Contents

Chapter 1

Fronthauling for 5G and Beyond

(Anvar Tukmanov, Maria A. Lema), (Ian Mings, Massimo Condoluci, Toktam Mahmoodi, Zaid Al-Daher), Mischa Dohler

The cellular network concept envisaged by D.H. (Doug) Ring in an internal Bell Labs journal [1]. In a 1947 paper, not publicly available at the time, fundamentals of a wide area coverage system for New York were laid out, and the designs of various coverage and service layers, nowadays referred to as HetNets, were described. The concept of deploying small cells in areas of high traffic, usage of frequency reuse plans, frequency discrimination in adjacent cells, and interference management are some of the techniques presented at the time to improve coverage and capacity radio planned network in New York. The key messages from this particular paper were that operators need to manage cellular interference whilst (i) increasing the number of sites and (ii) shrinking the reuse distance, Both of these concepts are still applicable even in todays mobile networks and in the forthcoming 5G radio networks.

Small cells, which are one the most prominent way to improve a cellular network performance, is a relative definition. In 1947 small cells had a radius of about 8km, nowadays a small cell radius is typically 500m or less; that is a reduction by a factor of at least 16. Furthermore, small cells and indoor type femto base stations necessitate techniques such as Self-Organised Networks (SON) to optimise and manage the interference levels amongst each other and amongst higher metro and macro layers. LTE in particular is designed to optimise the network spectral efficiency and utilises a frequency reuse of 1 (i.e. all cells use the full channel bandwidth available to an operator) introduces mechanisms for mitigating interference between cells and layers such as Inter-cell interference coordination (ICIC) and enhanced Inter-cell interference coordination (eICIC).

SON functionality gained popularity during the early days of 4G. Being a high performance, data centric network when compared to 3G, 4G LTE resulted in a vast increase in user traffic data and evidently an vast increase in the number of users migrating to 4G, which placed huge demands on operators to continuously enhance their networks typically in terms of coverage and capacity planning, and various access infill and in-band backhaul mesh solutions. In the early days of 4G, the increase in the number of users migrating from 3G and the deployment of new sites meant that operators had more cell-edge users than ever. SON has capabilities of dealing

with such interference problems to some extent. SON also has several other advantages including automation and rapid deployment all of which are highly desirable feature from an operators prospective, however, there are a number of implementation challenges currently been address by the likes of Small Cell Forum and NGMN. Some of these challenges to name a few include (i) reliability and interoperability in a multi-vendor, multi-operator environment, (ii) complexity in parametrisation, optimisation and algorithm development (iii) network performance variations based use case scenarios and (iv) the added complexity of network topologies all of which are key requirements for achieving optimum performance. The success of SON is also dependent on X2, which is the standard interface between base stations. X2 is currently used for handover and for basic interference coordination techniques. More advanced cooperative and interference mitigation techniques such as CoMP and Further eICIC (FeICIC) in addition to any further enhanced coordination techniques currently been proposed for 5G and the anticipated cell densification within different HetNeT layers, would require a much more stringent latency constraints on X2 and also S1 links.

To date, one of the approached to cell densification and interference, as depicted by many operators, is the concept of CRAN which was first introduced by IBM [2] and further developed by China Mobile Research Institute [3]. Base stations typically consist of a baseband unit (BBU) and a remote radio head (RRH) as illustrated in Figure 1.1. BBU which is usually located at the bottom of the base station contains RLC, PDCP, MAC and PHY layers and the RRH is usually placed near the antennas at the top of the mast contains the RF layer. The interconnection between the BBU and RRH which carries transmissions of In-phase and Quadrature (IQ) digitised data is based on a Common Public Radio Interface (CPRI), a specification developed by a number of vendors including Ericsson, Huawei Technologies, NEC Corporation, Alcatel Lucent and Nokia Solutions.

C-RAN is innovative network architecture designed to potentially overcome traffic and cellular interference challenges in dense networks. In CRAN architecture, the BBUs from a cluster of base stations is co-located, typically in a central location or in a data centre, and therefore the processing and transmissions are centralised in one place supporting up to hundreds of cells simultaneously. This enables more efficient coordinated transmissions amongst the base stations which are now referred to as remote radio heads (RRH) since they contain mainly the RF functions and not a complete eNB LTE stack. Coordinated scheduling in CRAN directly enhances the networks performance as naturally, it provides a more direct way to manage interference at cell edges than traditional networks such as DRAN in addition to the ability to adapt capacity resources to different traffic trends more efficiently and supporting some LTEs advance features and transmission schemes such as MU-MIMO and CoMP. There are numerous other benefits of CRAN, see Figure 1.2, such as scalability, reduced operational and energy costs and virtualisation of the RAN, the details of these however are out of the scope of this chapter. A recent trial from China Mobile for example has shown 53% and 30% savings in OPEX and CAPEX respectively [China Research] In order to understand the tradeoffs associated with the delivery

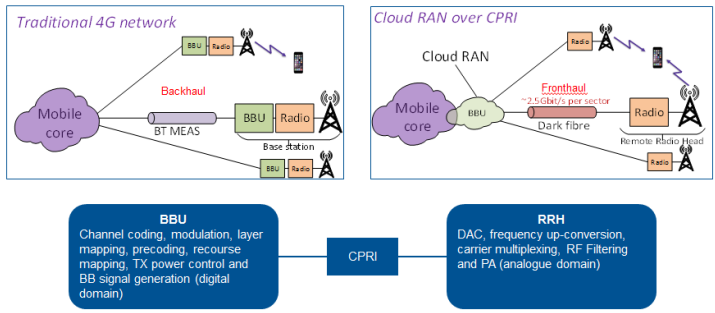

Figure 1.1: Base stations typically consist of a baseband unit (BBU) and a remote radio head (RRH).

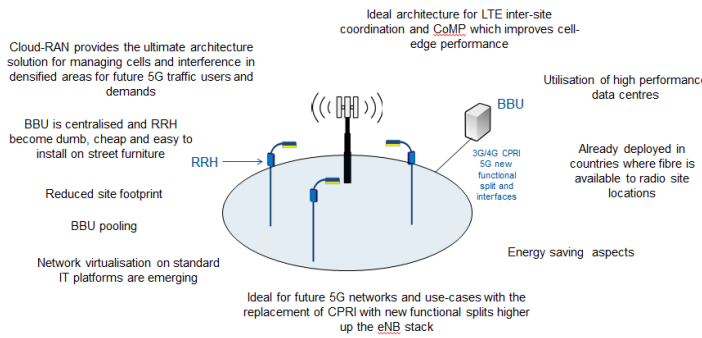

Figure 1.2: CRAN benefits include scalability, reduced operational and energy costs and virtualisation of the RAN.

of CRAN, the next section describes options for RAN functional splits between the functions performed centrally, and the functions retained in the RRH or RU.

The rest of the chapter is organised as follows. Section 1.1 provides a deeper look into the operation of current RAN protocols in a base station, also highlighting the typical challenges a fronthaul link serving a split 5G base station is likely to inherit from 4G systems. We then provide a review of key future network architecture elements and radio features affecting fronthaul transmissions in Section 1.2. Sections 1.3 and 1.4 respectively focus on the current fronthaul solutions that will be available for the first 5G systems, and on the industry view on aspects of 5G fronthauling.

1.1 RAN functional split options

Large channel bandwidths and lower end-to-end latency are some of the likely characteristics of 5G air interface, entering an active definition stage in standards at the time of writing. The requirement for the channels exceeding the current LTE max-

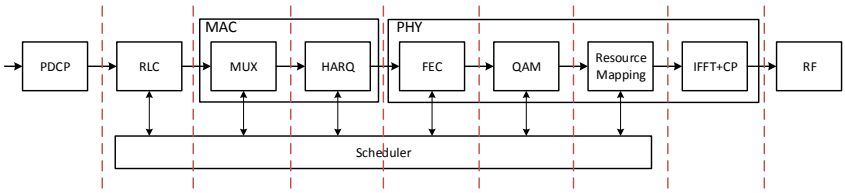

Figure 1.3: Candidate functional splits

imum bandwidth of 20 MHz originates from ever growing demand for broadband speeds, and may be realised through the use of microwave and millimetre wave (mmW) spectrum together with spectrum aggregation techniques. Reduced latency, in the order of a few milliseconds, in turn is required for some of 5G use cases featuring various system control applications, such as robotics and virtual reality, and may be achieved using shorter and more flexible radio frame structures.

Functional separation of RAN with these features is explored in this section, focusing on the coupling between functions and on the potential impact on fronthauling for split RAN architectures.

1.1.1 Splitting RAN air interface protocols

Efficient multiplexing of traffic and control data for multiple users within a single air interface frame is likely to remain the core function of RAN in 5G. In LTE this is realised through a scheduling function, present in every base station and working to achieve different objectives, such as consistent service performance for all connected users or minimisation of effects from interference. Scheduling algorithms defining the eventual mapping of information bits to waveform states are not standardised, however a scheduler is typically coupled with a number of base station's radio functions, affecting fronthaul requirements for different base station split options.

Figure 1.3 illustrates the sequence of LTE RAN protocols utilised in downlink transmission, their potential relationship to scheduling, and options for splits between the functions performed in a central unit (CU) and a remote unit (RU) connected with a fronthaul link. Operation of individual protocols in the radio stack of a base station is well described in literature, e.g. [4, 5], hence we provide a brief description of RAN protocols, focusing more on their interaction through scheduling and potential effects on fronthaul.

User plane radio protocol stack in LTE base stations consists of four main functional blocks. Packet Data Convergence Protocol (PDCP) is responsible for encryption, header compression and in-sequence delivery of user and control data. Radio-Link Control (RLC) layer performs segmentation of PDCP protocol data units (PDUs) and slower-rate retransmissions through ARQ mechanisms. Segmented data is then passed to the Medium Access Control (MAC) layer for multiplexing and scheduling, resulting in Transport Blocks (TBs). Physical layer (PHY) applies for-

ward error correction coding and maps encoded TBs onto resource elements to be transmitted over the air.

In order to understand the role of scheduling in the RAN protocol stack and its possible interaction with four protocol layers above, consider an example of a base station mapping information bits entering the PDCP layer onto resource elements (REs) within the LTE time-frequency resource grid at PHY layer. Each RE represents a subcarrier carrying one OFDM symbol represented by a state of the carrier waveform. Depending on the type of transmitted information and channel conditions, between 2 and 8 coded bits per symbol can be carried in one RE.

However note that the bits mapped onto REs are *coded* bits, comprising codewords that are derived from the TBs at the PHY layer. Therefore codeword bit grouping information based on the modulation scheme selected at MAC layer is required at the PHY modulation stage in an explicit form. While this information can be derived from control messaging such as Transport Format, the scheduler may need to be aware of granular channel quality in order to benefit from adaptive modulation and coding. Depending on implementation, these selection decisions may have dependencies spanning as far as RLC layer. For example, the base station may implement a scheduler adapting to near real-time channel conditions, or optimise device's memory requirements by reducing the need for data buffering. In this case user traffic multiplexing decisions within MAC and segmentation in RLC may depend on the distribution of supported modulation orders across REs in the PHY.

1.1.2 PDCP-RLC split

Forms of RAN split between PDCP and RLC layers already have applications in such features as Dual Connectivity and LTE-WiFi Aggregation (LWA). In particular, both technologies are based on the idea of providing additional transmission capacity by means of a connection point to the device. In case of Dual connectivity, some of PDCP PDUs are transmitted to another cell's RLC layer, similar to CU-RU interaction. In the case of LWA, PDCP PDUs are sent to an RU operating a different air interface technology. Main advantages of such RAN partitioning are modest bandwidth and latency fronthaul requirements. In fact, Figure 1.4 illustrates potential reduction in the required fronthaul bandwidth between RU and CU assuming a reduction in overhead per PDU due to header compression utilised in PDCP layer.

Advantages include centralised over-the-air encryption, greater potential for coordination of mobility and handover procedures, e.g. data forwarding from the old serving cell to a new serving cell can be simplified in a CU hosting base station functions above RLC layer. One of the drawbacks of PDCP-RLC split is limited potential for coordinated scheduling between multiple RUs. Due to the nature of the split, main scheduling operations are performed at the RU without an explicit exposure to PDCP processing.

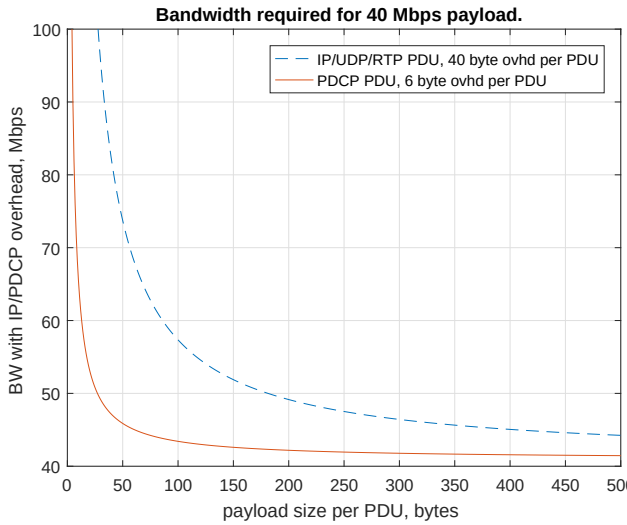

Figure 1.4: Illustration of the required bandwidth to transfer 40 *Mbit/s of user payload over PDCP PDUs with header compression and over datagrams with IP/UDP/RTP headers.*

1.1.3 RLC-MAC split

RLC produces segments containing PDCP layer PDUs on demand based on the notifications from MAC layer about the total size of RLC PDUs that MAC layer can transport. A functional split between RLC and MAC layer therefore can be impractical especially in the case of shorter subframe sizes in 5G air interface compared to LTE. In particular, shorter subframe sizes allow for more frequent decisions by the scheduler, adapting better to traffic demands or channel conditions, however this results in more frequent notifications to RLC from MAC specifying the size of the next batch of RLC PDUs. Flow control mechanisms and pre-emptive transmission of RLC PDUs of fixed size ahead of the MAC requests, as described in [6], can alleviate the resulting fronthaul latency and jitter requirements, however implementation of additional flow control mechanisms reduces the gains from centralising RLC processing.

1.1.4 Split MAC

MAC layer is responsible for multiplexing RLC data from logical channels onto transport channels provided by PHY layer, retransmissions via HARQ and nominally for scheduling, although as mentioned before, scheduling in practice may have dependencies across the whole base station radio stack. The case for separation of MAC layer is typically motivated by the desire to achieve performance gains and reductions in equipment complexity through offloading most MAC functions, such

as multiplexing and some of the scheduling decisions to CU, leaving in the RU only those functions that require real-time communication with PHY, such as HARQ and fine scheduling decisions.

Fronthaul links in this case would transport some form of pre-multiplexed higher-layer protocol datagrams along with scheduling commands. While the bandwidth requirements for this split should be comparable to the requirements for RAN splits at higher layers, the latency requirements are highly dependent on the realisation and interaction of scheduling functions in the CU and RU. One possible realisation of this split could involve CU delivering data to the RU in advance in the order decided in the higher-MAC scheduler, leaving the tactical decisions to the lower-MAC scheduler based on the HARQ and fine channel measurement reports [6].

While the final protocol architecture of 5G RAN is not known at the time of writing, implementation specifics and inter-dependency of lower and higher MAC layer scheduling functions may pose an interoperability challenge for base station equipment vendors and operators.

1.1.5 MAC-PHY split

The main motivation for this type of split is in enhanced capabilities from joint scheduling and coordination among RUs connected to a common CU, in addition to the benefits realised by splits at higher layers. The output from the CU to RU in this case would consist of TBs for further FEC encoding in PHY layer, and Transport Format (TF) carrying necessary information for correct processing of a TB. There are two main challenges associated with realisation of this split.

First, channel measurements utilised in the scheduling decisions need to be transported from RUs to the CUs in a timely manner in order to benefit from coordinated multi-cell scheduling and adaptive modulation schemes. For schedulers in 5G eNBs, potentially making decisions more frequently, fronthaul delays may reduce the ability fully benefit from shorter subframes and wider channel bandwidth, however exact performance trade-offs would need to be studied for specific 5G RAN technologies.

Second, data retransmissions handled through HARQ processes in LTE were designed based on compromises between performance and complexity of implementation. In particular, uplink HARQ responses from the eNB to the UE were designed to occur at pre-determined moments, hard-limiting time budget for baseband processing at the UEs and eNB. Based on LTE cell size of 100 km, the UE and eNB baseband processing needs to complete respectively within 2.3 and 3 ms in order for the UE uplink to be correctly acknowledged [5], with any transport latency reducing these budgets. Fronthaul requirements for next generation RAN would therefore depend on such factors as the type and number of HARQ processes expected to be supported by the base stations and UEs, as well as the typical cell sizes and use cases.

1.1.6 PHY split: FEC performed at CU

Assuming the latency, an possibly jitter, requirements on the fronthaul can be satisfied through the design of RAN protocols, appropriate choice of transport technology, or both, the PHY layer splits become possible offering opportunities for tighter coordination of transmissions and RU hardware simplification.

Among all PHY functions, the first candidate for a move to CU is FEC functionality. In LTE, FEC operates closely with MAC, providing multiple levels of CRC procedures, turbo or convolutional encoding on TBs that is very tightly coupled with MAC HARQ processes through redundancy version indications necessary for soft combining at the receiver side. The result of FEC procedures is further scrambled to provide an additional level of protection against interference, and resulting codewords are mapped onto the symbols of the modulation scheme chosen by the scheduler based channel conditions and other system indications.

Fronthaul transport for this split would therefore carry codewords together with additional scheduling information required to perform further PHY functions in the RU such as MIMO processing, precoding, antenna mapping and power allocation. As in the case of split MAC, bandwidth requirements would be comparable to higher-layer splits, while latency requirements would depend on the implementation of the scheduling functions in the RU and CU. However the latency and jitter requirements are likely to be tighter compared to split MAC or MAC-PHY split since the RU now has much less flexibility in the scheduling decisions and has to execute CU's commands on symbol-by-symbol basis.

1.1.7 PHY split: Modulation performed at CU

This is the first split where fronthaul would transport quantised in-phase and quadrature (IQ) components of the symbols from the modulation scheme chosen within CU to carry user and control data that was eventually encoded into codewords. The RU responsibilities would now be limited to conversion of the sampled frequency-domain modulation symbols onto time-domain through IFFT operation, followed by CP insertion, parallel-to-serial and digital-to-analogue conversions.

Fronthaul bandwidth requirements for this split are dependent of the number of symbols transmitted and on the number of bits used to quantise each symbol, plus any control information necessary for further PHY processing in the RU. Calculations for LTE 20 MHz channel bandwidth suggest approximately 900 Mbit/s throughput would be required to transport 150 Mbit of user data [6]. Channel bandwidths expected for 5G are likely significantly exceed current conventions, especially in mmWave frequencies, resulting in multi-gigabit throughput requirements to support this split.

1.2 Radio access network technologies, architecture and backhaul options

This section outlines some of 5G radio features and components of the emerging 5G architectures with potential impact on fronthaul/midhaul.

1.2.1 Modern network architecture

New architectural advancements in 5G need to be designed in accordance to the requirements of the majority of services and applications that will be carried out in the network. To create one network that is able to simultaneously satisfy the needs of multiple devices of different nature, it is necessary to introduce flexibility, programmability and virtualisation.

This section goes through the main features considered to be key in the design of the 5G architecture, with a specific focus on how a C-RAN with fronthauled access will impact the applicability or the final performance of such features.

1.2.1.1 Backhaul, Fronthaul and Midhaul

Cloud or C-RAN is a logically centralised set of eNB baseband and higher layer functionalities. The baseline architecture of a fully centralized Cloud RAN is shown in Figure 1.5. Based on the NGMN definitions [7], the Fronthaul spans distances between the RRH and the BBU. The classical form of fronthaul is a point to point link that transports baseband radio samples, also known as common public radio interface (CPRI), which is a synchronous interface transporting digitised base band signals over a symmetric high speed physical or radio link.

Based on the Metro-Ethernet Forum definition [8], Midhaul is the interconnection of a small-cell and a macro-cell via Ethernet links, with the assumption that the small cell is covered by an eNB. Without loss of generality, in this chapter the term midhaul is used to refer to a point to point link or network that transports signals beyond the physical layer; the term fronthaul is used to refer to the point to point link or network that transports physical layer signals.

As part of the core network, the Backhaul spans the section between the baseband and the evolved packet core (EPC) elements: MME, SGW, PGW, etc.

1.2.1.2 Network Function Virtualization

Network Function Virtualisation (NFV) has already shown great potential in the virtualization of core network functionalities, since it increases the flexibility of the core network implementation. In this way, control or user plane functions related to mobility management or gateways can be virtualised and placed in data centres anywhere in the network. The possibility of decoupling network functions allows configuring dedicated core networks, so that the system can better meet the service requirements [9, 10].

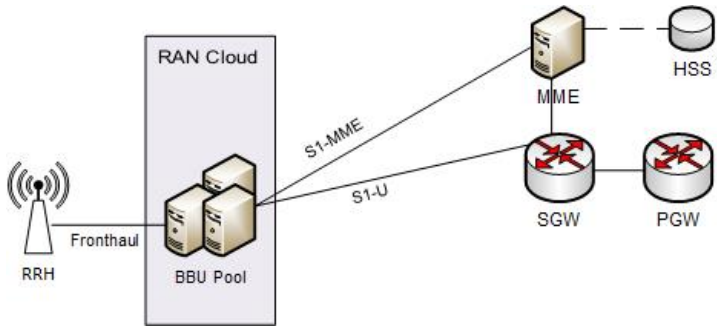

Figure 1.5: Cloud RAN General Architecture Overview

Thus, some of the NFV general advantages are that the virtualised version of the functionality can be placed in standard IT servers and switches, reducing the cost of proprietary platforms, and allowing for more flexible implementations.

Virtualization is very challenging in RAN low layers due to its real-time nature; synchronization requirements that ensure a good wireless signal processing performance are in the order of microsecond and nanosecond. However, the base band radio processing may be virtualized with the use of software defined radio (SDR) techniques, and allow the mobile network to introduce the advantages of NFV in the RAN.

1.2.1.3 Network Slicing

5G networks should have a high degree of programmability, configurability and flexibility to support heterogeneous deployment solution, scenarios and applications: to this aim, a key aspect of 5G systems is considered the network slicing [11]. According to NGMN, a network slice (a.k.a. 5G slice) is composed of a collection of 5G network functions (i.e., specific features enabled on Control/User-planes) and specific radio access technologies (RAT) settings that are combined together for the specific application or business model [12]. This means that all slices will not contain the same functions, i.e., a 5G slice will provide only the traffic treatment that is necessary for the service, and avoid all other unnecessary functionalities [13].

A 5G slice can span all domains of the network:

- specific configurations of the transport network;
- dedicated radio configuration or even a specific RAT;
- dedicated configuration on the fronthaul and the backhaul.

For any of the mentioned domains, network slicing determines reservation of resources to handle the enabled slices in order to fulfil the QoS constraints/features of the slice. This might introduce issues in case of fronthaul solutions which do not offer mechanisms for resource reservation/prioritisation.

Slicing based on fronthaul or midhaul availability

A slice is commonly considered as a set of QoS features to be guaranteed in the whole network domain, where QoS from the slicing point of view is considered as the whole set of features which involves not only the typical QoS parameters (such as data rate, latency, jitter, reliability) but also other service-related functionalities (such as mobility, security, authentication, etc.) characterising the slice.

From one perspective, network slicing poses challenging requirements on the fronthaul/midhaul interface(s) which should be open and flexible, offering multi-vendor operation and good forward and backward compatibility, while providing options for transport bandwidth reduction. From another perspective, the characteristics of a given fronthaul/midhaul interface (latency, supported data rate, reliability, etc.) poses constraints on the possible slices to be enabled on that segment of the network. Under this point of view, the availability of different fronthaul, midhaul and backhaul solutions drastically influences the slices to be enabled: this might introduce slice holes in the provisioning of network slices only in those areas where the fronthaul/backhaul characteristics fulfil the requirements of the slice with consequent implications on the business models of the provider.

Slicing for multiple RAT aggregation

5G wireless systems are likely to joinly utilise 3GPP and non-3GPP RATs [14]. In this field, the fronthaul or midhaul technology available for each different RAT influences the network slicing in terms of selection of the best RAT as well as in terms of inter-RAT management. Usually the best RAT is the one which is able to guarantee the best performance on the radio interface (e.g., lowest interference, highest data rate or lowest congestion); nevertheless slicing takes into consideration the whole network domain and this might introduce issues when selecting a specific RAT to handle a specific slice (e.g., a RAT able to offer high data rates but with limited capabilities in terms of latency on the underlying infrastructure is not suitable for slices that need to guarantee low-latency handover mechanisms). From another point of view, the aggregation of multiple RATs might introduce issues for inter-RAT communications: in this case, the latency experienced over the fronthaul or midhaul (as well as on the backhaul) could avoid to achieve the goal of a seamless handover procedure because of the need to perform additional authentication/security mechanisms when moving from one RAT to another one.

1.2.1.4 Mobile Edge Computing

Mobile edge computing (MEC) can be considered as a service provided through slicing. The idea behind this approach is that content, services and applications might be available at the edge of the network, i.e., as close as possible to the end-user. This will help the operator to generate additional revenue by saving the utilisation of network resources.

Although there is not an obvious connection at first glance, mobile edge computing and Cloud RAN are synergistic technologies. Cloud RAN with virtualisation of the eNB functions uses general purpose processors (GPPs) to run baseband func-

tions, at a data centre, or locally at the venue itself. The same GPP platform can be used for MEC applications which run at the edge and not the core, even though the edge is centralised.

From a capacity point of view, mobile edge computing does not introduce any additional requirements compared to legacy data provisioning procedures on the fronthaul. The main benefits are in terms of backhaul offload (as once data is available on the edge, no transmission in backhaul will be further necessary). Nevertheless, the backhaul might limit the provisioning of mobile edge computing services because of its capacity. Indeed, when a content, service or application needs to be moved to the edge, the backhaul is expected to handle the traffic needed to move the service to the edge plus all the related signalling traffic necessary to manage this shift. As a consequence, low-capacity backhaul links might limit the provisioning slices involving mobile edge computing.

From a C-RAN (and thus, fronthaul and midhaul) perspective, it is worth to underline the following aspect. When a data/service/application is available on the edge it is expected to be provided quicker to the end user, as the traffic will not reach the core network. In this case, the delay that may limit the service latency is the Layer 2 reliability procedure (hybrid automatic repeat request HARQ), which poses a high requirement to the fronthaul not to further increase that delay, given that is the most susceptible part of the network when increasing latency.

1.2.1.5 Service Virtualisation

Service virtualisation allows network functionalities to become a software package that might be moved in the network according to the specific requirements/conditions of the network in a given instant/period. Service virtualisation puts constraints in terms of reliability of the fronthaul/backhaul, to avoid retransmissions when a service is moving from one location to another one. In addition, fronthaul/backhaul latency needs to be considered as it will affect the time to run of the virtualised service.

A specific scenario for service virtualisation is the inter-connection with non-3GPP RANs, where solutions such as moving authentication functionalities close to the edge (i.e., AAA services run in a C-RAN instead of being provided by the PGW) can be exploited to offload the backhaul links and to cut delays. Nevertheless, this is only applicable in case the fronthaul is suitable to offer high capacity and low latency.

1.2.2 5G technologies and use cases

The new generation of RAN that will enable 5G is going to consider most of the current innovative features that are being discussed today by both standardization bodies and research community. 5G is going to co-exist with the evolution of the LTE-A standard, and new radio access technologies (RATs) are also going to be integrated, to form a unique heterogeneous network, capable of providing multi-connectivity from different points of view: multiple scenarios, multiple radios, and multiple cells.

In general, 5G enablers are all those technologies that can contribute to a large scale cooperative network, and at the same time enable the main features of 5G, such as enhanced mobile broadband experience or mission critical machine type communications with ultra-reliable and ultra-low latency communications. It is particularly interesting to identify the applicability of such 5G enablers in a C-RAN architecture and how the underlying fronthaul network impacts the overall network performance.

1.2.2.1 Integration of multiple air interfaces

Besides the enhancement of the already mature LTE-A, a set of new radio access technologies is required to satisfy future requirements in terms of spectral efficiency and availability and throughput. In particular, these new RATs need to exploit efficiently higher frequency bands and bandwidths, which require the use of new time and frequency numerology:

- Support for low latency: Shorten transmission time interval (TTI) to 0.2-0.25 ms [15]
- Low overhead to cope with time dispersion: Cyclic Prefix of 1 s [16, 17]
- Common clock with LTE to support different RATs
- Robustness against phase noise and frequency offset: use of large subcarrier spacing [15].

Since the new waveform requirements are set to satisfy the 5G ambitions, the challenges to be imposed in a C-RAN architecture are then transferred: very low latency to satisfy the TTI reduction and high capacity to satisfy the increase in spectral efficiency.

In a 5G network, the RAN architecture needs to enable the aggregation of multiple RATs, including new air interfaces and legacy ones i.e., 4G/3G and possibly fixed services. In this sense, system convergence and integration are two musts for 5G.

Hence, high frequency bands, in the range of millimetre wave (mmW), need to be integrated with the use of low frequency bands, such as LTE and 4G communications. Work in [18] discusses the integration of LTE and new 5G air interfaces, where UEs are capable of simultaneously transmit and receive in both radio technologies. In a centralised RAN context the architecture that supports the integration needs to be evaluated, since sharing resources is more than likely. This work proposes an architecture relying on common protocols, called integration layers. Due to difficult synchronisation at lower layers, the integration point is recommended to be located at least at the PDCP and RRC layers. It is assumed that the LTE and the 5G air interface are in a co-located RAN. Indeed, similar ideas as in the Dual Connectivity architecture discussed in [19] are presented for the user plane aggregation; single data flows can be aggregated over multiple air interfaces, or different flows may be mapped to different air interfaces. In a scenario with RAN cloudification, where PDCP and RRC layers of integration are centralized, there is a strong requirement

on the midhaul network to support multiple RAT data integration, mainly in terms of capacity and latency.

Recent work by the 3GPP involve two options of data offloading, one is to the wireless LAN via WiFi/LTE aggregation and another one is using LTE in unlicensed spectrum. 3GPP has defined several WLAN offloading mechanisms which rely on the connection between the LTE core network and WLAN. The recent work on data aggregation at the LTE base station allows for better control of offloading with improved system and user performance while leveraging the existing LTE features of carrier aggregation (CA) and Dual Connectivity [20, 21].

1.2.2.2 Support for Massive MIMO

With massive MIMO, the system uses antenna arrays, with a few hundred antennas, that simultaneously serve many tens of terminals in the same time-frequency resource. The basic premise behind massive MIMO is to capture all the benefits of conventional MIMO, but on a much larger scale [22].

In a centralized RAN scenario, the use of Massive MIMO would dramatically increase the data rate requirement in the fronthaul, proportionally to the number of antennas. In fact, if digitised radio signals are transmitted through the fronthaul links capacity requirements would increase to the order of 2 Tbps for 500 MHz bandwidth with the use of mmW, as shown in [23]. As a matter of fact, in [24] authors estimate the line rate to transmit sampled radio signals over the fronthaul for a 20 MHz bandwidth using 64 antennas: nearly 80 Gbps are required.

Therefore, to allow for efficient C- RAN implementation, other architectures need to be evaluated. Work in [23] suggests that beamforming operations can be shifted close to the RRH to alleviate these data rate requirements. In the same line, [25] suggests that MIMO precoding, detection and modulation/demodulation functions should be located at the RRH; thus the information to be transported are modulation information bits. In this case the required bandwidth for a C-RAN is one order of magnitude lower that transporting sampled base band signals.

1.2.2.3 Massive Cooperation: Multi-Connectivity Networks

Multi connectivity is a disruptive technology where devices will simultaneously transmit and receive to and from different access points. One of the key issues when moving towards a user or service centric network is to provide the mobile network with sufficient flexibility to select the serving cell(s) that better suits the device or service requirements.

In fact, the whole concept of cellular association is believed to change with 5G, and one device has no singular connection, but a set of antennas that provides service. In such a context, cooperative and decoupling techniques are massively exploited. Typical cooperative features include, joint transmission and reception, coordinated scheduling and beamforming, enhanced inter-cell interference coordination, among others. As well, decoupling techniques being considered by the literature are the Control and User plane split and the Uplink and Downlink split.

Both, the 3GPP and the research community have contributed actively with new architectural alternatives that enable easy cooperation among base stations. Cood-

inated multi-point (CoMP), Dual Connectivity or inter-site Carrier Aggregation are some examples in the available literature. It is true that to enable efficient cooperative networks the information exchange among the different access points is essential. Current LTE-A standards consider this information exchange through the X2 low latency interface, which at some extent can limit the performance of cooperative applications. In this sense, centralised processing of control information and management facilitates the implementation of such features by localising the information exchange. In this sense, centralised cooperative processing requires a fronthaul or midhaul network to aggregate traffic from multiple access points.

Inter-cell interference is also a crucial aspect in massive cooperation. Each cell autonomously restricts the resource allocation with the objective of limiting interference between adjacent cells. To this end, cells require data exchange which can tolerate significant latency. Nowadays inter-cell interference coordination (ICIC) techniques may be applied with any level of centralization, since either fully centralised or distributed architectures allow for either static or dynamic ICIC algorithms. However, the fast coordination among medium access contron (MAC) layers in a C-RAN environment can result in the integration of smarter enhanced ICIC (eICIC) techniques that allow to efficiently carry out high-speed and ultra-reliable communications in the cell edge (i.e., where the UE suffers from strong interference).

Coordinated Multi-Point Transmission and Reception

CoMP refers to the wide range of techniques that enable dynamic coordination among multiple cells that belong to the same cluster. Two major forms of CoMP can be recognised in 4G LTE-A: Joint Processing (JP) transmission or reception, and Coordinated Scheduling or Beamforming (CS or CB, respectively).

In particular, CS/CBs main goal is to identify the worst interferer and avoid collisions by preventing the use of the most destructive precoding matrices (precoding is the process in which the incoming layered data is distributed to each antenna port). To this end, cells need to negotiate their beams and exchange MAC layer information; if this is done through the X2 link then bandwidth and latency requirements may not be sufficient to satisfy a high number of users. In [26] it is concluded that a MAC level centralization can strongly limit the performance of CoMP due to additional latencies for data paths and channel state information (CSI) feedback over the midhaul network. On the other hand, PHY layer centralization can allow all kind of CoMP schemes. Along the same lines, work in [27] discusses that most of the CoMP schemes (Downlink joint transmission and CS/CB) can be achieved with a MAC layer centralisation, provided that information on the CSI is forwarded to the central scheduling entity. The CSI acquisition process for multiple RRHs is quite demanding in terms of overhead and reporting delays and both have a direct impact on the system performance.

The particular cases of interference rejection combining (IRC), or successive interference cancellation (SIC) the PHY layer should be centralised at some extent to avoid additional communication among the central and remote unit for CoMP purposes. Similarly, work in [28] proposes an architecture with physical layer centralisation to support UL JP techniques.

Dual Connectivity

Dual connectivity is one of the 3GPP potential solutions to improve user performance by combining the benefits of the Macro cell coverage and the Small cell capacity. This new technology introduced in Release 12 [19] is defined as the simultaneous use of radio resources from two eNBs connected via non-ideal backhaul link over the X2 interface. One of the new advances is the introduction of the bearer split concept, which allows a UE to receive simultaneously from two different eNBs, known as Master eNB and Secondary eNB, MeNB and SeNB respectively. The 3GPP has proposed several architectural alternatives for downlink dual connectivity in [19], an architecture with a centralised PDCP layer can effectively support the dual connectivity with user plane bearer split.

Downlink and Uplink Decoupling

The UL and DL split, or DUDe, has been covered by the literature recently as a means to reduce the UL and DL imbalance that occur in heterogeneous networks, due to the transmit power disparities between small and macro cells. The DUDe technology is the most device-centric enabler feature being investigated so far; it allows to have two different serving cells, one for the DL and another for the UL. As well, the UL feasibility of adopting the bearer split (i.e., dual connectivity) has been argued by the literature in terms of power consumption, and UL data should be either transmitted directly to the best cell in terms of received power [29].

Architecture solutions that allow to support DUDe while maximizing the capacity must include at least a shared MAC layer among both serving cells, since Layer 2 control information needs to be forwarded from one serving cell to another (i.e., HARQ protocol acknowledgements). As well, Layer 3 RRC ought to be centralised and shared among both serving cells, since parallel RRC connections would add too much complexity in the UE side [30].

Device to Device Communications Integration

Devices itself can as well collaborate in the RAN, by allowing direct transmission between devices controlled by the serving eNB. One of the key aspects in Device to Device (D2D) is the control plane information, managed by the eNB. Control information sharing in the network can improve the spectral and energy efficiency of the devices having direct communications. Since cognitive and instantaneous decisions can be made in the RAN, to allow for improved management of resources and the contention of the interference levels, a centralized control and resource management can potentially improve the outcomes of D2D.

1.2.3 Practical backhaul technologies

The fronthaul interface (as the transmission of base band sampled signals) distance is limited by the implementation of the HARQ protocol in the uplink in LTE (i.e., 8ms) since is the lowest round trip time (RTT) timer imposed by the MAC layer. Due to the synchronous nature of the HARQ in the UL and to its explicit dependency

with the sub-frame number, is the MAC procedure that poses the most stringent requirement on latency. Relevant fronthaul solutions must fulfil the requirements outlined for CPRI transmissions, and assure a correct performance of all procedures. Fronthaul options available for native CPRI transport discussed in [4], are classified into technologies that can either multiplex or perform addressing to the native CPRI signal:

Dark Fibre

Upon availability, it is a very straightforward solution, but requires high CAPEX. Point to point distance is limited by HARQ timers. Only one link can be transmitted since no multiplex is carried out.

WDM type

- Passive WDM: allows for transmission rates up to 100 Gbps, and distance is limited by latency requirements (i.e., HARQ). Performance is similar to dark fibre but better reuse of facilities due to the multiplexing capabilities.
- CWDM: A single fibre with bidirectional transmission can be used to reduce costs.
- DWDM: Good for large aggregate transport requirements.
- WDM-PON: Alternative to DWDM, WDM PON with injection-locked SFPs.

Microwave

In Millimetre radio solutions the distance is capped due to processing and modulation (few hundred meters to 7 km). Capacity is typically between 1.25 Gbps to 2.5Gbps. However, [5] highlights that considering channel sizes, physical modulations and coding rates that are supported by recent implementations, a capacity in the range of 10 Gbps can be achieved. Microwave solutions require high bandwidth availability and high spectrum.

Optical Transport Networks

Good network solution that meets the jitter requirements of CPRI. One good advantage is forward error correction (FEC) which makes links less sensitive to bit errors, however the FEC added latency would further reduce the achieved distance.

XGPON and GPON

GPON is used in connection with FTTH, which is available in many urban areas. Distances are limited to 20 km, and it is impractical for fronthaul applications because it is asymmetric and bandwidth limited.

Network solutions that are able to transmit native CPRI are those that can cope with the increased capacity demands and very low jitter; table 1 summarised the most typical fronthaul network options.

Table 1.1: Transport Network Capabilities

Transport	Throughput	Latency	Multiplexing capabilities
Dark Fiber	10 Gbps+	5 μs/km	None
DWDM/CWDM	100 Gbps+	5 μs/km	High
TDM - PON	10 Gbps	Dynamic BW allocation $>$ 1 ms	High
GPON (FTTx)	DL: 2.5 Gbps UL:1.25 Gbps	$<$ 1 ms	High
EPON (FTTx)	1 Gbps – 10 Gbps	$<$ 1 ms	High
OTN (FTTx)	OTU4 – 112 Gbps	FEC latency	High
Millimeter Wave	2.5 Gbps – 10 Gbps	0.5 ms – 100 μs	High
xDSL (G.fast)	10 Mbps – 100 Mbps (1 Gbps)	5 – 35 ms (1 ms)	High

1.3 Current fronthaul solutions

1.3.1 CPRI in C-RAN

CPRI is the most common transmission mode between the BBU and the RRH, and it carries sampled base band signals. Thus, capacity demands for native CPRI transmission are based on several factors. The fronthaul bandwidth is proportional to the systems available bandwidth, the number of antennas and the quantisation resolution (the number of bits per I or Q sample are 8-20 bits for LTE [31]) and in any case is dependent on the cell load and the user data rates [32]. For example, macro sites generally have three to six sectors combining different mobile RATs (i.e., 2G, 3G and 4G in multiple frequencies). According to the NGMN alliance report, one MCell generates approximately 15 Gbps of uncompressed sampled base band signals [26].

Hence, the basic fronthaul requirements to be considered for the transmission of native CPRI are [7]:

- Capacity: from CPRI option 3 to 9, i.e., 2.457 Gbps to 12.165 Gbps [32]. Note that this capacity has been calculated considering LTE bandwidth configurations and different number of antennas: CPRI option 3 capacity link is for an LTE bandwidth of 10 MHz with 4 antennas, or a 20 MHz bandwidth with 2 antennas. As remarked in section 1.2 the use of multiple antennas or mmW can increase capacity demands to the order of terabytes.

- Jitter: in the range of nanoseconds, according to the physical layer time alignment error (TAE) (i.e., 65ns)

- Latency: maximum round trip delay excluding cable length 5 s, to assure the efficient implementation of frequency division duplex (FDD) inner loop power control [32].

- Scalability Support for multiple RATs and RAN sharing

- Distance: 1-10 km for most deployments, 20-50 km for large clouds

Also, the fronthaul must deliver synchronisation information from the BBUs to the RRHs, which is natively supported by CPRI through the control and management. Given these stringent requirements, new scalable and efficient solutions need to be explored in the context of CPRI. One option is the CPRI compression that allows to significantly reduce the required bitrate while assuring that it meets the transparency requirements of the CPRI. . Also the literature is exploring the so-called functional split, where some of the eNB functionalities remain in the remote unit, allowing to relax the bandwidth and delay constraints.

Also, the fronthaul must deliver synchronisation information from the BBUs to the RRHs, which is natively supported by CPRI through the control and management.

Given these stringent requirements, new scalable and efficient solutions need to be explored in the context of CPRI. One option is the CPRI compression that allows to significantly reduce the required bitrate while assuring that it meets the transparency requirements of the CPRI. Also the literature is exploring the so-called functional split, where some of the eNB functionalities remain in the remote unit, allowing to relax the bandwidth and delay constraints.

1.3.2 CPRI Compression

Compressed CPRI may be used to reduce capacity requirements in places where fronthaul bit stream transport is limited, CPRI compressed and decompressed function may be used, which can provide 2-3 times more utilization.

Point to point fronthaul compression methods, where the central unit independently compresses each remote unit baseband signal, includes techniques such as filtering, block scaling and non-linear quantization [33]. These solutions allow to remove redundancies in the spectral domain by down sampling the input signal, mitigate peak variations. Compression reduces the data rate by a factor of three with respect to uncompressed CPRI signals. For example, the work in [34] applies the same techniques obtained good system performance with and 1/3 compression rates in a practical propagation environment. In particular, results show that compression can effectively reduce the amount of data rate transmission on the CPRI without comprising the actual baseband data at low compression ratio.

Moreover, [35, 36] presents another approach for compression in the downlink, named multivariate fronthaul compression. The proposed solution is based on joint design of precoding and compression of the baseband signals across all base stations; results show that this CPRI compression scheme outperforms the conventional approaches of point to point compression. As a rule of thumb, compressed CPRI techniques are seen to reduce the fronthaul rate by a factor around 3 [27].

1.3.3 Fronthaul or Midhaul over Ethernet

When evaluating CPRI as the transport service for the C-RAN fronthaul some issues arise:

- provides no statistical multiplexing gain (due to the continuous data transmission)
- how to manage and provide service level agreements in the fronthaul service [32, 37].

Recently, in both research community and industry it has been discussed if Ethernet networks can be used to transmit the physical layer signals, which would initially imply some more framing overhead and struggle to meet the requirements. However, on the other hand, using Ethernet links to encapsulate sampled base band signals brings several advantages [23]:

a Lower cost-industry standard equipment

b Sharing and convergence with fixed networks

c Enables statistical multiplexing gains when signal has a variable bit rate

d Enables the use of virtualisation and orchestration

e Allows network monitoring

f Allows managing the fronthaul network (i.e., Fronthaul as a service). Path management enables the use of virtualisation and software defined networks (SDN)

Originally, Ethernet is a best effort based technology, and it is not designed to meet the low jitter and latency requirements for base band signals transmission, i.e., CPRI. In general, works considering CPRI over Ethernet [38] suggest dedicated links between RRH and BBU, and the Ethernet network is enhanced with additional features to satisfy stringent latency and jitter constraints.

In particular, the IEEE is actively working in new standardisation efforts. In particular, IEEE 1904.3 Radio over Ethernet (RoE) for encapsulation and mappers [39], the main objectives are, (a) to define a native encapsulation transport format for digitised radio signals, and (b) a CPRI frame to RoE (i.e., Ethernet encapsulated frames) mapper. Also, several enhancements to allow the transport over time sensitive traffic have been presented in the IEEE 802.1CM Time-Sensitive Networking for Fronthaul Task Group [40]. Solutions such as frame pre-emption or scheduled traffic allows to better manage packets in Ethernet and reduce jitter [38].

Other solutions to support CPRI over Ethernet is the use of timing protocols, such as Precision Time Protocol (PTP) that provides synchronism through the exchange of time stamped packets; however, [23] remarks that bitrate as well as delay and delay variation requirements result in significant implementation challenges when using Ethernet in the fronthaul to transmit digitised radio signals, i.e., CPRI.

1.3.4 C-RAN Integration in 5G: Feasibility Discussion

5G can be seen as a tool box of enabling technologies,where a specific set of these technologies can potentially provide the requirements of a given service. Therefore, when assessing the suitability of C-RAN with a fronthaul network in 5G, the main conclusion is that a flexible network architecture should be considered. Depending on the service and the underlying network infrastructure the RAN can be configured to employ any set of technologies and split functionalities. Cloud RAN and fronthaul/midhaul is all about heterogeneity. An overall conclusion of the transport network for fronthaul or midhaul and functional split trade-off discussion is that there is no configuration that can satisfy all the requirements for 5G simultaneously:

- High levels of processing in the BBU may increase latency due to fronhtual links, and pose challenging capacity requirements for CPRI transmission, especially if higher number of antennas are employed;
- High levels of processing in the RRH may impair the cooperative capabilities and neglect the multi connectivity features;
- Low latency or congested fronthaul networks may impair the correct function of reliability algorithms (i.e., HARQ in Layer 2), and increase latency and decrease user throughput;
- The lower layers are extremely latency critical whereas the higher layers are not;
- Network slicing capabilities and virtualisation outcomes are very much dependent on fronthaul transport networks;
- All these requirements in latency to maintain the high synchronicity between the lower layers are going to be more demanding when implementing centralisation with 5G radio access technologies.

Based on this, no rule of thumb can be proposed to efficiently implement a partially of fully centralised RAN. There is a strong trend from industry and academia to leverage packet switched networks, such as Ethernet, to provide a cost-effective transport network solution that allows to converge backhaul with fixed backbone networks. In the practical environment, Korean operators KT big picture of 5G and re-design of the RAN involves the use of functionality split to support packet networks that deliver Ethernet frames. The lowest level functionality split option would involve the transmission of MAC PDUs over the fronthaul, meaning that the entire PHY layer is kept in the radio head side [41]. Table 1.2 summarizes the main 5G enablers described in this document with its optimal Cloud RAN configuration.

Splits within the PHY layer, using either CPRI for sampled radio signals or packet switched networks after resource de/mapping, have the main advantage of maintaining the PHY, MAC and RLC layers operating together and keep full cooperation gains of the centralised architecture. Conversely, 5G latency is the key

Table 1.2: 5G Enablers mapped to the optimal Cloud RAN configuration

5G Enabler	Cloud RAN Configuration
Low latency RAT	Lower layers close to radio unit
Multiple RAT integration	Common PDCP
LTE in Unlicensed band	Common MAC
Massive MIMO	Lower layers close to radio unit
Multi Connectivity	High level of centralisation
Network slicing and Service Virtualisation	High dependency on transport network performance and availability

performance indicator, which may indicate that these layers need to stay as close as possible to the radio site.

In this sense, to be able to integrate all the 5G enablers and satisfy its individual requirements for implementation, base stations (RRH+BBU) functionalities should be completely flexible; allowing efficient configurations of slices that satisfy the services requirements and the UE needs at all times, considering each key performance indicator separately.

1.4 Market direction and real-world RAN split examples

Backhaul networks to connect mobile radio basestations to core networks are a key area of interest for both mobile operators (as they form a significant cost element) and fixed operators (who often provide the backhaul infrastructure). Over the lifetime of LTE, we have seen an evolution from the initial S1 backhaul interface to the CPRI based "fronthaul" technology adopted by Cloud RAN (CRAN) and finally the proposal of alternative basestation splits to reduce the overhead that CRAN fronthaul imposes.

1.4.1 Mobile backhaul

The architecture of a 4G mobile network as standardised by 3GPP is a flat structure simply involving a network of radio basestations (EnodeBs) which are linked to the Evolved Packet Core network via backhaul connections (Figure1.6). The EnodeB sites incorporate both radio and baseband processing functionality linked to antennas located at the top of the associated masts by means of co-axial feeders. However, these can exhibit high losses and therefore some vendors moved to deployments which placed a Remote Radio Head or Radio Unit (RRH or RU) incorporating digital to analogue and analogue to digital conversion and power amplification next to the antenna and connected it via optical fibre to the Baseband Unit or Central Unit (BBU or CU) providing the packet processing and scheduling functions at the base of the mast.

The backhaul connections linking the remote ENodeBs to the EPC form the important S1 reference point which is standardised by 3GPP RAN3 working group.

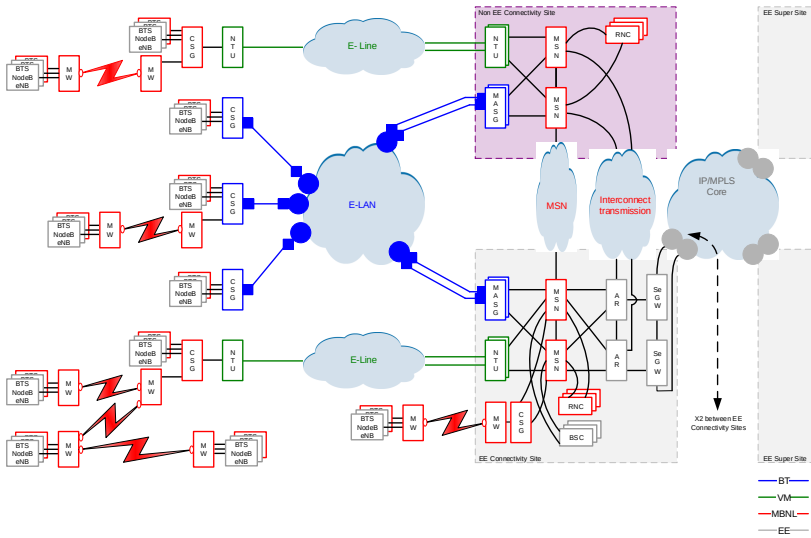

Figure 1.6: Practical end-to-end mobile network architecture. Less commonly used terms are: MSN – Multi-service network, NTU – network termination unit, MW – micro/millimeter-wave backhaul, SeGW – security gateway. Diagram credit: Andy Sutton, Principal architect at BT.

In addition, eNodeBs may be connected to each other by means of the X2 interface. Although the RAN backhaul network would appear to be simply a collection of point-to-point links, the reality is usually much more complex. A practical radio mobile network architecture is illustrated in Figure 1.6, with key features including:

1. Shared RAN – various aspects of the RAN are shared between different mobile operators, in this particular example through a joint venture MBNL. This includes the 3G radio, plus the towers and backhaul at the majority of sites.

2. The shared RAN arrangement is facilitated by a fibre ring providing resilient high-capacity connectivity between the core network and RAN Connectivity Sites (in the order of 10 to 20 serving the UK).

3. From these connectivity sites, point-to-point fixed connections provide backhaul connectivity to major basestation sites.

4. Finally, second-tier, smaller basestation sites are server from the major sites mainly using point-to-point microwave links.

1.4.2 Centralised or Cloud RAN

As discussed earlier, the C-RAN architecture was first proposed in 2010 by IBM [2], and then described in detail by China Mobile Research, [3] extends the BBU-RRH concept described above by moving the BBU from the basestation site to a centralised site and co-locating or pooling with BBUs from a number of other basestations to form a Centralised-RAN.

In a further step, the BBU functions may be realised in software components deployed within a shared computing platform or compute cloud. This virtualised BBU pool running on general purpose processors is referred to a Cloud-RAN.

The key barrier to implementation of C-RAN is the bandwidth and latency requirements of the fronthaul connection. However, for operators, C-RAN approach offers operators a number of potential advantages:

1. Simplified basestation installation: The BBU is the most complex element of the basestation and its removal can facilitate a reduction in the physical footprint and simplify the installation of the basestation.

2. Reduce Power Consumption: China Mobile have estimated that 72% of the power used by the network is expended in the RAN and that, nearly half of this is consumer by air conditioning. Consolidating the BBU functions would allow most of this power to be saved as RRH elements can typically be air-cooled.

3. Increased spectral efficiency: Due to the characteristic of LTE that all cells generally operate on the same frequency (or set of frequencies), inter-cell interference is often the limiting factor on cell capacities and throughput. This manifests itself as a difference of up to 10x between cell centre and cell edge throughput. Two approaches may be taken to mitigate interference effects minimising interference and exploiting it constructively.

   One approach is to use inter-cell interference control which allows eNodeBs to co-ordinate with neighbouring cells over the X2 interface to ensure that the resources that are being used to communicate with a cell-edge mobile are not used by neighbours at their cell edge. eICIC (enhancedICIC) also addresses the time domain by introducing Almost Bland Subframes (ABS) which allows eNodeBs to negotiate (again via X2) an interval when its neighbour cell will mute its signal allowing it to send information to a cell-edge UE. CRAN does not directly improve either ICIC or eICIC performance but it facilitates the establishment of the X2 interfaces on which they depend.

   An alternativeis in utilisation of Co-Operative Multi-Point (CoMP) techniques which attempt to utilise interference. In CoMP several cells co-operate to serve cell-edge UEs – for example by transmitting the data to a specific mobile from more than one cell so that the signals combine additively. However, this requires very tight synchronisation between the cells in the CoMP group, which can be achieved if all the cells are served from a centralised BBU pool.

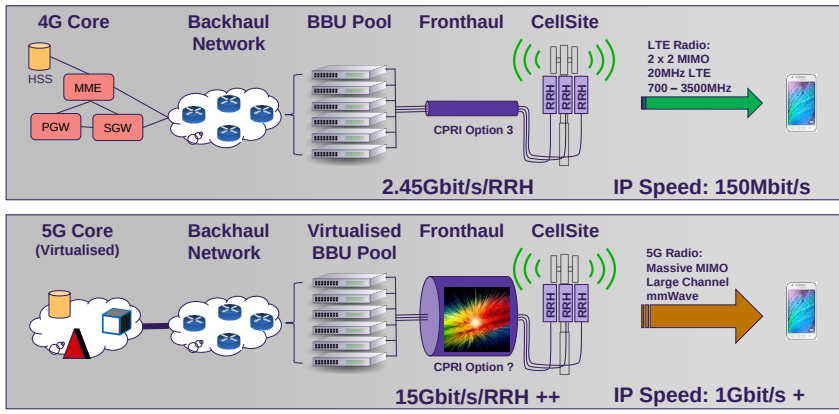

Figure 1.7: Correspondence between fronthaul and backhaul throughput requirements for 4G and 5G. High-Speed Radio ambitions to increase throughput by over 100x may result in similar increases needed on Backhaul/Fronthaul.

4. Reduced upgrade and maintenance overheads: The CRAN architecture co-locates BBU functions at a much smaller number of locations and therefore repair and upgrade activities are also concentrated at these locations, significantly reducing overheads such as travel time. In addition, if equipment failure does occur, there are many more opportunities for absorbing the problem through re-configuration within the BBU pool with automation reducing further the need for human intervention.

5. Increased BBU resource flexibility and utilisation: The transition to a centralised, virtualised deployment makes it possible to pool BBU resources which can be shared across basestations. This can greatly increase the utilisation efficiency of these functions.

1.4.3 Forward view to 5G

As shown in Figure1.7, a typical deployment uses 20MHz LTE bandwidth with 2T2R MIMO and therefore the CPRI rate to each radio head is 2.45Gbit/s. Assuming that the basestation site comprised three sectors, at total of 7.37Gbit/s is required (typically across 3 fibre pairs).

Compared to the usable data rate of 150Mbit/s, the figures shown above indicate that CPRI introduces an overhead of over 1600%. Taken together with the stringent latency requirements of CPRI, this make direct point-to-point optical fibre connections the only suitable technology for fronthaul links. Moreover, the introduction of 5G radio technologies such as massive MIMO and wider channel bandwidths will also make this situation even worse. Figure 1.7 also shows a possible 5G fronthaul

scenario with a 5G radio systems supporting a real-world capacity between 1 and 3Gbit/s – translated into a CPRI fronthaul rate up to 45Gbit/s. Such speeds will drive up the costs of the termination devices for the optical fibre connections and delay or prevent the implementation of the required signal processing in virtualised software functions.

1.4.4 Industry 5G fronthaul initiatives

The bandwidth issues described above, taken together with the other challenges of current fronthaul protocols such as strict limits on latency and jitter and the inflexibility of the deployment architectures that they impose, have driven operators to look for alternatives which can deliver the benefits of CRAN without the costs.

An important vehicle delivery of this ambition is the work on the IEEE 1914 Next-Generation Fronthaul Interface (NGFI) specification. This will be based on Ethernet to deliver a packet-based, multi-point to multi-point interconnect using statistical multiplexing which will handle data security, quality of service and synchronization. The NGFI activity is supported by China Mobile, AT&T, SK Telecom, Nokia, Broadcom, Intel and Telecom Italia.

In addition, a number of operators have announced details of their plans covering C-RAN and fronthaul. These typically include the support need for further study of RAN architectures to better understand both the trade-off between centralisation/virtualisation gains and fronthaul capacity and how efficient interworking between 4G, 5G and WLAN might be achieved in a CRAN architecture. In particular, investigations into the number of required splits between a software-based Control Unit (CU) and a remote hardware-based Access Units (AU) are ongoing. One important component from operators' perspective is that the interface between the CU and DU units is open and standardised to support multiple transport solutions required in a practical network deployment.

1.4.5 Split MAC trials

The BBU-RRH functionality separation in this case is within the MAC layer itself. Typically the MAC scheduler, i.e. the upper part of the MAC, is centralised within the BBU. This enables the possibility of coordinated scheduling amongst cells, whilst the lower MAC and HARQ functionality, given the 8ms critical timing aspects of the HARQ cycle, are placed on the RRH in order to remove any additional latency from the fronthaul link. Having the lower MAC and HARQ on the RRH allows the remote scheduler to operate with the PHY layer in a semi-autonomous way and at a sub-frame level. Some tight coordination in the form of scheduling commands and HARQ reports is required for the communications between the MAC and HARQ scheduling. The required capacity and latency is 150/75Mbps (similar to S1/X2 rate) for a 20MHz carrier and ¡6ms, respectively. Standard packetized technologies for fronthaul transport such as Ethernet can be sufficient to ensure good coordination between the two scheduling processes for some of the current basic

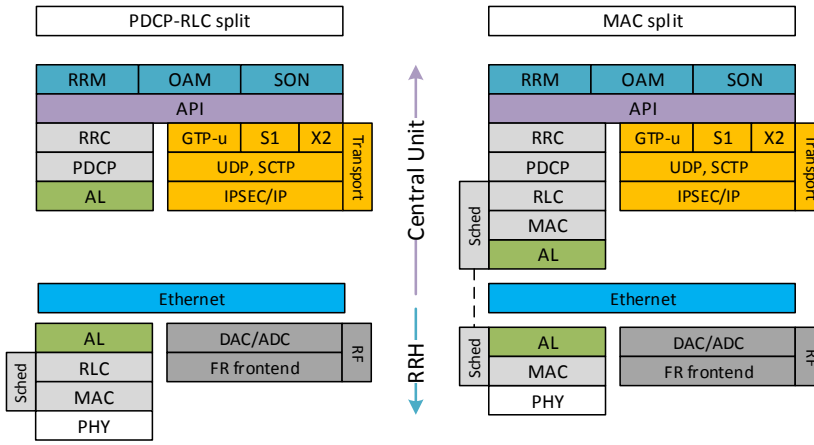

Figure 1.8: Protocol architectures for experiments with base station PDCP-RLC and MAC splits. A split at higher layer typically has lower fronthaul requirements, while split at lower layers provides more coordination opportunities at the CU. AL here stands for an adaptation layer necessary to transport protocol data units over transport networks.

LTE networks and not for LTE-A and 5G networks as the fronthaul requirements are typically higher.

BT and Cavium have conducted trials on split cell options and on the suitability of various transport technologies for fronthaul since 2015. In an experiment that is believed to be a world first, BT has demonstrated Caviums MAC split solution over LTE 2x2 MIMO 15MHz carrier running successfully over 300m of copper using G.FAST and 52km of GPON under realistic load conditions with speeds of up to 90Mbps (using standard UE devices) with only 10-15% overhead associated with a CAL layer interface which is independent of the number of antennas, bandwidth and carriers. Further trials are ongoing on a further MAC/PHY split and denser deployments. The basic setup consists of Caviums ThunderX blade, a virtual machine consisting of a complete EPC, eNB stack and application layer, capable of pooling up to 128 BBUs with varies splits running simultaneously. The connectivity between the BBU and RRH uses standard Ethernet protocol which can connect directly to G.FAST (see Figure 1.9) or GPON distribution point. The latency on both technologies was less than 3ms and 1.5ms respectively making them both suitable for MAC and MAC/PHY splits.

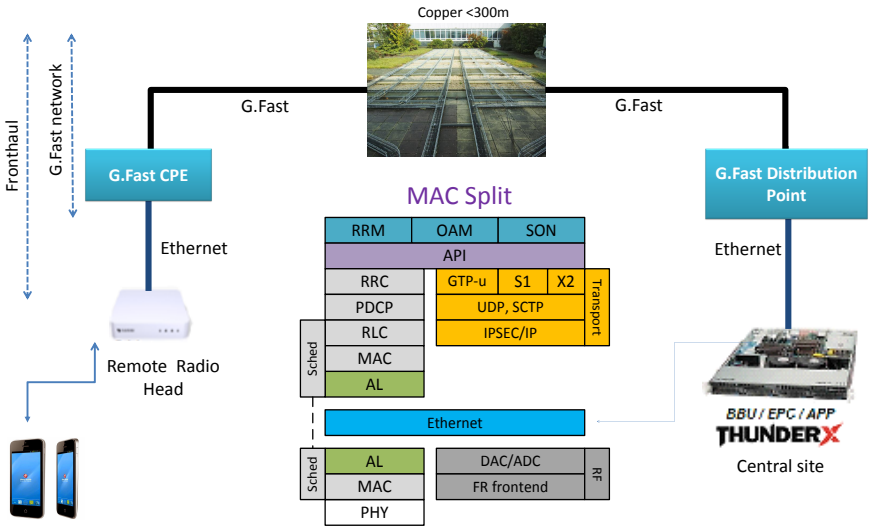

Figure 1.9: Split MAC experiment setup

1.5 Conclusion

This chapter focused on the practical aspects of fronthauling the next generation radio access networks. One of the important characteristics of 5G RAN is likely to be an increased degree of coordination between RAN nodes in networks that are becoming more dense. In this context CRAN becomes an interesting concept with overlaps with other emerging concepts such as NFV and MEC, however realisation of CRAN and associated requirements on fronthaul are dependent on the distribution of the RAN functions between central unit and the remote unit. This chapter provided an overview of the interdependence between RAN functions through the scheduling functions within a base station.

One of the main observations highlighted in this chapter is that many backhaul technologies can support 5G fronthaul requirements, however there is a significant opportunity for optimisation. In particular, while the use of CPRI can be viable for point-to-point links, associated transport overheads may be prohibitive for a dense network, especially where base stations utilise some of the proposed 5G air interface techniques such as higher order MIMO over large channel bandwidths. Provided examples of real-world split RAN operation illustrate the viability of alternative split RAN architectures that can be more efficient from deployment perspective, and invite further research in the area.

Bibliography

[1] D. H. Ring, "Mobile telephony – wide area coverage – case 20564," *Bell Telephony Lab. Tech. Memoranda*, vol. 47-160-37, pp. 1–22, December 1947.

[2] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM Journal of Research and Development*, vol. 54, no. 1, pp. 1–12, January 2010.

[3] C. M. R. Institute, "C-RAN the road towards green ran," White Paper, October 2011.

[4] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*. Elsevier Science, 2013.

[5] S. Ahmadi, *LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies*. Elsevier Science, 2013.

[6] "Small cell virtualization functional splits and use cases," 2016.

[7] "Fronthaul Requirements for Cloud RAN," White Paper, NGMN, March 2015.

[8] M. E. Forum, "MEF 22.1.1 Implementation Agreement Mobile Backhaul Phase 2 Amendment 1 Small Cells," MEF, TR, Jul. 2014. [Online]. Available: https://www.mef.net/Assets/Technical_Specifications/PDF/MEF_22.1.1.pdf

[9] 3GPP, "Architecture enhancements for dedicated core networks; Stage 2 (Release 13)," 3rd Generation Partnership Project (3GPP), TR 23.707, Dec. 2014. [Online]. Available: http://www.3gpp.org/dynareport/23707.htm

[10] ETSI, "Network Functions Virtualisation (NFV); Architectural Framework," ETSI, TR GS NFV 002, Oct. 2013.

[11] "NGMN 5G White Papter," White Paper, NGMN, February 2015.

[12] N. Nikaein, E. Schiller, R. Favraud, K. Katsalis, D. Stavropoulos, I. Alyafawi, Z. Zhao, T. Braun, and T. Korakis, "Network store: Exploring slicing in future 5g networks," in *Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture*. ACM, 2015, pp. 8–13.

[13] M. Iwamura, "Ngmn view on 5g architecture," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[14] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5g mobile and wireless communications: the vision of the metis project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, May 2014.

[15] G. Berardinelli, K. Pedersen, F. Frederiksen, and P. Mogensen, "On the design of a radio numerology for 5g wide area," *ICWMC 2015*, p. 24, 2015.

[16] E. Lhetkangas, K. Pajukoski, J. Vihril, and E. Tiirola, "On the flexible 5g dense deployment air interface for mobile broadband," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*, Nov 2014, pp. 57–61.

[17] J. Vihriala, N. Ermolova, E. Lahetkangas, O. Tirkkonen, and K. Pajukoski, "On the waveforms for 5g mobile broadband communications," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[18] I. D. Silva, G. Mildh, J. Rune, P. Wallentin, J. Vikberg, P. Schliwa-Bertling, and R. Fan, "Tight integration of new 5g air interface and lte to fulfill 5g requirements," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[19] 3GPP, ""Study on small cell enhancements for E-UTRA and E-UTRAN-Higher Layer Aspects(Release 13)," 3rd Generation Partnership Project (3GPP), TR 36.842, Dec. 2014. [Online]. Available: http://www.3gpp.org/dynareport/36842.htm

[20] "LTE Aggregation & Unlicensed Spectrum," White Paper, 4G Americas, Nov 2015.

[21] 3GPP, "Study on Licensed-Assisted Access to Unlicensed Spectrum," 3rd Generation Partnership Project (3GPP), TR 36.889, Mar. 2015. [Online]. Available: http://www.3gpp.org/dynareport/36889.htm

[22] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

[23] N. J. Gomes, P. Chanclou, P. Turnbull, A. Magee, and V. Jungnickel, "Fronthaul evolution: from cpri to ethernet," *Optical Fiber Technology*, vol. 26, pp. 50–58, 2015.

[24] S. Park, C. B. Chae, and S. Bahk, "Before/after precoding massive mimo systems for cloud radio access networks," *Journal of Communications and Networks*, vol. 15, no. 4, pp. 398–406, Aug 2013.

[25] J. Liu, S. Xu, S. Zhou, and Z. Niu, "Redesigning fronthaul for next-generation networks: beyond baseband samples and point-to-point links," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 90–97, October 2015.

[26] "Further Studies on Critical Cloud RAN Technologies," White Paper, NGMN, March 2015.

[27] U. Dtsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for lte," *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 105–128, June 2013.

[28] D. Boviz, A. Gopalasingham, C. S. Chen, and L. Roullet, "Physical layer split for user selective uplink joint reception in sdn enabled cloud-ran," in *2016 Australian Communications Theory Workshop (AusCTW)*, Jan 2016, pp. 83–88.

[29] Huawei, "Handling of UL Traffic of a DL Split Bearer," 3GPP TSG-RAN, Tech. Rep. R2-140054, February 2014.

[30] ZTE, "Comparison of CP Solution C1 and C2," 3GPP TSG-RAN, Tech. Rep. R2-132383, 2013.

[31] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *Journal of Communications and Networks*, vol. 18, no. 2, pp. 135–149, April 2016.

[32] ETSI, "CPRI Specification v6.0," , TR , Aug. 2013.

[33] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela, "Compressed transport of baseband signals in radio access networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3216–3225, September 2012.

[34] B. Guo, W. Cao, A. Tao, and D. Samardzija, "Lte/lte-a signal compression on the cpri interface," *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 117–133, Sept 2013.

[35] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint Orecoding and Multivariate Backhaul Compression for the Downlink of Cloud Radio Access Networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5646–5658, 2013.

[36] S. H. Park, O. Simeone, O. Sahin, and S. S. Shitz, "Fronthaul Compression for Cloud Radio Access Networks: Signal Processing Advances Inspired by Network Information Theory," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 69–79, Nov 2014.

[37] "RAN Evolution Project Bachkaul and Fronthaul Evolution," White Paper, NGMN, March 2015.

[38] Huawei, "A Performance Study of CPRI over the Ethernet," January 2015. [Online]. Available: http://www.ieee1904.org/3/meeting_archive/2015/02/tf3_1502_ashwood_1a.pdf

[39] J. Korhen, "Radio over Ethernet Considerations," February 2015. [Online]. Available: http://www.ieee1904.org/3/meeting_archive/2015/02/tf3_1502_korhonen_1.pdf

[40] IEEE, "Time Sensitive Networking Task Group." [Online]. Available: http://www.ieee802.org/1/pages/tsn.html

[41] Harrison., J. Son, and M. Michelle, "5G Network as envisioned by KT Analysis of KTs 5G Network Architecture," November 2015.