GTI 5G Network Architecture White Paper

GTI

http://www.gtigroup.org

5G Network Architecture White Paper

V 1.0

GTI

Version	V1.0
Deliverable Type	□Procedural Document √Working Document
Confidential Level	□ Open to GTI Operator Members □Open to GTI Partners √Open to Public
Program Name	5G eMBB
Working Group	
Project Name	Architecture
Source members	Ericsson, ZTE, Nokia
Support members	
Editor	
Last Edit Date	12-02-2018
Approval Date	DD-MM-2018

Confidentiality: The GTI documents may contain information that is confidential and access to the documents is restricted to the persons listed in the Confidential Level. This document may not be used, disclosed or reproduced, in whole or in part, without the prior written authorisation of GTI, and those so authorised may only use this document for the purpose consistent with the authorisation. GTI disclaims any liability for the accuracy or completeness or timeliness of the information contained in this document. The information contained in this document may be subject to change without prior notice.

Document History

Date	Meeting #	Revision Contents	Old	New
2018/02/12		The 1st version		V 1.0

Executive Summary

Unlike the exiting architecture that was designed for the delivery of personal communication services and content, 5G need support a wide variety of new applications and services, so the architecture evaluation is necessary, the network needs to be more flexible and scalable to cope with a broader range of business domains with a variety of different characteristics. The whitepaper describes the challenges to meet these service demands, and introduce a series key technologies to solve these problems. Finally, here are some feasible solutions to enable typical use cases.

Table of Contents

References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

[1] 3GPP, TS 23.501, Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage-2, v15.0.0

[2] 3GPP, TS 23.502, Technical Specification Group Services and System Aspects; Procedures for the 5G System; Stage-2, v15.0.0

[3] 3GPP, TS 37.340, Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Stage-2, v15.0.0

[4] 3GPP, TS 38.401, Technical Specification Group Radio Access Network; NG-RAN; Architecture description; v15.0.0

[5] 3GPP, TR 38.806, Study of separation of NR Control Plane (CP) and User Plane; (UP) for split option 2; v15.0.0

1. Network evolution and 5G architecture

1.1. Network evolution from 4G to 5G

Traditional networks have provided mobile broadband connectivity for smartphones, tablets and laptops. The exiting architecture was designed for the delivery of personal communication services and content, such as voice, video, and web browsing. However, to open the door for supporting a wide variety of new applications and services, the architecture evaluation is necessary to support a broader range of new services and applications with a variety of different characteristics. Beyond connectivity, 5G will offer operators unique opportunities to create new business models and enable new use cases for consumers, enterprises, and industry specific services, as well as content and application providers. Therefore, current telco architecture will transform to being cognitive, cloud optimized and seamless in operation.

1.2. Use cases and main requirements for 5G

IMT for 2020 and beyond defines three families of usage scenarios and each scenario requires a completely different network service and poses requirements that are radically different, sometimes even contradictory.

eMBB (Enhanced Mobile BroadBand): focuses on services characterized by high data rates, such as the new enterprise services and applications along with the exploding consumption of multimedia and collaborative working and social communications, such as AR/VR and video in all its various forms and formats.

mMTC (Massive Machine Type Communications): focuses on services that have high requirements for connection density, such as those typical for smart city and smart agriculture use cases. The smart homes, smart cities and smart factories, all containing billions of sensors require access to a flexible and scalable infrastructure.

URLLC (Ultra-Reliable and Low Latency Communications): focuses on latency-sensitive services, such as self-driving, remote surgery, or drone control. A large number of real-time applications will demand end-to-end network latency of single digit milliseconds to avoid perceivable lags in browsing or videos, or to control drones and robots. In addition, critical machine communications require very high reliability and very low latency; for example, in public safety, autonomous vehicles and telemedicine.

1.3. Main challenges in the network and problems to be solved

The demand for mobile broadband will continue to increase. By 2025, there will be 10-100 times more devices connected than humans. The objects will range from cars and factory machines, appliances to watches and apparel, and new uses will continue to arise, thus communications networks need to be able to meet flexibly and cost effectively in order to support operator profitability and the wider ecosystem.

To deliver the network requirements demanded by this broad range of services will require the fundamental change to all aspects of network architecture including both fixed and wireless access, which opens the potential for a wider range of services and applications.

5G radio revolution will be driven by innovations in radio transmission, algorithms and architecture, e.g., the new spectrum options, scalable Massive MIMO to realize massive capacity, improved end-user experience and cell edge throughput. 5G is also aiming an architecture to have common higher layer architecture and protocol stack considering the multi-layer and inter-working among different radio technologies including 5G, LTE, etc. With such architecture, the coordination for interference management and traffic aggregation could be performed to achieve the most performance gains across such complex network. In addition, the optimal RAN architectural options are also required to support for high throughput and robustness for spotty 5G coverage and multi-connectivity use cases, the signaling overhead inside RAN and between RAN and Core will then need to be optimized.

The traditional radio access network (RAN) consists of standalone, all-in-one base stations where the radio functions and baseband processing are co-located and integrated at the cell site. The baseband units (BBUs) at these sites are connected to the mobile core through the mobile backhaul transport network. With 5G, the new radio technologies including Massive MIMO and mm-Wave require more antennas and wider bandwidth. The traditional fronthaul between the BBU and Radio Remote Head (RRH) may be bottleneck for 5G due to the requirement on expensive high throughput low latency fronthaul.

A 5G RAN architecture is to allow deployments using Network Function Virtualization, which implies that some of the higher layers of the protocol may run on a virtualized pool of hardware resources. In order to gain efficiencies from NFV, the enhancement to the network architecture would be desirable to pool hardware resources across multiple NR nodes.

In general, 5G architecture will need to be flexible to benefit from all available spectrum options, utilizing licensed, shared access and unlicensed spectrum. There will be a need to balance the requirement for high data rates or low latency with massive device densities as well as wide geographic coverage. More importantly, as it is difficult to foresee all future uses, applications and business models, the network needs to be flexible and scalable to cope with the unknown.

2. Solutions and proposals for 5G network architecture

2.1. The evolution of network architecture

5G, compared to 4G, needs to be more flexible and scalable to enable a wider range of scenarios and services. In order to meet the challenges caused by 5G requirements a new architecture is expected with technical innovation and evolution to support more wider scenarios, provide higher user data rate, lower latency.

For the core network, it will be reconstructed based a more convenient and flexible framework vertical having the following characteristics:

- Virtualization and NF modularization;
- Unified service based architecture and interface;
- Control plane and user plane separation;
- Mobility management and session management function decoupling;
- New QOS architecture for introducing the new services.
- Network slicing for supporting the new business domains.

The following Figure 1 show the reconstructed non-roaming core network architecture and interfaces:

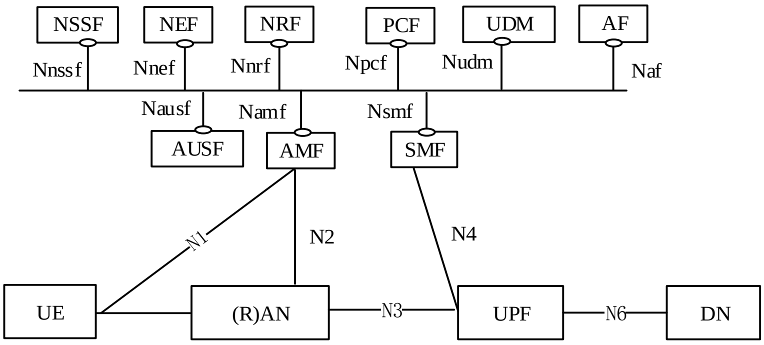

Figure 1 5G Core Service Based Architecture (No-Roaming)

Figure 2 show the core network architecture for home routed roaming in which roaming user must access the services from the UPF of Home Network. There is an N9 interface between UPF of Visited Network and UPF and Home Network for service data forwarding.

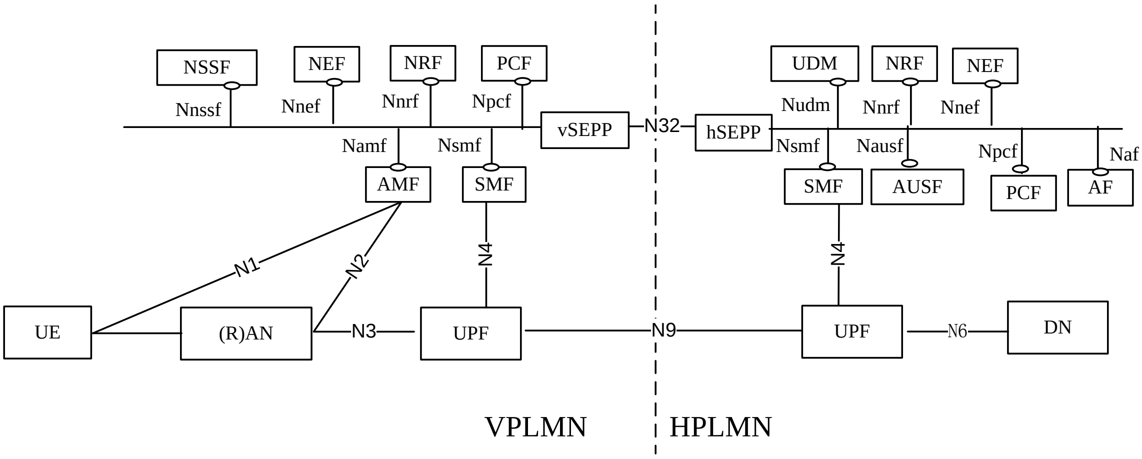

Figure 2 5G Core Service Based Architecture (Home Routed Roaming)

Figure 3 show the core network architecture for local breakout roaming in which roaming user can access the services from the UPF of Visited Network.

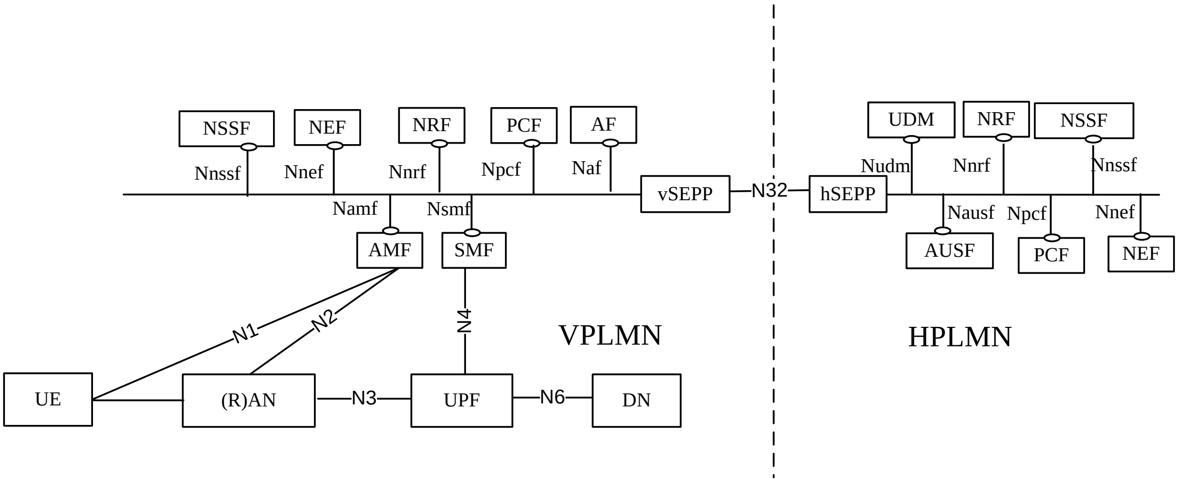

Figure 3 Core Service Based Architecture (Local Breakout Roaming)

As the radio access network, the NG-RAN will provide higher access performance and support new 5G characteristics. It consists of gNBs interconnected with each other by the Xn interface, are also connected by the NG interface to the 5GC. The NG-RAN architecture is illustrated in below:

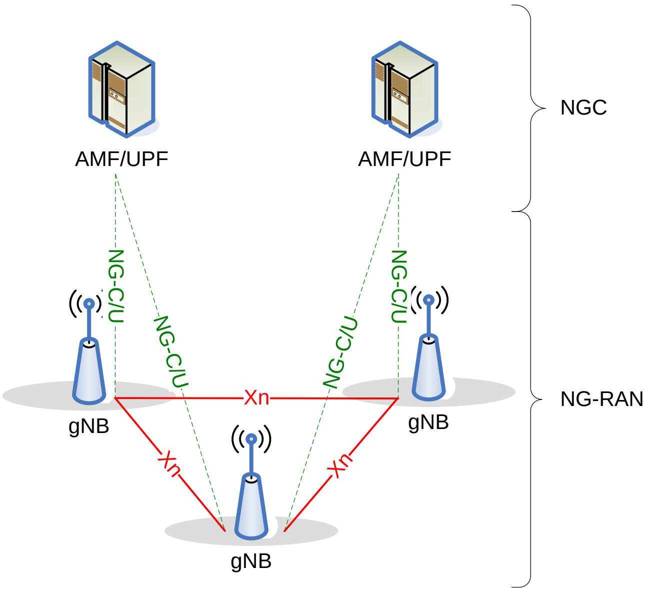

Figure 4 5G RAN Architecture and Interface

The introduction of 5G network cannot be achieved overnight, because the network deployment and evolution scheme will be influenced by many factors. The operators having different market and competitive position will select the different 5G deployment and investment strategy. Whether have the suitable spectrum (Sub6G or mmW), the existing network status and future technology planning are also the important factors.

According to the evolution pace of the core network and the upgrade process of the current LTE network, there may be a variety of network migration paths, such as:

1) Migration Path 1

5GC is launched in day 1, and 5G RAN will be introduced in the first time. 5G and 4G network are independent of each other. The migration mechanism has no impact on the existing LTE network, which is easy to introduce and quickly verify 5G performance, but NR must achieve continuous coverage, otherwise the user will fall back to the 4G impact experience. In the subsequent phase, LTE can continue to upgrade to support 5GC in order to provide new feature align to 5G except to air interface. This path is applicable for operators with advanced technology planning, rich investment and suitable spectrum resources. It also provide a possibility for new operators that has 5G bands but no 2G/3G/4G spectrum and license in the specific markets.

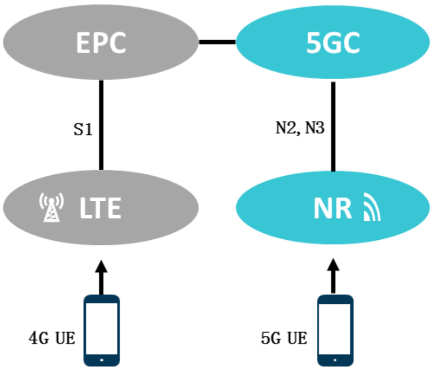

Figure 5 Migration Path for 5G SA Networking

2) Migration Path 2

In early 5G deployment phase, 5GC and 5G NR will be introduced, and LTE will be gradually upgraded to support 5GC. There are not the continuous coverage requirements for 5G NR. It also apply non-standalone networking architecture with LTE using dual connectivity mechanism. The migration can take full advantage of the existing LTE coverage. At the same time, it can speed up the deployment of some new services that can't be supported by the current EPC, such as URLLC etc.

Step 1: Early 5G Deployment utilizing the evolved LTE

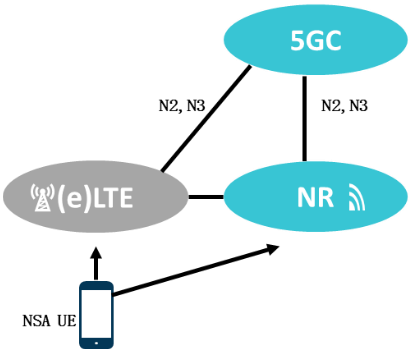

Figure 6 Migration Path to NSA Networking with LTE Upgrading

Step 2: Migrate to standalone NR networking.

3) Migration Path 3

In early 5G deployment phase, only 5G NR will be introduced. It will connect to the existing EPC via S1-U interface, and apply dual connectivity technology to support non-standalone networking architecture. In the subsequent phase, with the 5GC introduction, LTE RAN also be upgraded to support NG interface. At last, it will also achieve 5G standalone networking. The migration path can introduce 5G service as soon as possible with. But there are more steps of evolution in the future.

Step 1: Early 5G Deployment utilizing the exist LTE coverage.

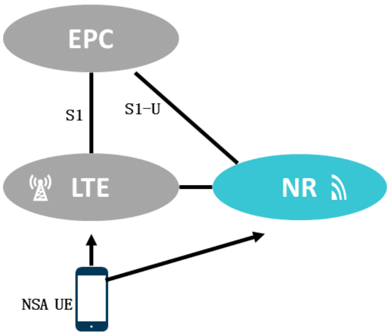

Figure 7 Migration Path to NSA Networking Utilizing the Exist LTE

Step 2: Migrate to support 5GC and EPC simultaneously

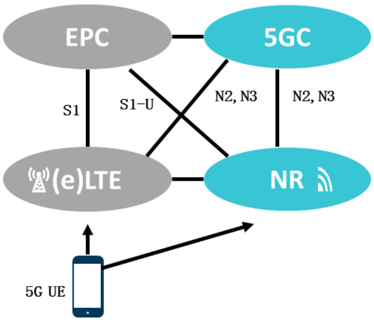

Figure 8 Migration Path to NSA Networking with LTE Upgrading

Step 3: Migrate to standalone NR networking.

In order to support both old and new terminals at the same time, EPC and 5GC will coexist for a long time, or EPC as a slice of 5GC. So when LTE is upgraded to eLTE, it needs to support the ability to route different UEs to different core networks based on UE capabilities.

We have listed two typical migration paths to introduced 5G, but there are many options more than the above. It is not excluded that some operators may take a specially tailored solution based on their specific scenario needs and the deployment of the current network.

2.2. 4/5G interworking

In the early deployment phase, if 5G is applied to standalone networking, it needs to support interworking with 4G that is similar to traditional inter-RAT interoperability including: mobility in CONNECTED STATE (handover, redirection), mobility in IDLE STATE (cell selection or reselection), load balancing, energy saving coordination, etc. In order to interwork with EPC, the UE that supports both 5GC and EPC NAS can operate in single-registration mode or dual-registration mode:

- In single-registration mode, UE has only one active MM state (either RM state in 5GC or EMM state in EPC) and it is either in 5GC NAS mode or in EPC NAS mode (when connected to 5GC or EPC, respectively). UE maintains a single coordinated registration for 5GC and EPC. Accordingly, the UE maps the EPS-GUTI to 5G GUTI during mobility between EPC and 5GC. To enable re-use of a previously established 5G security context when returning to 5GC, the UE also keeps the native 5G-GUTI and the native 5G security context when moving from 5GC to EPC.

- In dual-registration mode, UE can handle independent registrations for 5GC and EPC, and maintains 5G-GUTI and EPS-GUTI independently. In this mode, UE provides native 5G-GUTI, if previously allocated by 5GC, for registrations towards 5GC and it provides native EPS-GUTI, if previously allocated by EPC, for Attach/TAU towards EPC. In dual-registration mode, the UE may be registered to 5GC only, EPC only, or to both 5GC and EPC.

For service continuity for inter-system change, the N26 interface has been introduced in core network architecture. N26 interface is an inter-CN interface between the 4G MME and 5GS AMF in order to enable interworking between EPC and the NG core, and it is used to provide seamless session continuity. The Figure 9 shows the core architecture supporting inter-RAT mobility with N26 interface.

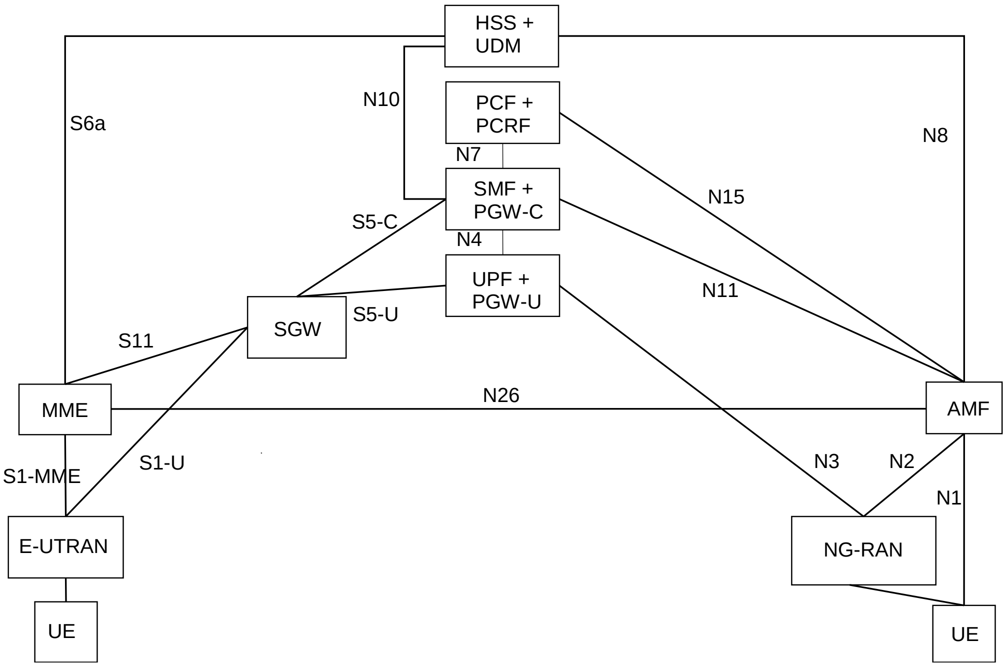

Figure 9 Architecture for interworking between 5GS and EPC/E-UTRAN

It should be noted that support of N26 interface in the network is optional for interworking. For mobility in dual-registration mode, the support of N26 interface between AMF in 5GC and MME in EPC is not required. For mobility in single-registration mode, either using or not using the N26 interface for interworking is viable. For interworking without the N26 interface, IP address continuity is provided to the UEs on inter-system mobility by storing and fetching PGW-C+SMF and corresponding APN/DDN information via the HSS+UDM. In such networks AMF also provide an indication that interworking without N26 is supported to UEs during initial Registration in 5GC or MME may optionally provide an indication that interworking without N26 is supported in the Attach procedure in EPC.

If 5G is applied to non-standalone networking, it should support interworking with LTE based NSA architecture using Multi-RAT dual connectivity (MR-DC) mechanism that a multiple Rx/Tx UE may be configured to utilize radio resources provided by two distinct schedulers in two different nodes connected, one providing E-UTRA access and the other one providing NR access. One scheduler is located in the Master Node (MN) and the other in the Secondary Node (SN). The MN and SN are connected via a network interface and at least the MN is connected to the core network. The two nodes cooperate with each other to support a various services. According to the different types of the core network and the access network, there are three ways for dual-connectivity.

- **EN-DC**

A UE is connected one eNB that act as a MN and one gNB that act as a SN. The eNB is connected to EPC via S1-C and S1-U interface, and the gNB is connected to EPC via S1-U interface. The eNB and the gNB connect each other via X2 interface. The MR-DC mechanism is called E-UTRA-NR dual connectivity (EN-DC).

The EN-DC architecture is suitable for deployment scenarios that only introduce 5G NR and do not introduce 5GC. In this architecture, the anchors of the control plane are always located in the LTE RAN, that is, the S1-C interface is terminated by eNB. It can be further divided into three types networking architecture in the sight of different user plane aggregation point and distribution flow types, as shown below:

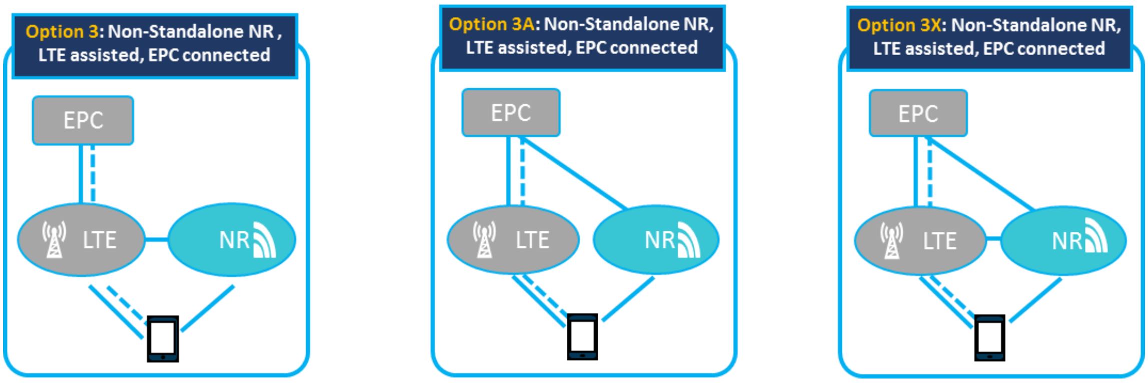
Option 3: Non-Standalone NR , LTE assisted, EPC connected
EPC
LTE
NR
Option 3A: Non-Standalone NR, LTE assisted, EPC connected
EPC
LTE
NR
Option 3X: Non-Standalone NR, LTE assisted, EPC connected
EPC
LTE
NR

Figure 10 EN-DC Scenarios

Option 3 networking: eNB also terminates S1-U interface, and NR gNB haven't any interface to EPC. The traffic flow is converged at eNB PDCP layer and divided from the eNB to the gNB via X2 interface. In this architecture, eNB nodes will carry a large amount of traffic and corresponding computing work. So the eNB hardware have to be upgraded to avoid becoming bottleneck. Correspondingly, the backhaul to the core network and between the two nodes also need to be improved.

Option 3A networking: gNB also has S1-U interface to EPC. Traffic flows are split in the core network, and the different service bearers can be carried in LTE or 5G NR. In this architecture, eNB can migrate the services that needs high throughput or ultra low latency to the NR Node, reducing the requirement of processing capacity for itself. And because there is no split mechanism, the demand for the X2 backhaul is easy to meet.

Option 3X networking: gNB has S1-U interface to EPC. The traffic flow is converged at gNB PDCP layer and divided from the gNB to the eNB via X2 interface. In this architecture, the new 5G NR with high performance carry on most of the traffic and avoids the excessive upgrade and transformation to the existing RAN and transport network. When the NR coverage is poor, the LTE side also can provide robustness using traffic flow split mechanism with smaller bandwidth requirement of X2 backhaul.

The operators can choose the appropriate architecture according to the existing network situation and evaluation of network transformation and upgrading cost.

- **NGEN-DC**

A UE is connected one evolved eNB that acts as a MN and one gNB that act as a SN. The evolved eNB is connected to new 5GC via NG-C and NG-U interface, and the gNB is connected to 5GC via NG-U interface.

The evolved eNB and the gNB connect each other via Xn interface. The MR-DC mechanism is called NG-RAN E-UTRA-NR dual connectivity (NGEN-DC).

The NGEN-DC architecture is suitable for deployment scenarios that LTE RAN have been upgraded to support 5GC. In this architecture, the anchors of the control plane are always located in the eLTE RAN. Like EN-DC, it can also be divided into three architectures as shown below:

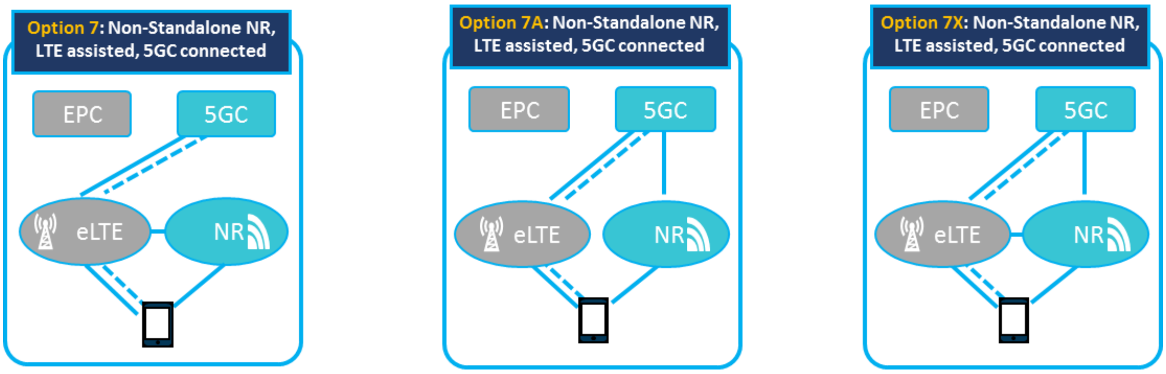
Option 7: Non-Standalone NR, LTE assisted, 5GC connected
EPC
5GC
eLTE
NR
Option 7A: Non-Standalone NR, LTE assisted, 5GC connected
EPC
5GC
eLTE
NR
Option 7X: Non-Standalone NR, LTE assisted, 5GC connected
EPC
5GC
eLTE
NR

Figure 11 NGEN-DC Scenarios

For Option 7 family(7/7a/7x), the bearer split mechanism, the requirement for network elements and transport are similar to Option 3 family (3/3A/3X). LTE has evolved to eLTE to support 5GC in this deployment phase.

- **NE-DC**

A UE is connected one gNB that act as a MN and one evolved eNB that act as a SN. The gNB is connected to new 5GC via NG-C and NG-U interface, and the evolved eNB is connected to 5GC via NG-U interface. The gNB and the evolved eNB connect each other via Xn interface. The MR-DC mechanism is called NR-E-UTRA dual connectivity (NE-DC).

The NGEN-DC architecture is suitable for deployment scenarios that 5G NR can provide the continuous coverage. In this architecture, the anchors of the control plane are always located in the 5G NR. Because 5G NR have stronger capability and greater capacity, so the architecture that traffic are divided from SN (eLTE) to MN (NR) is no longer needed. So Option 4 contains only include two architectures: Option 4 and 4A.

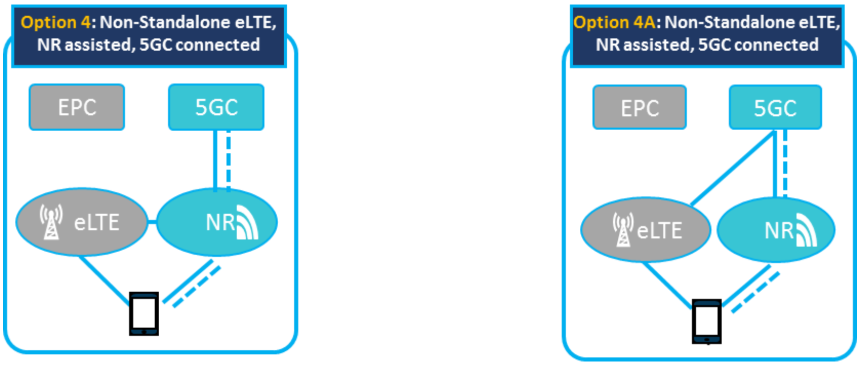
Option 4: Non-Standalone eLTE, NR assisted, 5GC connected
EPC
5GC
eLTE
NR
Option 4A: Non-Standalone eLTE, NR assisted, 5GC connected
EPC
5GC
eLTE
NR

Figure 12 NE-DC Scenarios

In addition, dual-connectivity technology can also be combined with SRB split, new SRB taken by the SN, carrier aggregation and other technologies to achieve more flexibility, reliability, and throughput gains.

As mentioned above, whether using SA networking or NSA networking, for various service scenarios, we can use the dual-connectivity technology between LTE and 5G system to improve throughput, reliability, latency and mobility performance.

2.3. A split architecture

The technology logic of RAN Split and Cloud RAN are not entirely chime. The former refers to RAN logical separation, and the later concern about cloud framework based virtualization technology. As a starting point, CU/DU split could be introduced first, and this would lead to run CU part in a cloud execution environment.

Driven by greater needs for coordination as well as increasing resource efficiency and advances in network virtualization, Cloud RAN architecture allows for the use of NFV techniques and data center processing capabilities such as coordination, centralization and virtualization in mobile networks. This supports resource pooling (more cost-efficient processor sharing), scalability (more flexible hardware capacity expansion), layer interworking (tighter coupling between the application layer and the RAN) and spectral efficiency.

2.3.1. High layer split

The cloud execution environment is a kind of natural way to deploy the CU part and cloud RAN should support the following:

- separation of control and user plane to support flexible scaling of capacity for different functions of the RAN;
- a variety of deployment options for anticipated network scenarios, including a wide range of transport network solutions, base station configurations and user applications;
- alignment with legacy deployments, which reduces the overall network complexity thanks to a unified network architecture.

The Cloud RAN will support different network architecture functional splits, as outlined in Figure 13, including different levels of functionality implemented as NFVs. By utilizing Cloud RAN, operators can centralize the control plane (seen together with PDCP split in Figure 13) – which does not have extreme bitrate requirements – to bring RAN functionality closer to applications, or further distribute the physical layer closer to the antenna (PHY split in Figure 13) to enable massive beamforming.

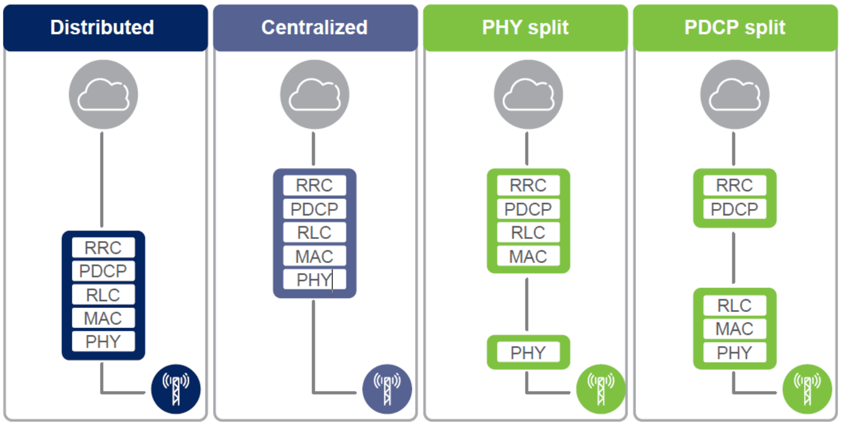

Figure 13 Examples of functional splits of the radio access protocol layer in a Cloud RAN

Cloud core and NFV frameworks also bring applications closer to the RAN, and this proximity enables scalable and shared common and commercial-off-the-shelf (COTS) execution platforms to be used and leveraged for cost-effectiveness and flexibility. For instance, if cloud core functions are pushed out into the network and RAN is centralized to some degree, there will eventually be some degree of colocation of core and RAN functionality – either with RAN and core together on a server in a distributed fashion, or with RAN and core executing in a centralized data center environment. This will enable substantially lower latencies for the interconnection between RAN and core.

From deployment point of view, the CU/DU separation could be realized as PPF (packet processing function), mainly PDCP layer, RCF (radio control function), mainly RRC layer, and RPF (radio processing function), mainly Layer 1. The PPF and RCF can be executed in a cloud environment in the case of CU/DU separation.

2.3.2. CP/UP split

This kind of selective centralization of the control plane – shown in Figure 15– can provide user experience benefits such as mobility robustness, while spectral efficiency can be ensured through a level of radio resource coordination across radio sites.

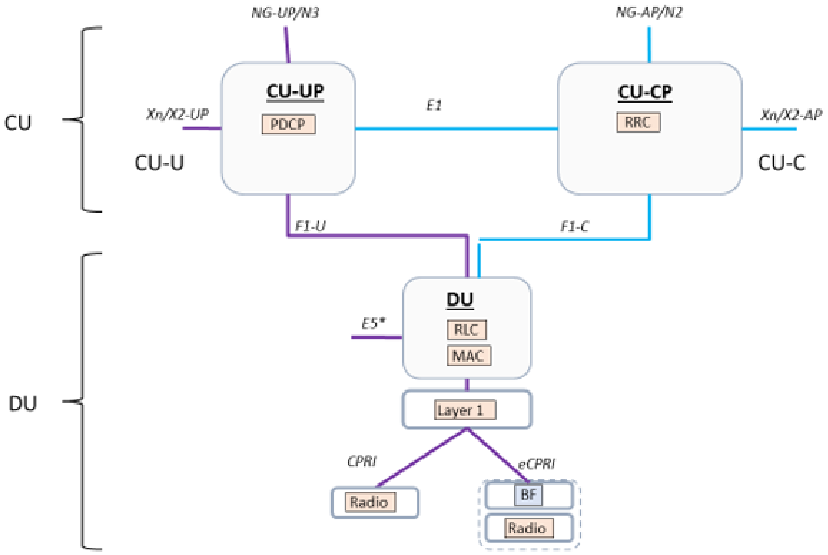

Figure 14 The CU/DU deployment view in a split architecture

From a user plane perspective, Cloud RAN can also provide optimization benefits for certain deployment scenarios driven by dual connectivity needs. With dual connectivity in a fully distributed deployment, data can be routed first to one site and then rerouted to the second site.

This results in what is referred to the trombone effect in the transport network, which means that data is sent inefficiently back and forth over the same transport network segment. This can be avoided by placing the routing protocol higher up in the transport network aggregation hierarchy, which improves user plane latency.

The L2 user plane packet data convergence protocol (PDCP) layer is predominantly a routing protocol, but it also includes a fair amount of processor-heavy ciphering. Optimized ciphering accelerators can be used to provide a low-latency and high-bandwidth performance implementation in a more energy- and cost-efficient way, as a complement to a more generic packet data processing environment.

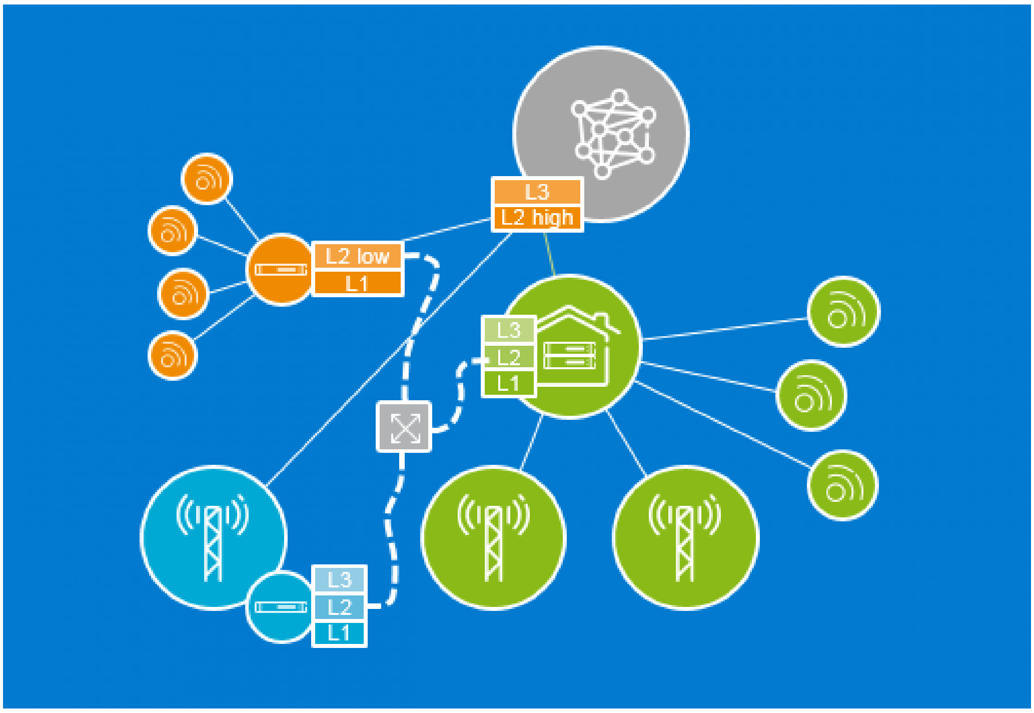

Figure 15 Showing scalability and modularity of functional allocation of the different protocol layers together with virtualization in the cloud.

Figure 16 illustrates a scenario where the CU-CP (denoted as RRM) is co-located with the DU to provide low latency for critical CP procedures, while the CU-UP (denoted as PDCP-U) is deployed in a centralized manner to take advantage of cloud technologies and to provide a central termination point for UP traffic in dual-connectivity (DC) configurations.

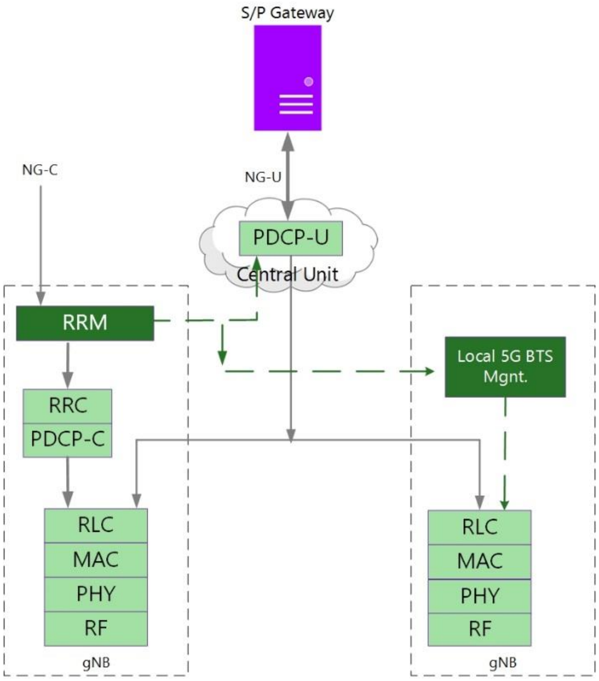

Figure 16 Network scenario with distributed CU-CP (RRM) and centralized CU-UP (PDCP-U)

Figure 17 illustrates a scenario where the CU-CP (denoted as RRM) is centralized to coordinate the operation of several DUs, which can potentially provide efficient load balancing. Also in this case the CU-UP (PDCP-U) is deployed in a centralized manner to take advantage of cloud technologies as provide a central termination point for UP traffic in DC configurations.

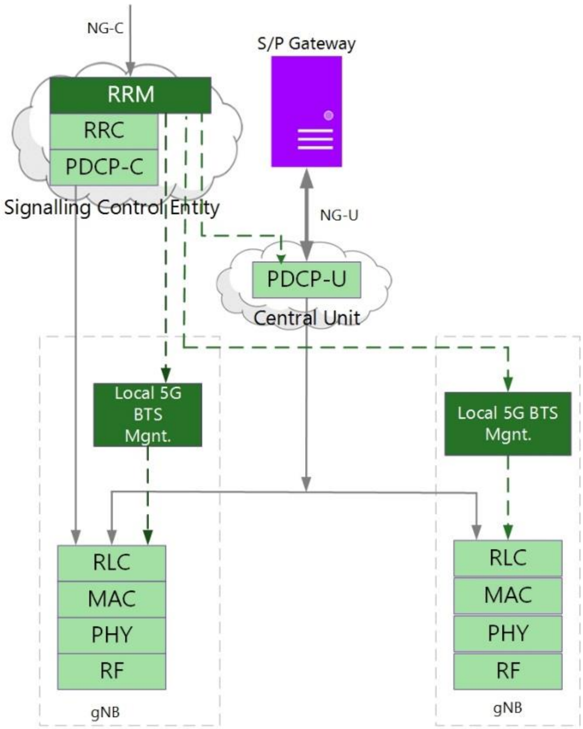

Figure 17 Network scenario with centralized CU-CP (RRM) and centralized CU-UP (PDCP-U).

2.3.3. Low layer split

A new common public radio interface called evolved CPRI (eCPRI) has been defined as new standard specification which will be applied to support 5G. The eCPRI specification will be based on new functional partitioning of the cellular base station functions that enables:

- A many-fold reduction of the required bandwidth
- Bandwidth requirements scale flexibly according to the user plane traffic
- The use of main stream transport technologies like Ethernet

eCPRI nodes will send either user data or user data control in Ethernet frames tagged as different eCPRI Services. Depending on actual functional split between baseband unit (REC) and radio unit (RE) that data sent between eCPRI nodes will be of different kind:

- Transport block data
- Modulated data in frequency plane
- IQ-data in time domain (CPRI split)

The new eCPRI Specification is promoting an intra-PHY functional split between baseband unit and radio unit, i.e. a low-layer split (see Figure 18).

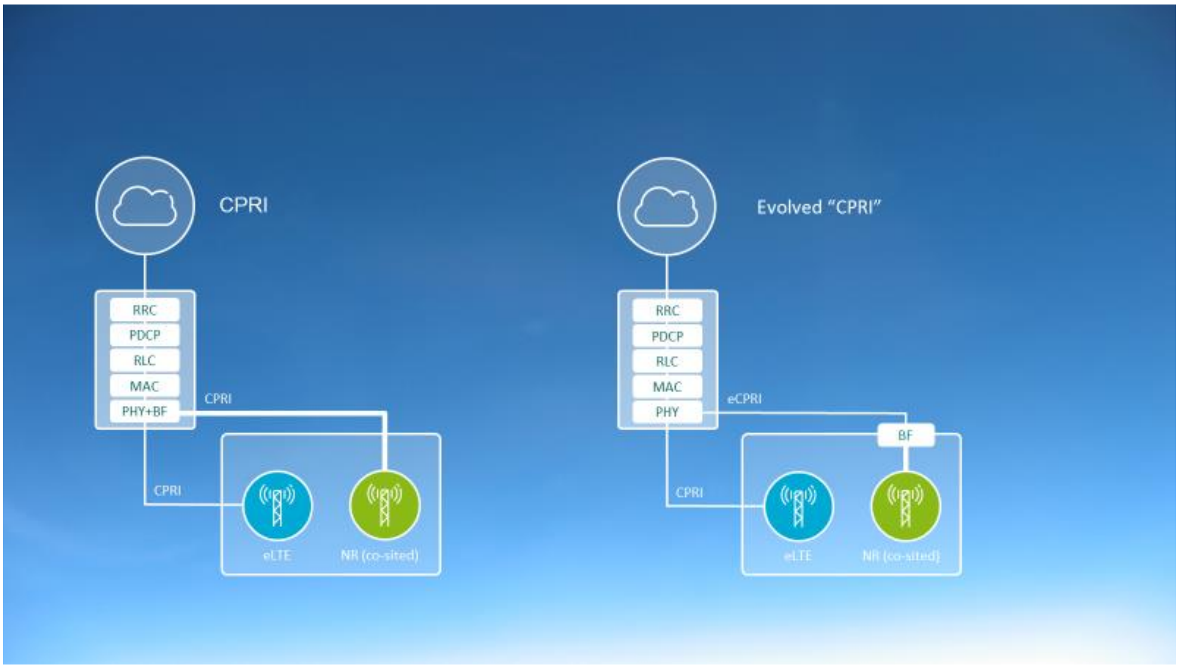

Figure 18 Illustrate digital beamforming is deployed in the radio unit in the case of eCPRI

2.4. The introduction of virtualization and cloud in RAN

The terms virtualization and cloud are often used interchangeably. They do work well together in many cases, including in a RAN context. However, each concept brings different things to the table.

In general, virtualization is a technique that can mean different things in different scenarios, and it is unlikely to mean the same thing in a RAN context as in, for example, a data server context. The reason for this is the substantial difference in real-time requirements imposed by the radio access protocol. Many of the synchronization requirements that ensure the performance of the radio access protocol are on the microsecond level and, in some cases, the nanosecond level. Thus, RAN functionality is not easily hosted by the so-called virtualized platform as a service (PaaS) model, as is possible with straightforward applications and server-type functions.

On the other hand, there is no need to virtualize all RAN functionality to provide the benefits of Cloud RAN. Virtualization as an execution environment technique can be used to provide isolation, scalability and elasticity, among other things, for the Radio Resource Control (RRC) protocol layer. When applied in this manner, virtualization can be used to simplify the management and deployment of the RAN nodes, for example, by allowing the definition of arbitrarily-sized base stations (in terms of the number of cells) and for more flexible scaling of higher layer functionality separate from the scaling of other layers.

Virtualization can also be used to leverage a common execution environment for RAN, core and application functionality, providing the ultimate in execution proximity and ensuring maximum responsiveness of, for example, a certain service, or, as it is sometimes called, a certain type of network slice. The possibility to virtualize network functions in this way makes it feasible to place the functionality on a more generic and generally available execution platform together with cloud core applications and other latency-critical services, sometimes even in a PaaS environment.

Centralizing base station processing with Cloud RAN simplifies network management and enables resource pooling and coordination of radio resources.

Pooling, or statistical multiplexing, allows an execution platform to perform the same tasks with less hardware or capacity. This is of greatest interest for tasks that require a large number of computational resources. It also means that the most desirable pooling configuration is a fully centralized baseband approach with a star connection long-haul CPRI between the pooled baseband and the distributed remote radio heads. This is because processing of the lower layers constitutes such a large part of the computational effort. As mentioned earlier, however, there are not many cost-efficient solutions for long-haul CPRI.

By using separate (data center) processing capacity for higher layers, new features can be introduced without affecting the performance and capacity of distributed baseband units. The introduction of massive MIMO configurations – which will be of increasing interest with the move into ever-higher frequency bands – will also further highlight the need for optimized transport and baseband processing for centralized baseband configurations. This requirement is mitigated if centralization is isolated to the higher layers of the protocol stack.

There are some other potential centralization advantages, including:

- Fewer X2 instances: centralizing the X2 control plane leads to fewer X2 instances and shorter;
- X2 distances, i.e. latencies;
- Fewer handover failures: centralizing mobility and traffic management decisions leads to fewer handover failures and less network control signaling in complicated (heterogeneous) radio network environments;
- Memory trunking gains: pooling of the user equipment contexts from users served by more.

2.5. The support of new QoS and network slicing

QoS (Quality of Service) is a network technology that enable network to solve network delay and block problems, providing better support for the special application services. Compared to 4G network, which is mainly used for mobile equipment, the 5G network address to interconnect all things, so it must support a variety of service types. Therefore, 5G's QOS architecture need to further enhanced comparing to 4G.

In LTE system, the basic granularity of QoS control is EPS bearer. The service IP packet needs to be mapped to different EPS Bearer, and all data flows on the same EPS bearer will get the same QoS guarantee. Further, eNB achieves a one-to-one mapping between the DRB bearer and the S1 bearer to provide the corresponding QOS guarantee.

The 5G new QoS model is based on QoS flows. There is a PDU session that provides a PDU connectivity service between a UE and a Data Network. The QoS Flow is the finest granularity of QoS differentiation in the PDU Session. User Plane traffic with the same QOS flow within a PDU Session receives the same traffic forwarding treatment (e.g. scheduling, admission threshold). Similar to 4G EPS bearer, the 5G QoS model also supports both GBR and non-GBR QoS flows.

For each QoS flow, a QoS profile will be derived by the SMF, and delivered to the RAN and UE via AMF for guarantee service QoS.

For each UE, the NG-RAN establishes one or more Data Radio Bearers (DRB) per PDU Session. The NG-RAN maps packets belonging to different PDU sessions for different DRBs. It is up to NG-RAN how to map multiple QoS flows to a DRB. The mapping can be made based on the service type, the bearer QOS profile, the RAN resource condition and the network slicing related configuration.

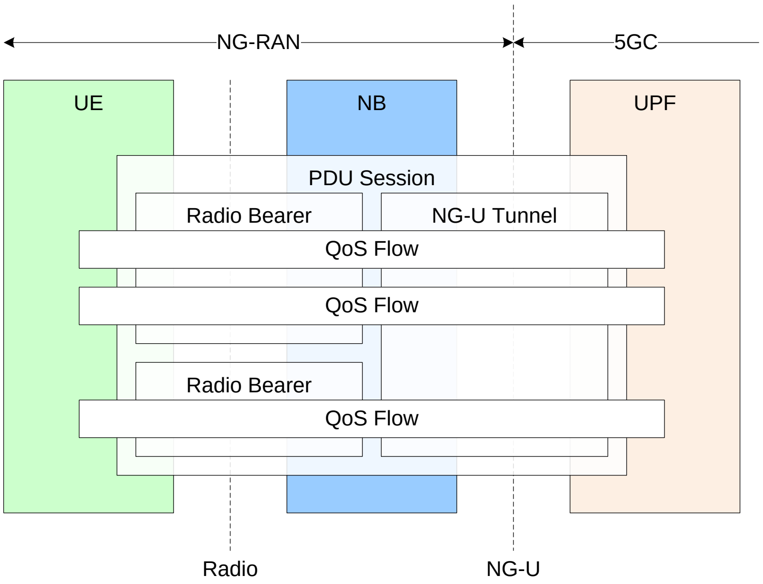

Figure 19 New 5G QOS Architecture

Network slicing is also one of the technical characteristics of 5G, which is used to isolate a logical network in physical network infrastructure according to different service characteristics, and provide better end to end support for special services. It has a similar technical goal with QoS mechanism. In order to ensure the coordination and consistency of the two technology, the core network will generate QoS configuration based on the slice type. And for each PDU session, the Network Slice Selection Assistance Information (S-NSSAI) will be carried in PDU session management messages from AMF to RAN . RAN will use the slice identity to ensure the consistency between QoS and network slice, including the resource allocation and scheduling policy, etc.

2.6. The requirements of 5G network architecture to underlying transport network

A virtualized RC architecture should aim to provide the same characteristics as the embedded legacy LTE product. This is under the condition that the provided transport interconnect is dimensioned to support the traffic model.

It’s highly likely that the separation into the smaller services introduce additional signaling, and an increased cost of signaling. This means that to fulfill the aim to be on par with legacy, the interconnect must be faster and dimensioned higher to support the same characteristics, at least when deployed in a cloud context.

The transport interfaces in a vRAN consists of eCPRI between RU and DU, F1 between DU and CU (see Figure 20).

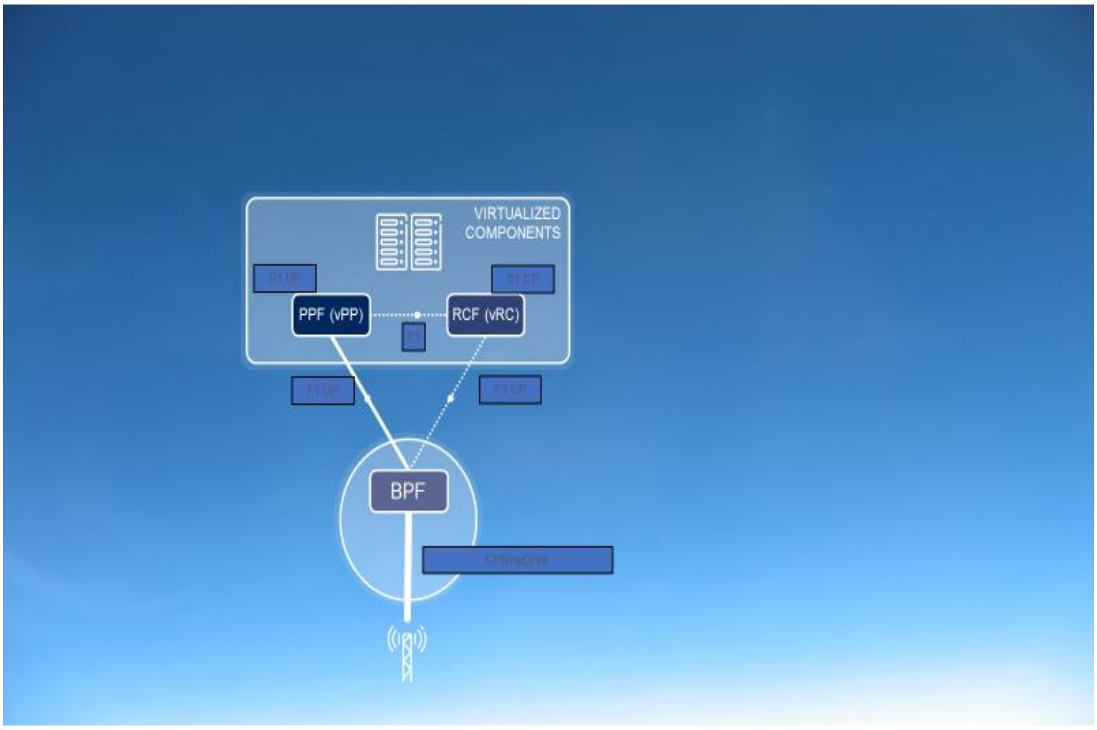

Figure 20 Illustrate the F1 and eCPRI transport interfaces

2.6.1. Transport Interface eCPRI

The transmission links between the central baseband units and distributed radio units use Common Public Radio Interface (CPRI) fronthaul over dedicated fiber or microwave links as one of the transport solutions. This CPRI fronthaul requires tight latency and large bandwidths.

State-of-the-art signal processing technology can enable large centralized baseband configurations that host a number of remote radio units. These remote units are fully coordinated with joint transmission and reception across all antenna elements, cells and bands.

The potential for better performance with a fully centralized baseband deployment is unmatched.

However, in many situations, CPRI connectivity requirements will be insurmountable, especially considering likely future 5G requirements such as extreme data rates and massive beamforming, which will feature many more individually-controlled antenna streams.

With this limitation in mind, most future networks will likely consist of a combination of distributed and centralized baseband deployments, mainly depending on availability of fiber and performance needs.

In order to properly benefit from centralization, operators will need a RAN architecture that both takes advantage of the strengths of a centralized architecture and allows for more affordable transport solutions.

2.6.2. Transport Interface F1

F1 is the interface between the lower (real-time) and the upper (non-real time) part of BBU processing. This interface is based on Ethernet or IP. It has relaxed throughput requirement like backhaul, but it has more stringent latency requirements due to the tight feedback loop between the Virtualized RAN and Radio Processor, where an optimized implementation can benefit very much to e2e performance.

Protocol is based on Ethernet/IP, expected throughput is in the range of 10 Gbps with the limitation of carrier bandwidth Sub6G, one way maximum latency 5ms, while frame length and jitter requirements are in line with backhaul applications.

3. How 5G network architecture enables the use cases

3.1. High bandwidth

Bandwidth = Spectrum * Spectrum effectiveness * Site density

Based on the above formula, high Bandwidth need more spectrum, more spectrum efficiency, and more intensive site deployment.

It is not easy to achieve the 3 targets, especially in the legacy network architecture such as LTE, so the 5G network architecture is required to give solutions for high bandwidth scenarios.

Get more spectrum

For operators, spectrum is always the most valuable, and the NR spectrum that operators can acquire must be finite.

By supporting MR-DC, including EN-DC, NGEN-DC and NE-DC, the 5G network architecture can make 5G users to use both LTE and NR spectrum.

Increase spectrum effectiveness

In 5G air technology, the best technique to improve spectrum efficiency is massive beamforming. But the deployment of massive beamforming will challenge the transport bandwidth of fronthaul.

By supporting low layer split, the 5G network architecture can significantly reduce the requirement for transport bandwidth and make the deployment of massive beamforming easier.

Increase site density

In mobile networks, the higher the site density, the more interference between stations, the more complex the coordination of radio resource management, and the lack of solution to these problems could bring more negative gain.

By supporting high layer split and centralized deployment of CU-CP and CU-UP , the 5G network architecture can get better coordination between DUs, get better load balancing and interference cancellation performance, and improve the benefits from increasing site density.

3.2. Low latency

E2E delay in the mobile network depends mainly on the time delay between application server ,CN and RAN. The best way to reduce the delay is to deploy the application server, CN and RAN together. However, in a legacy RAN architecture such as LTE, the cost is unacceptable to deploy the application server and CN function entity in the RAN node. Because eNB is made up of specialized hardware that is not applicable for the deployment architecture. In this case, additional hardware for deploying the application server as well as CN function must be provided. This would bring high cost for each distributed eNB.

In 5G network architecture, the application server and CN function entity and CU can be deployed together by reusing the CU execution environment. It can reduce the hardware cost of deploying. The virtualization of CU introduce the COTS hardware as resource pool carrying network function entities, and this would further reduce the deployment cost.

3.3. Large Connections

Large connections are mainly for the Internet of things (IoT) scenarios, and all kinds of terminals communication requirements of IoT are different, such as water meters generally deployed in the hidden space that makes more demanding on the mobile network coverage; the terminals used on the shared bikes require more energy efficiency for location tracking and communication, which would reduce the cost of maintaining the battery.

In order to meet the high demands of these new applications, new QoS framework and network slice mechanism are introduced in the 5G network architecture. The new QOS framework can define more dimensions of QoS, such as coverage, energy efficiency, delay, etc. And the network slicing can be achieved via dividing E2E network into slices for each industries and orchestrate functional entities depending on the characteristics of service requirements.

3.4. Multiple use cases coexist

The requirements of high bandwidth, low latency, large connections are very different. All these requirements are the typical scenarios and use cases of 5G network, and it is impossible to deploy a mobile network for each use case or scenario.

Similar to the large connections scenario mentioned above, by using network slices, 5G network architecture can support multiple user cases coexisting.

Operators can provide E2E network slice for each scenarios. Orchestration of network function entities and resource allocation can be performed according to the characteristics of each use case. As the operation and maintenance of network slicing is independent, the same cellular network can provide specific services for different vertical business domains.

4. Summary

5G is a revolutionary technology that will enable much more than just broadband services. A variety of joined-up services across a range of verticals, accessible from any device over any connectivity medium will be offered. The adaptable 5G network with a centralized and distributed cloud-based network architecture would enable the delivery of high-bandwidth, low-latency experiences and enhanced productivity to form a connected world.

5. Abbreviations

5GC	5G Core Network
AMF	Access and Mobility Management Function
BBU	Baseband Unit
CN	Core Network
CU	Centralized Unit
COTS	Commercial-off-the-shelf
CPRI	Common Public Radio Interface
DU	Distributed Unit
eMBB	Enhanced Mobile BroadBand
EN-DC	E-UTRA-NR Dual Connectivity
EPC	Evolved Packet Core
QFI	QoS Flow ID
QoS	Quality of Service
LAA	Licensed Assisted Access
mMTC	Massive Machine Type Communications
MN	Master Node
MR-DC	Multi-RAT Dual Connectivity
NE-DC	NR-E-UTRA Dual Connectivity
NGEN-DC	NG-RAN E-UTRA-NR Dual Connectivity
NG-RAN	NG Radio Access Network
NR	NR Radio Access
PaaS	Platform as a Service
RAN	Radio Access Network
RE	Radio Equipment
REC	Radio Equipment Control
RRH	Radio Remote Head
RU	Radio Unit
SN	Secondary Node
URLLC	Ultra-Reliable and Low Latency Communications
UPF	User Plane Function
vRAN	Virtualized Radio Access Network