Independent market research and competitive analysis of next-generation business and technology solutions for service providers and vendors

HEAVY READING
WHITE PAPER

Cloud RAN & the Next-Generation Mobile Network Architecture

A Heavy Reading white paper produced for Huawei Technologies Co. Ltd.

HUAWEI

AUTHOR: GABRIEL BROWN, PRINCIPAL ANALYST, HEAVY READING

CLOUD RAN ADVANTAGE

The radio access network (RAN) is the critical asset that underpins the world's largest technology platform: mobile communications. It is fundamental to wide-area connectivity and supports an enormous, and expanding, range of services. The industry has made tremendous progress in mobile network performance over the last 20 years in terms of coverage, price per bit and user experience. To extend these gains, operators now want to apply advances in the cloud and software domains to the RAN.

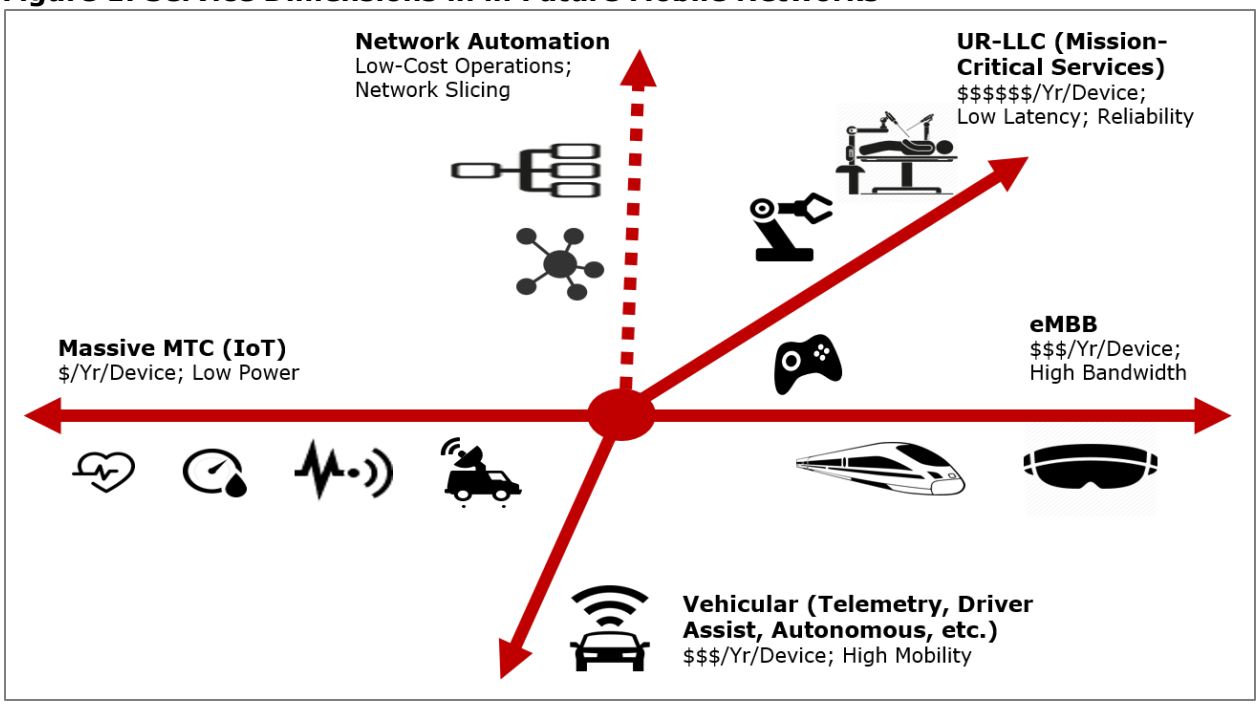
Cloud RAN is emerging as critical to the new wireless network architecture. This white paper investigates state-of-the-art commercial cloud RAN technology and looks ahead to new architectures, new deployment models and new services enabled by the next-generation RAN.

Services Opportunities Driving Cloud RAN

The reason to invest in the network, and consider disruptive cloud RAN architectures, is to enable new services. With LTE-Advanced Pro being deployed in 2017, and with 5G on the horizon for commercial operation in 2020, the mobile market is at an infection point. The market is moving from smartphone-dominated demand to an increasingly diverse range of services across many industries, supported on a common network platform.

Figure 1 identifies the major service categories being used to inform the development of 5G specifications, and which are driving investment in next-generation mobile networks.

Figure 1: Service Dimensions in in Future Mobile Networks



Source: Heavy Reading (after 3GPP SMARTER TR 22.891)

The major service categories – massive machine-type communications (mMTC), enhanced mobile broadband (eMBB) and ultra-reliable, low-latency communications (UR-LCC) services –

have very different performance requirements and traffic profiles. To serve these new markets and increase revenues substantially, operators need highly scalable and flexible networks. Support for these new use cases is inherent to cloud RAN development.

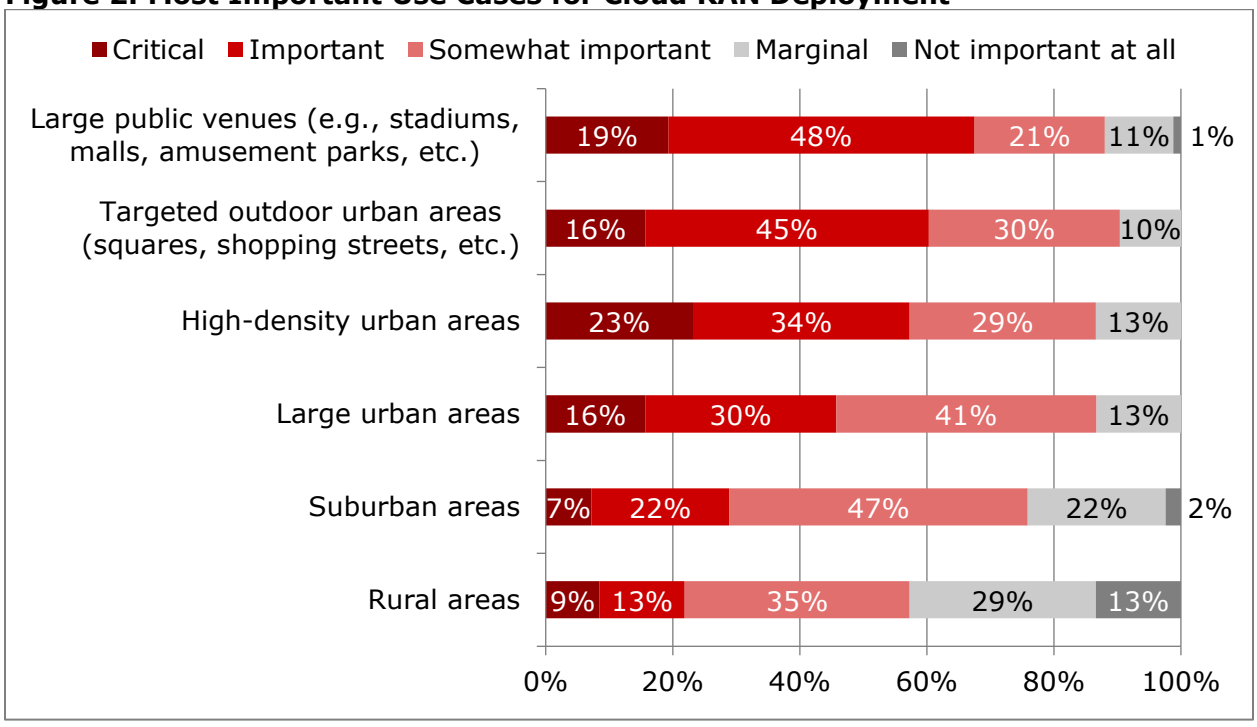
Cloud RAN Deployment

A large number of operators are now evaluating next-generation cloud RAN as a way to meet future service requirements, according to Heavy Reading's 2016 "RAN Strategies Operator Survey." Their reasons are, first and foremost, business-oriented.

Operator respondents ranked system efficiency as most important driver for cloud RAN, with 47 percent ranking the ability to improve scalability and resource utilization as "very important"; this is followed by improving the economics of RAN operation, at 43 percent.

The most popular use cases driving cloud RAN deployment are shown in **Figure 2**: large public venues (a combined 67 percent of operator respondents ranked this "critical" or "important" in our survey); targeted outdoor urban areas, such as public squares and shopping streets (61 percent); and high-density urban areas (57 percent). Suburban and rural coverage scored poorly, indicating that cloud RAN will be used first in hotspot areas with high-density demand.

Figure 2: Most Important Use Cases for Cloud RAN Deployment



Source: Heavy Reading's RAN Strategies Operator Survey, 2016 (n=82)

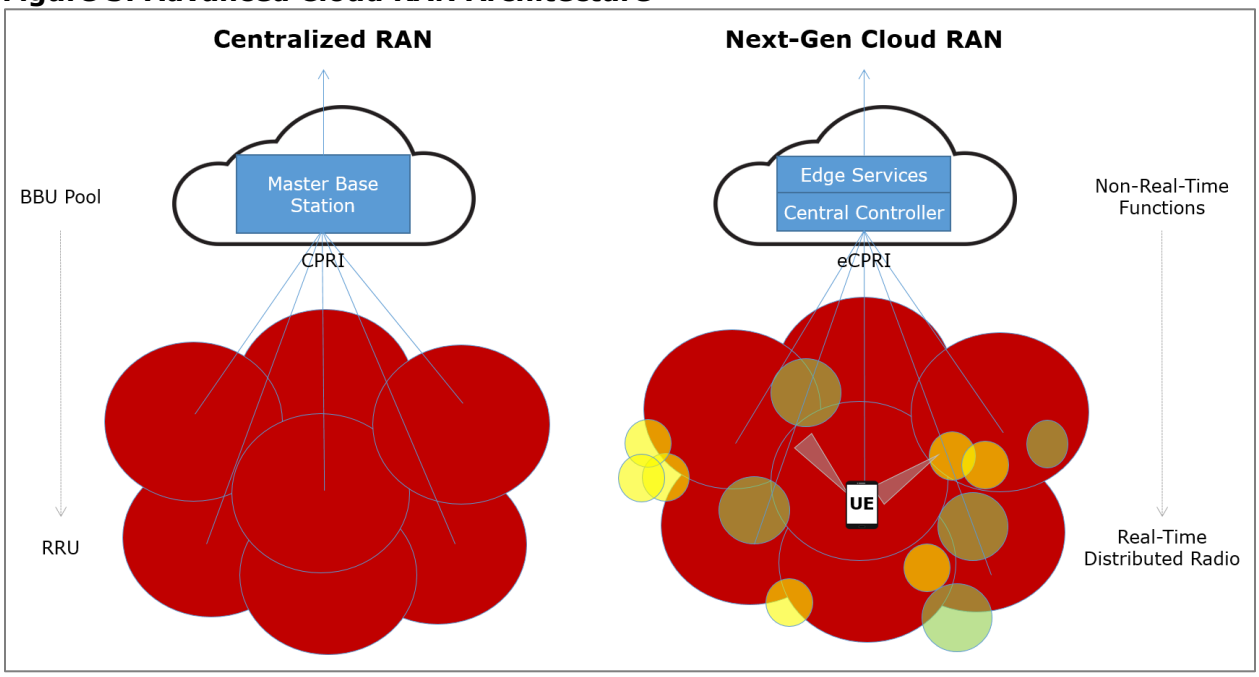
Applying Cloud Principles to RAN

The cloud RAN architecture has two main components: the distributed radio unit, deployed with the antenna, and the baseband unit, deployed centrally at a network-edge data center. The centralized unit aggregates multiple cell sites (typically 15+ cell site equivalents today;

increasing to hundreds in future). Although the implementation is RAN-specific, this architecture maps well to the cloud paradigm of centralized control and distributed processing.

This concept of centralized baseband and distributed radio units is well-known. Two typical network implementations of cloud RAN are shown in **Figure 3**: the classic centralized RAN architecture on the left; and the more advanced, next-generation cloud RAN that can support multi-layer, ultra-dense operation in many different deployment scenarios on the right.

Figure 3: Advanced Cloud RAN Architecture



Source: Heavy Reading

The next-generation architecture extends the cloud RAN concept to ultra-dense networks that incorporates macro, micro and small cells in diverse spectrum bands. By scheduling transmissions centrally, cloud RAN can improve cell edge performance by reducing inter-site interference and can serve users across different radio access bearers. With devices able to take advantage of inter-site connectivity and de-coupled uplink and downlink, cloud RAN introduces the concept of the "no-edge" network to ensure consistently good user experience in high-density networks.

The central controller can be designed using "cloud-native" software to run on general purpose telco cloud infrastructure. This location corresponds to an edge data center facility capable of running multi-access edge computing (MEC) services. One opportunity is to consider how radio access and cloud-based MEC services can be integrated – for example, to enable developers and content providers to optimize their service according to the desired, or available, radio bandwidth. This is expected to be useful for ultra-low-latency and mission-critical services in 5G.

Cloud RAN & 5G

Operator networks incorporate a mixture of radio access technologies. With 2G and 3G being phased out, cloud RAN will focus on 4G and 5G access. 5G is expected to deploy commercially

from 2019 onward. The technical report "Study on New Radio Access Technology; Radio Access Architecture and Interfaces" (TR 38.801) sets out various options for the RAN and its interfaces to the core network.

In the next-generation cloud RAN architecture, real-time (RT) functions are deployed at the antenna site to manage air interface resources, while non-real-time (NRT) control functions are hosted centrally to coordinate transmissions across the coverage area. In 5G, this is being formalized with the central unit (CU) and distributed unit (DU) functional split. This functional architecture is now "native" to the 3GPP specification.

The appropriate CU/DU split between RT and NRT functions is important in 5G for several reasons. On the DU side, for example, the use of massive multiple input/multiple output (MIMO) and associated beam-tracking and beam-switching techniques require highly accurate time estimation to support mobility, which points toward distributed real-time radio functions as optimal.

Elsewhere, centralized functionality makes more sense. The 5G RAN will consist of overlapping cells with multiple connectivity to the device and the network – for example, C-Band small cells and sub 6GHz macro cells will combine to provide the end-user service. This will need some kind of access network selection function that hosted in the CU by the central controller – i.e., in the NRT part of a cloud RAN architecture – that has view of the overall network state.

The following sections of this paper have more detail on how to introduce 5G NR into existing networks, with cloud RAN technology using dual connectivity and fast user-plane switching to manage multiple radio access bearers per device.

Automated, Intelligent OA&M

A critical capability of the next-generation cloud RAN architecture is the ability to accurately understand network status in real time across the coverage area. Coordinated scheduling can radically improve deployment flexibility and performance and enables operators to:

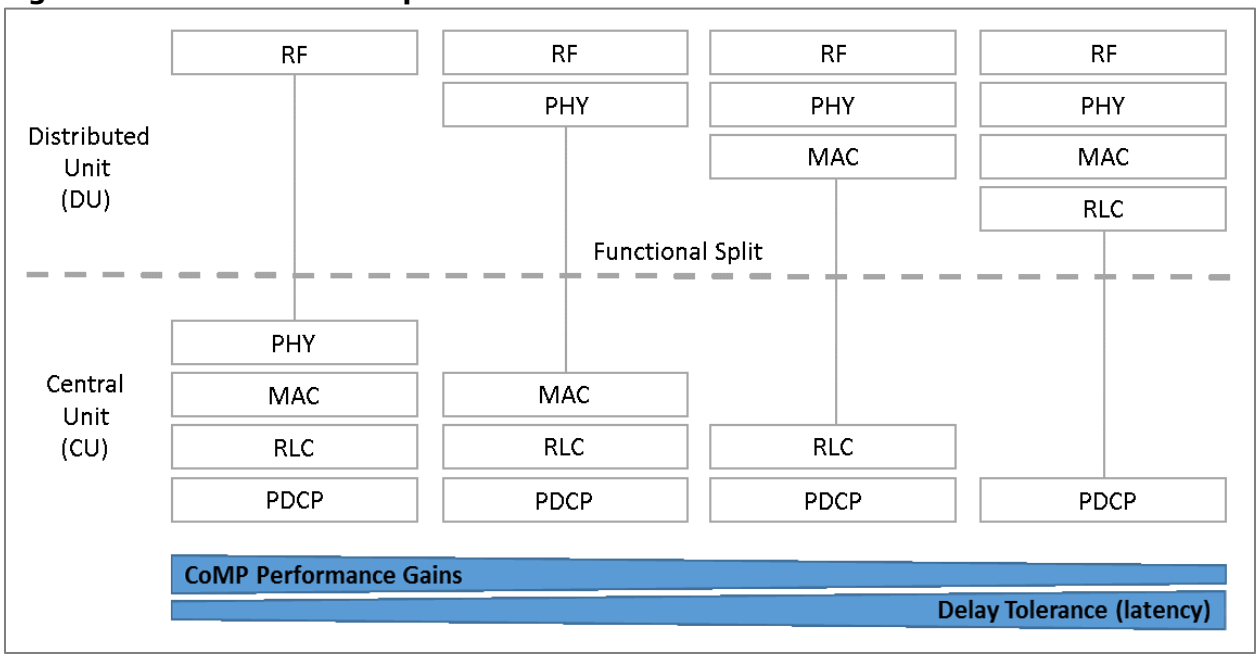
- **Automate continuous radio optimization:** Whereas self-organization network (SON) systems today are generally reactive and work on 15-minute feedback loops, a centrally-scheduled RAN will be able to make rapid and fine-grained changes.
- **Accelerate network expansion and deployment:** Automatic parameter configuration can accelerate deployment of additional capacity at hotpots – e.g., at venues for special events, or in downtown areas – without a complex cell planning exercise.
- **Adapt to changing user behavior:** The control function should analyze network and service performance to identify user behavior changes – for example, a surge of traffic for a news event or if a mission critical application needs to be prioritized – and then optimize then network accordingly.

Operations, administration and management (OA&M) automation is vital because of the significant operational cost dedicated to ongoing RAN optimization. To rapidly expand capacity – using a small cell underlay, for example – typically requires of expensive human skillsets to solve laborious problems. Automating these procedures can impact the cost of production and operator profitability.

ARCHITECTURE PRINCIPALS: RT & NRT SPLIT

Implementation of the cloud RAN architecture – and the subsequent deployment in the network – depends on the functional split between distributed radio and centralized control. There are several options for this in 4G and 5G; the main ones are shown in **Figure 4**.

Figure 4: RAN Functional Split



Source: Heavy Reading

Excluding Common Public Radio Interface (CPRI), a Packet Data Convergence Protocol (PDCP) layer functional split has the most support in the industry today. Several additional splits are proposed in the technical study on RAN architecture for 5G (TR 38.801), with consensus again going toward a PDCP or PDCP/upper radio link control (RLC) split as the optimal trade-off between flexibility and performance.

To the left of **Figure 4**, only the radio frequency (RF) module – which includes filters, power amplifiers, digital-to-analog converters (DACs), etc., is distributed, while all the digital processing is centralized. This model enables coordination at Layer 1. Moving to the right, progressively more functions are distributed, to the point where only the PDCP layer is centralized, which allows for Layer 3 coordination.

In principle, the split to the left of the chart, where the entire protocol stack is centrally controlled, offers the greatest performance gains from coordination across the coverage area. This requires very low-latency, high-bandwidth connection between the distributed unit and centralized unit. Today, the CPRI protocol is used to transport digital RF over fiber optic, with a dedicated channel for each radio.

Moving progressively to the right, the lower-layer functions migrate to the distributed unit. Because an increasing number of scheduling decisions are taken without central coordination, the performance gain declines. The advantage is that the transport link is more "forgiving"

in terms of performance, and multiple radios can be multiplexed onto a carrier Ethernet or IP service.

Essentially, there is a trade-off between system performance and the performance to the front-haul link between CU and DU. This also impacts hardware choices, and cost, at the CU and DU. The optimal balance is difficult to determine because it depends on technical factors and on commercial factors that are generally market- and operator-specific. If an operator owns lots of dark fiber and can easily connect them to radio units, a full centralized model by be appropriate. If the operator uses a lot of microwave to the cell site, a less centralized cloud RAN may be better. In practice, multiple models are likely to prevail globally.

Real-Time (RT) & Non-Real-Time (NRT) Components

To implement the cloud RAN architecture, it is useful to think in terms of NRT (higher MAC and Layer 3) and RT functions (PHY and lower MAC). The processes have different hardware requirements, and by determining a split appropriate to their network, operators can optimize hardware costs, deployment and operational simplicity, and the front-haul transport link.

Centralized Unit for NRT Functions

The central unit is an important part of the architecture. It contains a collection of NRT functions, which include inter-cell handover, cell selection and reselection, user-plane encryption and multi-connection convergence. These functions do not need ultra-low latency and can tolerate dozens of milliseconds of delay. They are therefore ideal for centralized deployment on general-purpose processors. Typically, the idea is to run CU functions on edge cloud infrastructure. New cloud RAN systems will incorporate cloud-native components to optimize scaling, resiliency, upgrade cycles, etc. There is an opportunity to deploy this on common edge cloud infrastructure, such as MEC, as discussed below/above, and to connect to third-party apps via application programming interfaces (APIs).

Distributed Unit for RT Radio Functions

Real-time functions involve radio network scheduling, link adaptation, power control, interference coordination, retransmission, modulation and coding, which require a large amount of specialized computing. Today, these functions are more efficiently processed on custom silicon than on general-purpose processors. Because radio units are a relatively high volume and performance-oriented, power efficiency is critical, and this means specialized hardware will typically be deployed at the distributed site.

There is potential to introduce some very interesting products – for example, using active antenna systems with integrated radio unit – with this functional split. As has been shown in the handset market, tight integration of the radio components can reduce cost and power consumption very significantly. Note also that these radios can still be software configurable even though they do not use general purpose processors.

Edge Cloud Deployment & Integration

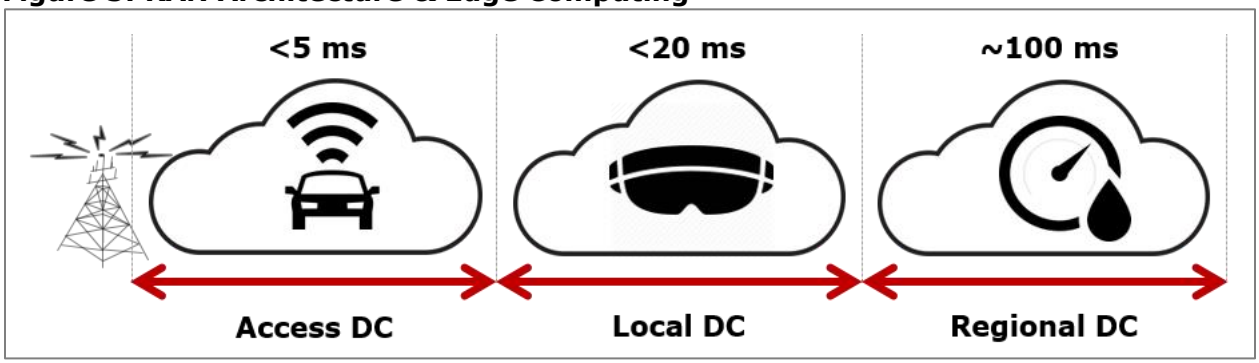
Cloud RAN offers an opportunity to integrate radio access with the rest of the telco cloud-enabled network – for example, using MEC. In this model, the application logic or content is hosted on cloud infrastructure at the same edge data center facility as the centralized control functions. This particularly useful where an application is delay-sensitive. Where

the service is not time-critical, it may be more efficient to process data centrally. This is shown in **Figure 5**.

To the left of the chart, the different levels of distributed data center mapped to transport latency. The closer to the edge, the lower the latency. The access data center is also the facility where the central unit is deployed. With under 1 ms of latency between the RAN and application, this might be suitable for very low-latency services, such as assisted driving, or machine automation in industrial facilities.

For services that are delay-sensitive but don't quite require real-time processing, such as virtual reality, augmented reality, fast image processing (e.g., for facial recognition) or venue services (e.g., for video replay at a sports stadium), the local data center may be more appropriate from the cost performance perspective. For other applications, a large centralized data center may be more appropriate.

Figure 5: RAN Architecture & Edge Computing



Source: Heavy Reading

MULTI-RAT ARCHITECTURE FOR 4G/5G CLOUD RAN

The Technical Report "Study on New Radio Access Technology; Radio Access Architecture and Interfaces" (TR 38.801) sets out various options for the RAN and its interfaces to the core network. This includes extensive analysis of the RT/NRT functional split options as discussed above – currently, we expect at least two functional splits across the DU and CU sites: one higher layer split and one lower layer split.

This decision is expected to be settled shortly in order to enable the 3GPP to achieve its target of a first 5G standards freeze for non-standalone (NSA) mode by the end of 2017, to enable commercial services from 2019. 5G in standalone (SA) mode, with a new core network and without dependencies on LTE, is scheduled to freeze, just six months later in mid-2018, to support commercial service launch from 2020 onward.

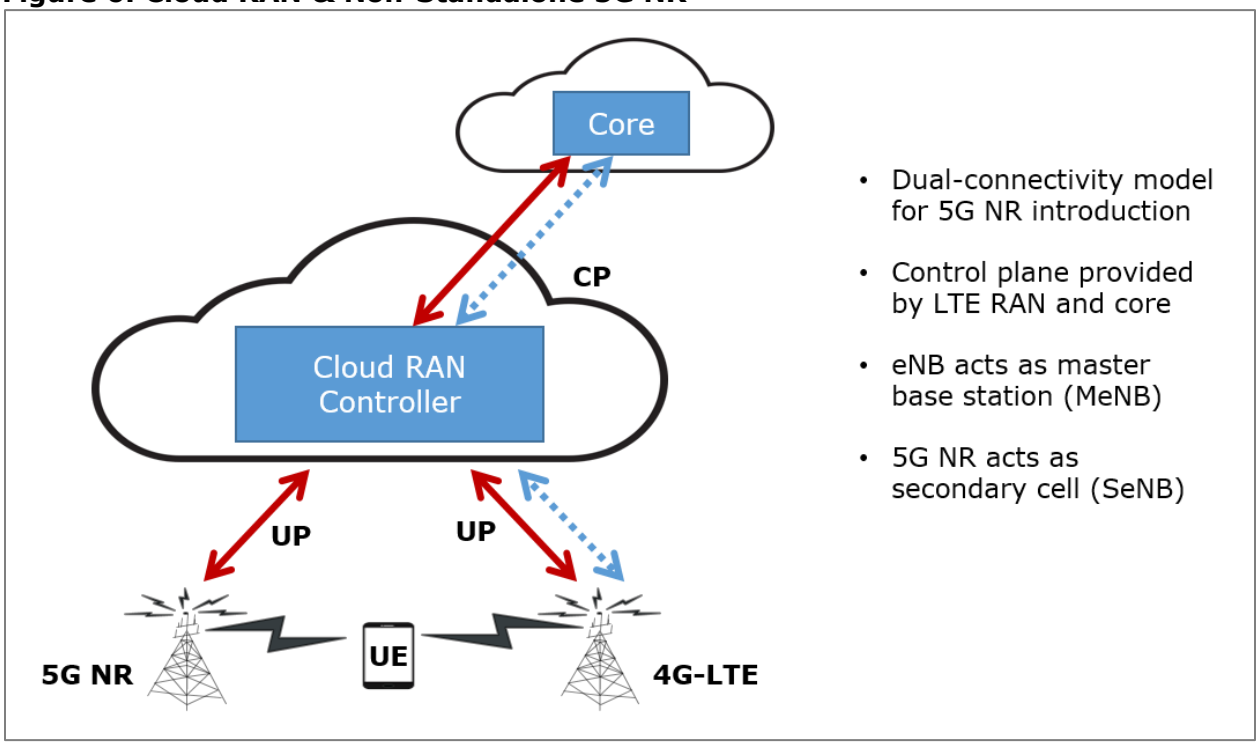
5G NR With Dual-Connectivity to LTE

Non-standalone mode is a way to introduce 5G New Radio into an LTE network. The LTE RAN is used as a master base station (MeNB) and provides control signaling to establish and manage the connection with the device. The 5G radio acts as a secondary base station (SeNB) to provide user-plane throughput. This is likely to be the way 5G is introduced into

commercial networks by some operators. Deploying 5G using non-standalone mode in a cloud RAN architecture is shown in **Figure 6**.

In this model, the UE (the customer device) is served by both the 5G and LTE radios. Rather than homing the 5G base station to the LTE base station, independent bearers, controlled by the centralized unit, are used in cloud RAN. This is more efficient from a transport perspective and avoids the LTE base station becoming a bottleneck. Over time, this cloud RAN architecture with centralized control enables superior performance because it can support fast user-plane switching between radio access technologies (RATs) at the PDCP level. In a multi-band, multi-RAT environment, this enables the cloud RAN to "orchestrate" resource allocation across the coverage area.

Figure 6: Cloud RAN & Non-Standalone 5G NR



Source: Heavy Reading