eNodeB is routed via the master eNodeB (right part of the Figure 16.3).

The first alternative is less demanding on the backhaul but results in dual connectivity being visible above the radio-access network which can be a drawback from a mobility perspective, for example, in scenarios with frequent changes of the secondary cell group. User-plane aggregation can be achieved with mechanisms such as multipath TCP [60]; that is, aggregation of the two data flows is done in the TCP transport layer above the core network and not in the radio-access network.

The second alternative, while more demanding on the backhaul as data between the secondary eNodeB and the device is routed via the master eNodeB, can provide higher performance as the aggregation is done closer to the radio interface and usage of multipath TCP is not necessary. Mobility resulting in a change of the secondary eNodeB is invisible to the core network which can be beneficial as handover between pico cells may be common in a heterogeneous deployment.

The protocol architecture for the user plane is shown in Figure 16.4 and is identical to the architecture for the single eNodeB case with the addition of support for split bearers. There are three different types of radio bearers: bearers transmitted from the master eNodeB, bearers transmitted from the secondary eNodeB, and split bearers transmitted across both eNodeBs. As seen in the figure, there is a single PDCP protocol entity for the split bearer, located in the master eNodeB, in which the data are split up and parts of the downlink data are forwarded over the X2 interface to the secondary eNodeB. There is a flow control mechanism defined on the X2 interface to control the amount of data forwarded to the secondary eNodeB.

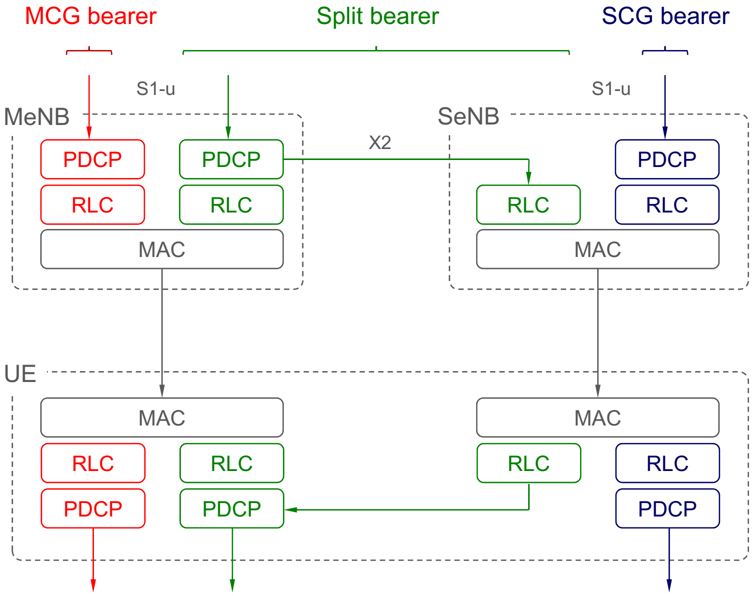

FIGURE 16.4

User-plane architecture.

In release 12, split bearers are supported in the downlink only and semi-static configuration assigns handling of uplink traffic of a split bearer to either the master eNodeB or the secondary eNodeB. This restriction is removed in release 13 where split bearers are supported also in the uplink.

On the device side, there are two MAC entities, one for each of the master and secondary cell groups. For split bearers, the received data packets may be received in the incorrect order, for example, if the secondary eNodeB had to perform retransmissions while this was not needed for the master eNodeB. Therefore, for RLC-AM the PDCP entity for a split bearer performs reordering in a similar way as the RLC handles reordering after the hybrid ARQ mechanism, see Chapter 8.

For the control plane, only the master eNodeB is connected to the MME[2] via the S1 interface, see Figure 16.5. This means that the secondary eNodeB is not visible to the MME, which is beneficial from a signaling perspective toward the core network as handovers are expected to be more frequent between secondary eNodeBs handling pico cells than for master eNodeBs handling macro cells in a heterogeneous deployment.

Each of the master and secondary eNodeBs independently handles scheduling and controls all the resources in the cells belonging to the respective eNodeB. Radio-resource control

[2]As discussed previously, this is from the perspective of a single device. The secondary eNodeB is typically connected to the MME as well as it may be a master eNodeB for another device.

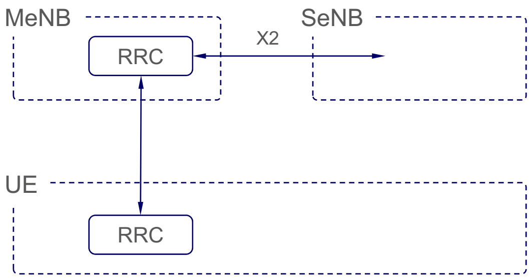

FIGURE 16.5

Control-plane architecture.

(RRC) is handled by the master eNodeB, transmitting RRC messages to the device. To handle radio resources in cells belonging to the secondary eNodeB, inter-eNodeB RRC messages have been defined. The master eNodeB does not modify the contents of the messages from the secondary eNodeB but simply encapsulates them into the RRC messages sent to the device.

16.2 PHYSICAL-LAYER IMPACT

Dual connectivity is mainly affecting the architecture while the impact on the physical-layer structure is modest. The two physical layers, one for each of the cell groups, operate independently of each other with their own data transmission, associated control signaling, random access, and so on. Each of the cell groups also has its own MAC layer which schedules physical-layer transmission independently. However, there are some aspects, in particular timing and power handling, where there is a mutual dependence between the physical layers in the two cell groups.

16.2.1 TIMING

LTE is designed to work both with and without inter-cell synchronization[3] and dual connectivity is not different in this respect. Within a cell group, the same requirements as for any other carrier aggregation scenario are applied—that is, all component carriers received are within a 33 µs window. Between cell groups, on the other hand, there is no requirement on synchronization, resulting in two cases:

- synchronous dual connectivity, where subframe boundaries of the two cell groups are aligned at the device within 33 µs;
- asynchronous dual connectivity, where subframe boundaries of the two cell groups have an arbitrary[4] timing relation at the device.

[3]In carrier aggregation all the involved cells are time synchronized.
[4]Actually, the specifications mention a 500-µs limit; time differences larger than this are handled by an offset in the subframe numbering.

In principle, there is no fundamental difference between these two cases but the actual implementation may differ. For example, if the carriers from the master cell group and secondary cell group are sufficiently close in frequency, a single RF chain may be capable of handling both if they are synchronized but not otherwise. Setting of the uplink transmission power is one example, which in many implementations can occur at the start of a subframe only. Such an implementation would be capable of synchronous dual connectivity where the subframe boundaries coincide but not asynchronous dual connectivity. To allow for different implementations, a device reports for each band combination whether it is capable of asynchronous operation or not. An asynchronous device can operate in either synchronous or asynchronous dual connectivity while a synchronous device can operate in synchronous dual connectivity only.

From a network perspective, it is up to the implementation to account for differences in SFN and subframe offset; for example, in conjunction with discontinuous reception in the device. Release 13 adds the possibility for the device to measure and report the difference between the two cell groups to assist network operation.

16.2.2 POWER CONTROL

Transmission power is the main area where there is an inter-cell-group dependency in the physical layer. Although the power setting is individual for each cell group, regulations specify the maximum transmission power *per device* and hence create a dependency between the cell groups when it comes to power sharing. Thus, when the device reaches its maximum transmission power there is a need to scale the power of the individual channels in the different cell groups. This may sound straightforward but the fact that the cell groups can be unsynchronized complicates the picture. For a given cell group, changes in the transmission power should occur at the subframe boundaries only as the receiver may assume that the transmission power is constant across a subframe.

The synchronous case, illustrated to the left in Figure 16.6, has all the subframe boundaries aligned across the cell groups. When setting the transmission power for subframe m in the master cell group the activity in the overlapping subframe in the secondary cell group is known and scaling the transmission power for the different channels is straightforward. Furthermore, changes in the transmission power occur at the subframe boundaries only.

FIGURE 16.6

Synchronous and asynchronous operation.

The asynchronous case, illustrated to the right in Figure 16.6, is more complicated. As an example, consider the master cell group (the situation is similar for the secondary cell group). The available transmission power in subframe m of the master cell group may now depend on *two* subframes in the secondary cell group, the ongoing subframe n and the future subframe $n+1$.

Since the power setting for the master cell group can be done at the corresponding subframe boundaries only, some margin for what might happen in the secondary cell group is necessary.

With the situation described earlier in mind, two methods for sharing the transmission power across cell groups are defined. They mainly differ in whether the scaling in case of a power limitation is done across all cells in all cell groups or separately per cell group. The power-control mode to use is configured by RRC signaling.

Dual-connectivity power-control mode 1 scales the power across cell groups as illustrated to the left of Figure 16.7. In case of a power limitation, the transmission power is scaled across all cells, regardless of the group they belong to, in the same way as in carrier aggregation. The only exception is that uplink control information in the master cell group is prioritized over uplink control information in the secondary cell group in case the same UCI type is used in both cell groups. In essence, this power-control mode does not differentiate between the cell groups and treats all cells in the same way. Power-control mode 1 is possible in synchronous operation only as the transmission power can be changed at subframe boundaries only. In an asynchronous scenario, the power of the master cell group would need to change as a result of power allocation done at the beginning of the subframes in an secondary cell group and vice versa, something which is not possible.

Dual-connectivity power-control mode 2 scales the power across carriers within each cell group but not between cell groups as illustrated to the right of Figure 16.7. The minimum guaranteed power available per cell group, expressed as a fraction of the maximum power, is configured through RRC signaling. In case of power limitation, each cell group is given at least its minimum guaranteed power. The remaining power is then first given to the cell group associated with the earlier transmission. In Figure 16.6 this means that, at the beginning of

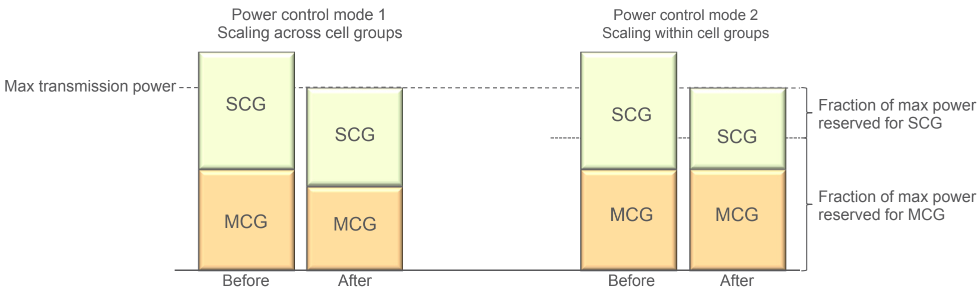

FIGURE 16.7

Power-control behavior for mode 1 and 2 when reaching the maximum transmission power.

subframe m, the secondary cell group can use the amount of the remaining power required to maintain the power constant during subframe n. Any power left after this is given to the master cell group in subframe m. Similarly, at the beginning of subframe $n+1$ in the secondary cell group, the master cell group uses the amount of power from the remaining power required to maintain the transmission power during subframe m. Since asynchronous operation implies that the subframe boundaries are not time aligned, transmission power for one cell group may need change at the subframe boundary for that cell group while it should be kept unchanged for the other cell group, power-control mode two is the only mode supported for asynchronous operation.

16.3 SCHEDULING IN DUAL CONNECTIVITY

Scheduling in dual connectivity is done independently in each eNodeB in the same way as described in earlier chapters—that is, the master cell group and the secondary cell group can independently transmit a scheduling request when needed. Similarly, discontinuous reception, DRX, is configured separately for each of the cell groups. Thus, as the two cell groups are scheduled independently, there is no need for tight coordination between the two eNodeBs although various forms of inter-cell coordination mechanism can be implemented if desired—neither is there a need to specify how to split downlink data across the two cell groups for a split bearer.

However, there are some aspects of scheduling, in particular handling of split bearers in the uplink in release 13, that are impacted by the introduction of dual connectivity. To determine how data for an uplink split bearer should be split across cell groups a threshold can be configured in the device. If the amount of data in the PDCP buffers for the split bearer is larger than the threshold, data is transmitted across both cell groups, otherwise it is transmitted only on one of the cell groups similarly to release 12.

Device reports used to support scheduling are also impacted by the introduction of dual connectivity. More specifically, the amount of data awaiting transmission and the available power, has to be split across the two cell groups with the corresponding impact on the reporting mechanisms.

Buffer status is reported per cell group. This is a natural choice as an eNodeB is only interested in the cell group it is scheduling. Knowing the buffer status for the master cell group does not help the eNodeB scheduling the secondary cell group and vice versa. For a split bearer, release 12 relies on semi-static configuration to determine in which cell group the split bearer should be included and reporting buffer status separately for the two cell groups works fine also for split bearers. In release 13, if the amount of data exceeds the previously mentioned threshold, the full amount of data is reported for the split bearer for both cell groups and it is up to the implementation to coordinate the scheduling decisions if needed.

Power headroom reporting is somewhat more complicated than buffer status reporting since the power is a resource shared across both cell groups as described earlier. Power

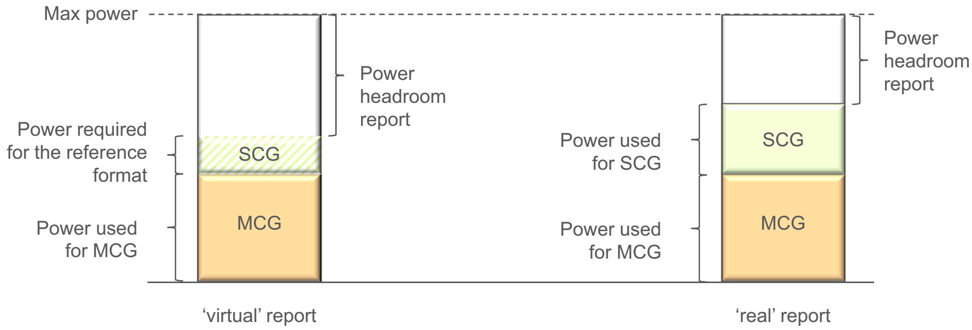

FIGURE 16.8

Power headroom report for master cell group (MCG) [the report for secondary cell group (SCG) is similar].

headroom reports for one cell group therefore need to take the activity in the other cell group into account. There are two possibilities for this with higher-layer configuration determining which one to use:

- the power used by the other cell group is given by a configured reference format (virtual report);
- the actual power used by the other cell group is used if it is transmitting (real report).

This is illustrated for the master cell group in Figure 16.8 (the same principle applies for the secondary cell group). The power headroom is the maximum transmission power of the device from which the actual master cell group transmission power and, depending on the reporting type, the actual power used (or virtual power assumed to be used) for the secondary cell group is subtracted.