5G Network Architecture

A High-Level Perspective

HUAWEI TECHNOLOGIES CO., LTD.

Contents

A Cloud-Native 5G Architecture is Key to Enabling Diversified Service Requirements

Through persistent effort and determination Telecom operators are implementing a digital transformation to create a better digital world. To provide enterprises and individuals with a real time, on demand, all online, DIY, social (ROADS) experience requires an end-to-end (E2E) coordinated architecture featuring agile, automatic, and intelligent operation during each phase. The comprehensive cloud adaptation of networks, operation systems, and services is a prerequisite for this much anticipated digital transformation.

The "All Cloud" strategy is an illuminated exploration into hardware resource pools, distributed software architecture, and automatic deployment. Operators transform networks using a network architecture based on data center (DC) in which all functions and service applications are running on the cloud DC, referred to as a Cloud-Native architecture.

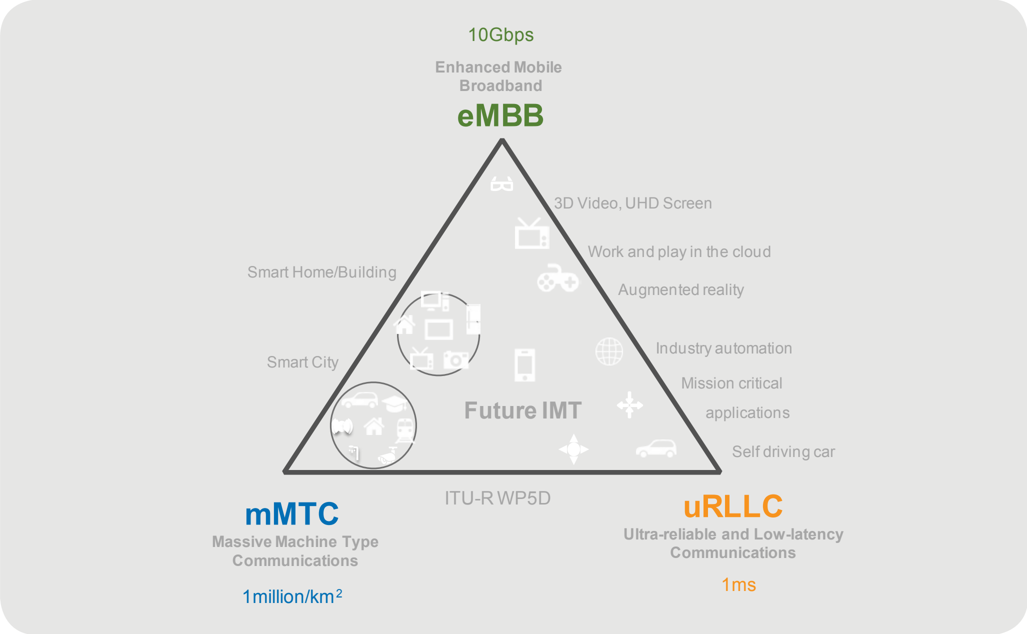
In the 5G era, a single network infrastructure can meet diversified service requirements. A Cloud-Native E2E network architecture has the following attributes:

- Provides logically independent network slicing on a single network infrastructure to meet diversified service requirements and provides DC-based cloud architecture to support various application scenarios.
- Uses CloudRAN to reconstruct radio access networks (RAN) to provide massive connections of multiple standards and implement on-demand deployment of RAN functions required by 5G.
- Simplifies core network architecture to implement on-demand configuration of network functions through control and user plane separation, component-based functions, and unified database management.
- Implements automatic network slicing service generation, maintenance, and termination for various services to reduce operating expenses through agile network O&M.

5G Will Enrich the Telecommunication Ecosystem

In the new exciting era of 5G, new communication requirements pose challenges on existing networks in terms of technologies and business models. The next-generation mobile network must meet diversified demands. The International Telecommunication Union (ITU) has classified 5G mobile network services into three categories: Enhanced Mobile Broadband (eMBB), Ultra-reliable and Low-latency Communications (uRLLC), and Massive Machine Type Communications (mMTC). eMBB aims to meet the people's demand for an increasingly digital lifestyle, and focuses on services that have high requirements for bandwidth, such as high definition (HD) videos, virtual reality (VR), and augmented reality (AR). uRLLC aims to meet expectations for the demanding digital industry and focuses on latency-sensitive services, such as assisted and automated driving, and remote management. mMTC aims to meet demands for a further developed digital society and focuses on services that include high requirements for connection density, such as smart city and smart agriculture.

The expansion of service scope for mobile networks enriches the telecom network ecosystem. A number of traditional industries, such as automotive, healthcare, energy, and municipal systems participate in the construction of this ecosystem. 5G is the beginning of the promotion of digitalization from personal entertainment to society interconnection. Digitalization creates tremendous opportunities for the mobile communication industry but poses strict challenges towards mobile communication technologies.


10Gbps
Enhanced Mobile Broadband
eMBB
3D Video, UHD Screen
Work and play in the cloud
Augmented reality
Smart Home/Building
Smart City
Industry automation
Mission critical applications
Self driving car
Future IMT
ITU-R WP5D
mMTC
Massive Machine Type Communications
1million/km²
uRLLC
Ultra-reliable and Low-latency Communications
1ms

A. The Driving Force Behind Network Architecture Transformation

The existing mobile network architecture was designed to meet requirements for voice and conventional MBB services. However, this previous organization has proven to be insufficiently flexible to support diversified 5G services due to multiple 3GPP version upgrades, a large number of NEs, complex interfaces. The driving force behind the network architecture transformation includes the following aspects:

· Complex networks incorporating multiple services, standards, and site types

5G networks must be able to provide diversified services of different KPIs, support co-existent accesses of multiple standards (5G, LTE, and Wi-Fi), and coordinate different site types (macro, micro, and pico base stations). The design challenge to create a network architecture capable of supporting such flexibility whilst meeting differentiated access demands is a brave endeavor to satisfy.

· Coordination of multi–connectivity technologies

5G is expected to co-exist with LTE and Wi-Fi for an extended period of time incorporating multi-connectivity technologies and the new 5G air interface. Multi-connectivity technologies must be coordinated based on traffic and mobility requirements of user equipment to provide sufficient transmission throughput and mobile continuity.

· On–demand deployment of service anchors

5G network architecture will be designed based on access sites and three-layer DCs. According to different service requirements, fiber/optic cable availability and network resource allocations, RAN real time and non-real time resources can be deployed on the site or on the access cloud side. This further requires that the service gateway location may also be deployed on the access cloud or on the core network side.

· Flexible orchestration of network functions

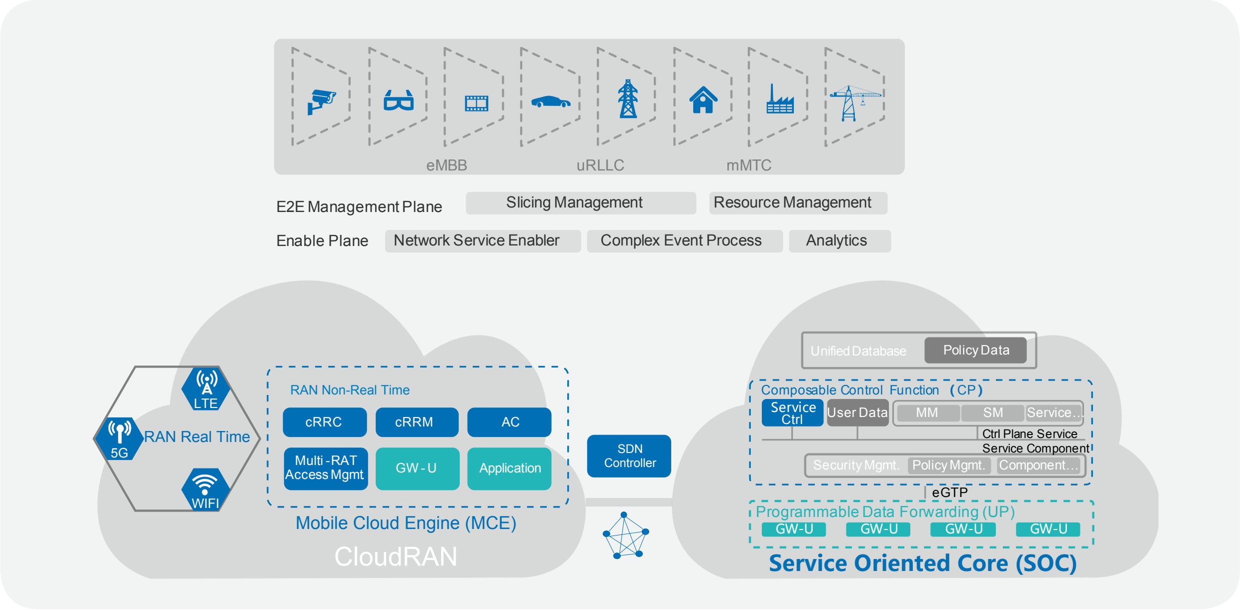
Service requirements vary with different network functions. eMBB requires a large throughput for scheduling. uRLLC requires ultra-low latency and high reliability. Networks must flexibly orchestrate network capabilities considering service characteristics, which significantly simplify network functions and increase network efficiency.

· Shorter period of service deployment

Various services have expanded the mobile network ecosystem and increased network deployment complexity. Rapidly deploying new services requires an improved set of lifecycle management processes involving network design, service deployment, and O&M.

B. The Service-Driven 5G Architecture

The service-driven 5G network architecture aims to flexibly and efficiently meet diversified mobile service requirements. With software-defined networking (SDN) and Network Functions Virtualization (NFV) supporting the underlying physical infrastructure, 5G comprehensively cloudifies access, transport, and core networks. Cloud adoption allows for better support for diversified 5G services, and enables the key technologies of E2E network slicing, on-demand deployment of service anchors, and component-based network functions.

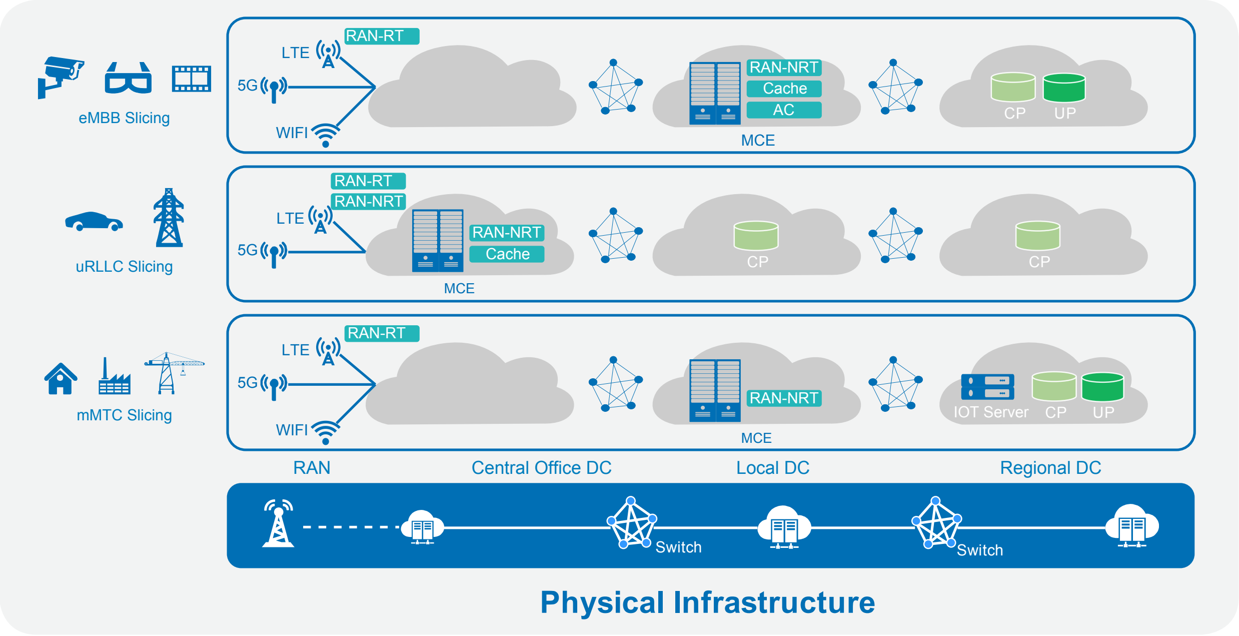


CloudRAN consists of sites and mobile cloud engines. This facility coordinates multiple services, operating on different standards, in various site types for RAN real time resources that require a number of computing resources. Multi-connectivity is introduced to allow on-demand network deployment for RAN non-real time resources. Networks implement policy control using dynamic policy, semi-static user, and static network data stored in the unified database on the core network side. Component-based control planes and programmable user planes allow for network function orchestration to ensure that networks can select corresponding control-plane or user-plane functions according to different service requirements. The transport network consists of SDN controllers and underlying forwarding nodes. SDN controllers generate a series of specific data forwarding paths based on network topology and service requirements. The enabling plane abstracts and analyzes network capabilities to implement network optimization or open network capabilities in the form of API. The top layer of the network architecture implements E2E automatic slicing and network resource management.

End-to-End Network Slicing for Multiple Industries Based on One Physical Infrastructure

E2E network slicing is a foundation to support diversified 5G services and is key to 5G network architecture evolution. Based on NFV and SDN, physical infrastructure of the future network architecture consists of sites and three-layer DCs. Sites support multiple modes (such as 5G, LTE, and Wi-Fi) in the form of macro, micro, and pico base stations to implement the RAN real time function. These functions have high requirements for computing capability and real time performance and require the inclusion of specific dedicated hardware. Three-layer cloud DC consists of computing and storage resources. The bottom layer is the central office DC, which is closest in relative proximity to the base station side. The second layer is the local DC, and the upper layer is the regional DC, with each layer of arranged DCs connected through transport networks.

According to diversified service requirements, networks generate corresponding network topologies and a series of network function sets (network slices) for each corresponding service type using NFV on a unified physical infrastructure. Each network slice is derived from a unified physical network infrastructure, which greatly reduces subsequent operators' network construction costs. Network slices feature a logical arrangement and are separated as individual structures, which allows for heavily customizable service functions and independent O&M.



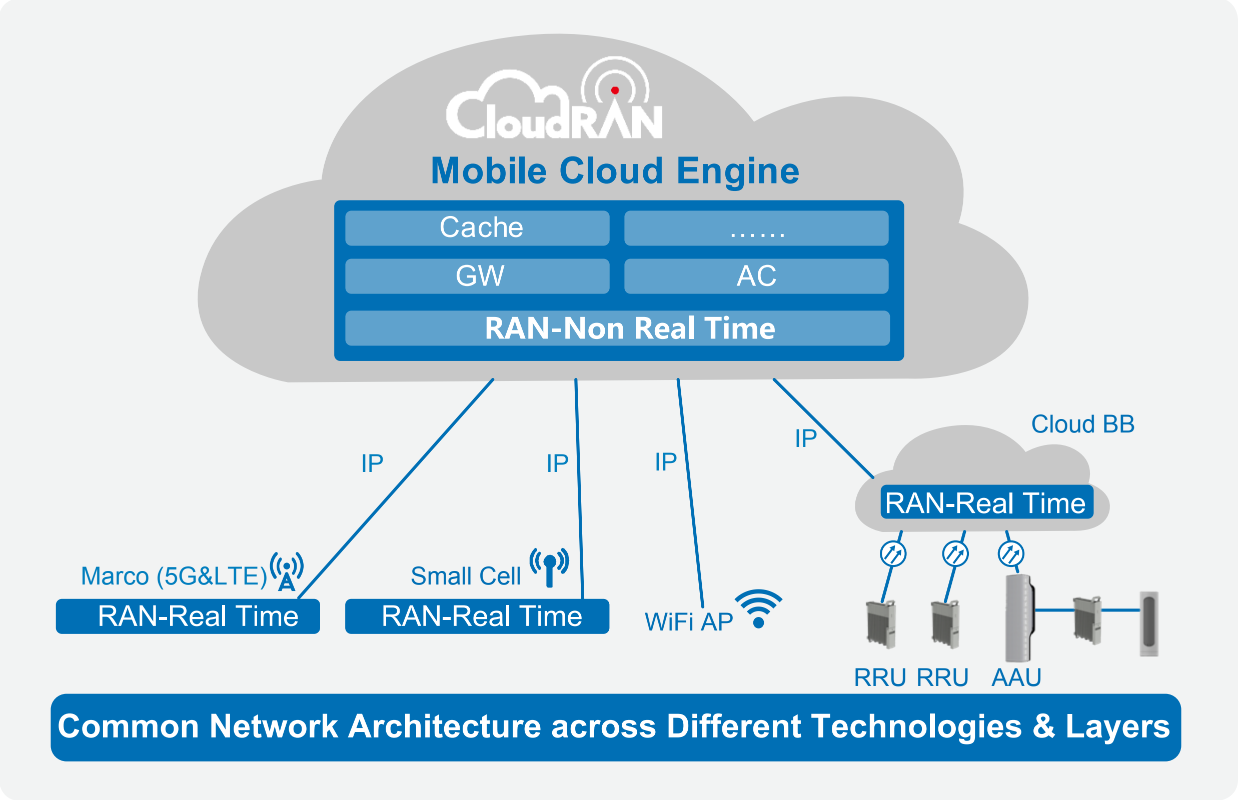
As illustrated in the preceding figure, eMBB, uRLLC, and mMTC are independently supported on a single physical infrastructure. eMBB slicing has high requirements for bandwidth to deploy cache in the mobile cloud engine of a local DC, which provides high-speed services located in close proximity to users, reducing bandwidth requirements of backbone networks. uRLLC slicing has strict latency requirements in application scenarios of self-driving, assistant driving, and remote management. RAN real Time and non-Real Time processing function units must be deployed on the site side providing a beneficial location preferably based in close proximity to users. V2X Server and service gateways must be deployed in the mobile cloud engine of the central office DC, with only control-plane functions deployed in the local and regional DCs. mMTC slicing involves a small amount of network data interaction and a low frequency of signaling interaction in most MTC scenarios. This consequently allows the mobile cloud engine to be deployed in the local DC, and other additional functions and application servers can be deployed in the regional DC, which releases central office resources and reduces operating expenses.

Reconstructing the RAN with Cloud

During the course of an evolution towards RAN2020, CloudRAN architecture is used on the RAN side to implement RAN Real Time functions, on-demand deployment of non-real time resources, component-based functions, flexible coordination, and RAN slicing. With Mobile Cloud Engine (MCE), CloudRAN can implement flexible orchestration for RAN Real time and non-real Time functions based on different service requirements and transmission resource configuration to perform cloudification of the RAN.

The RAN real time functions include access network scheduling, link adaptation, power control, interference coordination, retransmission, modulation, and coding. These functions require high real-time performance and computing load. The deployment of sites must include dedicated hardware with high accelerator processing specifications and performance, whilst located in close proximity to services. The RAN non-real time functions include inter-cell handover, cell selection and reselection, user-plane encryption, and multiple connection convergence. These functions require minimal real-time performance, latency requirements to dozens of milliseconds and are suitable for centralized deployment. A universal processor can be deployed in a MCE or site according to vast service requirements.

MCE can implement complex management while coordinating multiple processing capabilities based on regional time, frequency bands, and space. This upgraded management system allows CloudRAN to support 4G, 4.5G, 5G, and Wi-Fi, and implement coordination and scheduling of macro, micro, and pico site types. Network functions are deployed on radio, backbone, or core convergence nodes to maximize both network efficiency and additional capabilities.



A. Multi-Connectivity Is Key to High Speed and Reliability

Multi-connectivity is gaining a reputation as an underlying fundamental construct for the deployment of the future network architecture. CloudRAN can be seamlessly deployed in a unified network architecture. This is a huge leap in radio network deployment. In current fragmented networks, increasing speed and reducing latency can improve user experience. Reliable high-speed data cannot depend on a single frequency band or standard connections. In heterogeneous networks, multi-connectivity helps provide an optimal user experience based on LTE and 5G capabilities, such as high bandwidth and rates of high frequency, network coverage and reliable mobility of low frequency, and accessible Wi-Fi resources. In scenarios that require high bandwidth or continuity, a user requires multiple concurrent connections. For example, data aggregation from multiple subscriptions to 5G, LTE, and Wi-Fi is required to produce high bandwidth. An LTE network access is required to maintain continuity after a user has accessed a 5G high-frequency small cell.

In scenarios that source multiple technologies, CloudRAN serves as an anchor for data connection which noticeably reduces alternative transmission. In the traditional architecture integrating base stations as an anchor for data connection, LTE, 5G, and Wi-Fi data is aggregated into a non-real time processing module of a specific standard to be forwarded to each access point. In the CloudRAN architecture, non-real time processing function modules in access points of different modes are integrated into the MCE, which serves as an anchor for data connection. Data flows are transmitted to each access point over the MCE, which prevents alternative transmission and reduces transmission investment by 15%, and latency by 10 ms.

B. MCE

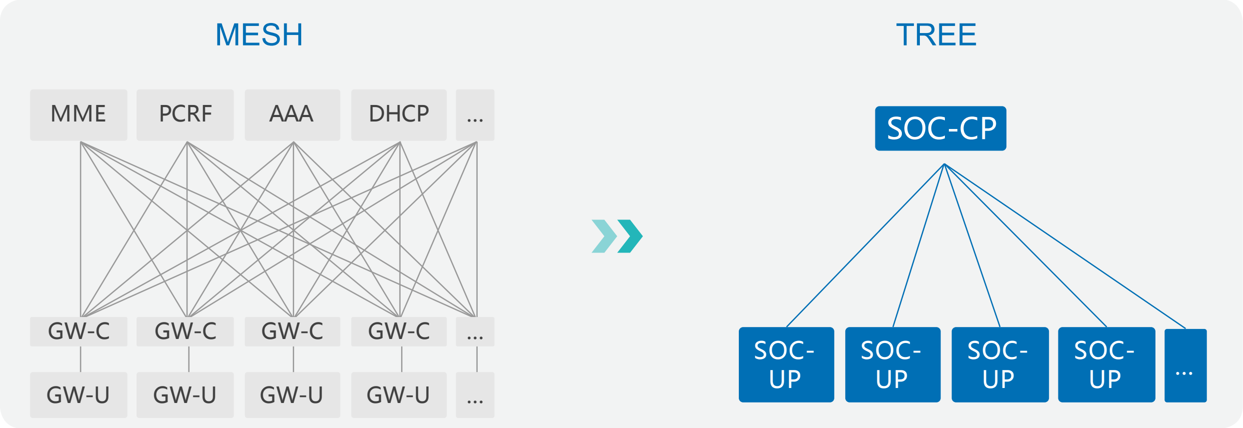
MCE is the logical entity of central control and management for CloudRAN, incorporating RAN non-real time functions, Wi-Fi AC, distributed gateway, service-related application distribution entity (App), and Cache. RAN non-real time functions include a general control plane (cRRC) to facilitate multi-connectivity and new technology deployment, and a centralized resource management module (cRRM) to ensure the efficient coordination of resources in heterogeneous networks. Cloud-based SON (cSON) is introduced to improve network capacity, coverage, and transmission resources to encompass vast extended areas and ensure the successful implementation of slicing management.

MCE can run on a dedicated platform and general COTS platform that are deployed above the Cloud OS and the COTS-based cloud infrastructure. This is to provide carrier-grade disaster recovery capability, on-demand deployment based on Cloud-Native architecture, flexible scale-in and scale-out functionality, and independent feature upgrades.

Cloud-Native New Core Architecture

A. Control and User Plane Separation Simplifies the Core Network

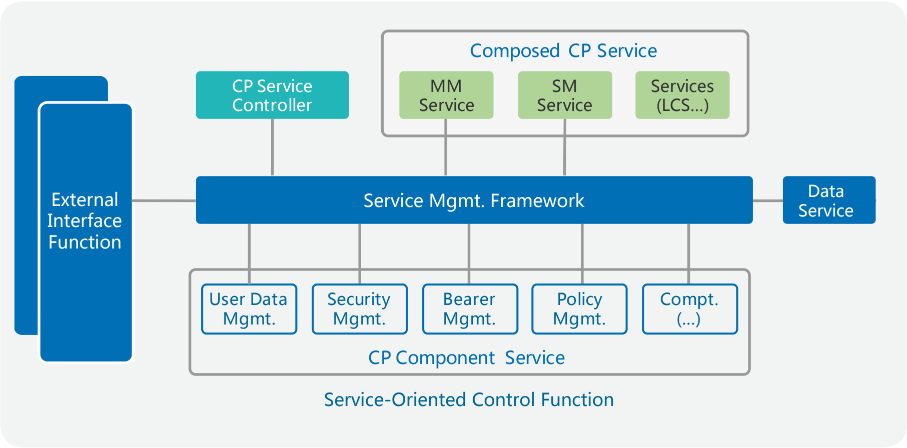
Existing network gateways integrate parts of both user plane and control plane functions. In the 5G era, many services with high requirements for latency require gateways to be relocated by a downward shift towards the local or central office DCs. This requires that the number of gateway nodes must increase by a factor of 20 to 30 times the original amount. If operators still opt to use the existing gateway architecture, complex gateway service configuration will significantly increase CAPEX and OPEX. In addition, if the control plane has subscribed reports of location and RAT information, a large amount of signaling will be generated between the site, distributed gateway, and network control plane. A large number of distributed gateways will result in heavy interface link load and handover signaling load on centralized control plane NEs.

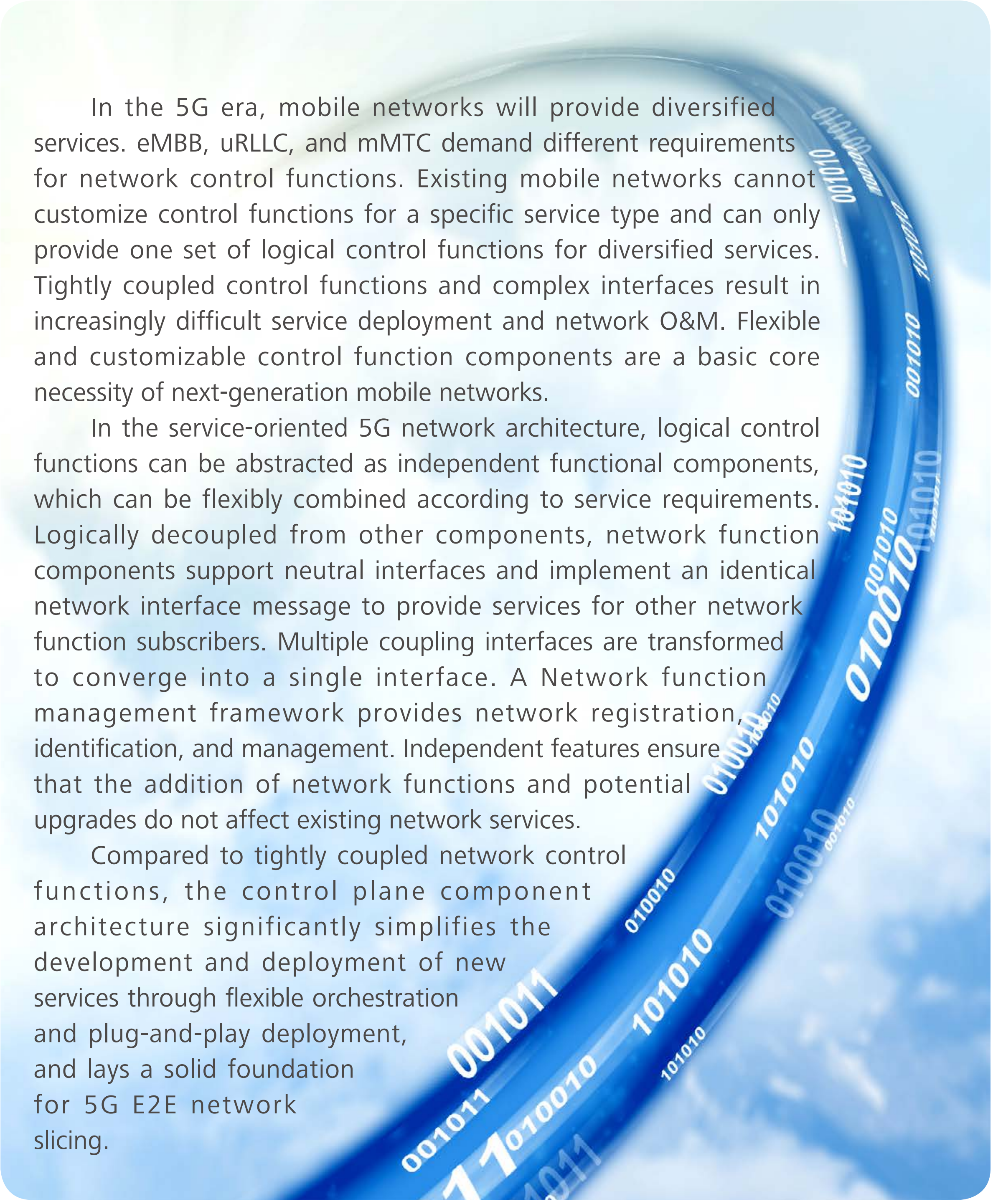


Gateway control and user plane separation divides complex control logic functions for convergence into control planes, which reduces the costs of distributed gateway deployment, interface load, and number of alternative signaling routes. In addition, the control plane and user plane separation supports scaling of the forwarding and control planes, which further improves network architecture flexibility, facilitates centralized control logic functions, and ensures easy network slicing for diversified industry applications. This segregation technique also decouples the forwarding plane from the control plane, which prevents frequent forwarding plane upgrades caused by control plane evolution. Two tasks must be completed to implement control and user plane separation. First, an implementation of lightweight functions to divide complex control logic functions. Second, the construction of models for the reserved core functions with the definition of a generalized template model complete with object-oriented interface for the forwarding plane to ensure that the forwarding plane is both programmable and scalable.

After the control and user planes are successfully separated, interfaces providing the associative link connections operate through the enhanced GTP protocol. Based on subscriber access types and subscription data, the control plane initiates an orchestration for service objects and atomic actions, and sends the request to the forwarding plane over the enhanced GTP interface. The forwarding plane then responds with a service-based event notification confirming receipt which is directed back to the control plane.

B. Flexible Network Components Satisfy Various Service Requirements



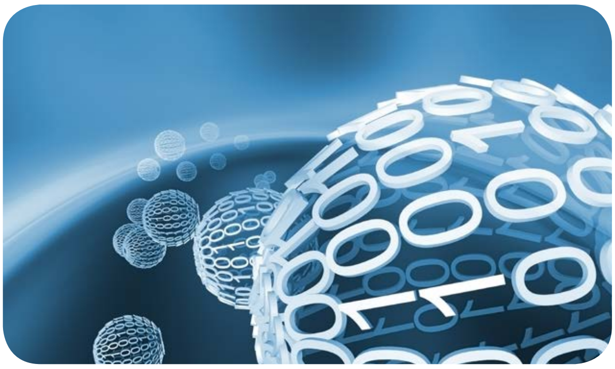


In the 5G era, mobile networks will provide diversified services. eMBB, uRLLC, and mMTC demand different requirements for network control functions. Existing mobile networks cannot customize control functions for a specific service type and can only provide one set of logical control functions for diversified services. Tightly coupled control functions and complex interfaces result in increasingly difficult service deployment and network O&M. Flexible and customizable control function components are a basic core necessity of next-generation mobile networks.

In the service-oriented 5G network architecture, logical control functions can be abstracted as independent functional components, which can be flexibly combined according to service requirements. Logically decoupled from other components, network function components support neutral interfaces and implement an identical network interface message to provide services for other network function subscribers. Multiple coupling interfaces are transformed to converge into a single interface. A Network function management framework provides network registration, identification, and management. Independent features ensure that the addition of network functions and potential upgrades do not affect existing network services.

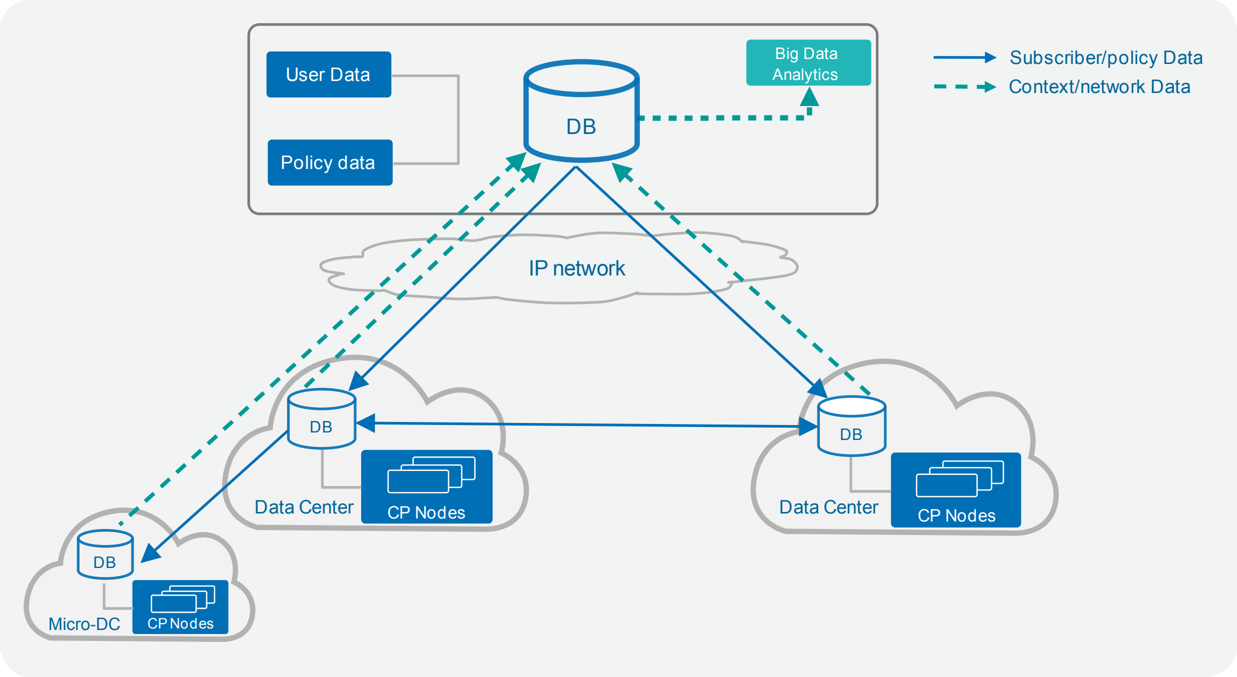
Compared to tightly coupled network control functions, the control plane component architecture significantly simplifies the development and deployment of new services through flexible orchestration and plug-and-play deployment, and lays a solid foundation for 5G E2E network slicing.

C. Unified Database Management



Rapid fault recovery is required for network data status information (such as user data and policy data shared across data centers), to meet network reliability requirements after the virtualization of functions. The traditional disaster recovery mechanism based on N+1 backup relies on private signaling interaction to implement status information synchronization, which produces system inefficiency and complex interaction of cross-vendor products.

With separated data and control logic, network status information can be centralized in a unified database. All network functions can access metadata models through standard interfaces and locally store dynamic user data. Thanks to the distributed database synchronization, network status information can implement real-time backup between data centers. With the help of the service management framework, the unified database simplifies the procedure for network information retrieval functions introduced by the component-based control plane to reduce the required signaling overhead for data synchronization.

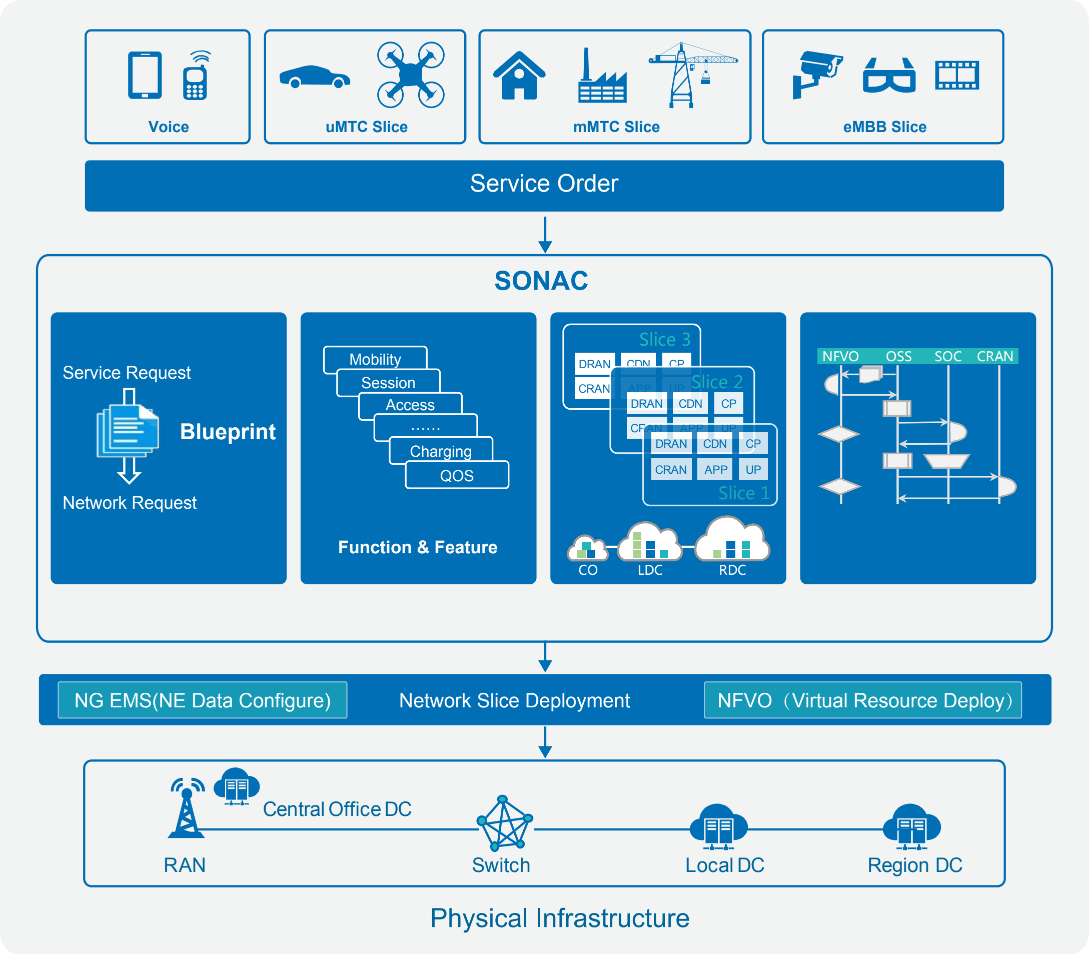


Self-Service Agile Operation

One of the targets and driving forces of network architecture evolution is to provide diversified services using mobile networks. E2E network slicing is a fundamental technology to achieve this target. In the 5G era, a network will contain multiple logically separated network slices. Each slice has a specific network topology, network function, and resource allocation model. If manual configuration is still used for network planning and deployment, operators' O&M system will potentially face a huge number of significant challenges.

5G networks will possess self-serving agile operation capabilities. Network slicing services can be automatically generated, maintained, or terminated according to services requirements, which significantly reduces subsequent operating expenses. Third-party vertical industries can input mobile network slicing requirements on an operation platform. The operator analyzes customer requirements based on current network status.

After a service level agreement procedure is complete, the operator maps



various service requirements on network requirements, and selects multiple network function components to generate a network slice. According to service features and deployment of data centers, the operator determines logical network function deployment nodes and defines a connection relationship, namely software-defined topology (SDT). After the network slicing topology is defined, an E2E protocol is defined, namely, software-defined protocol (SDP). According to service requirements, network resources are allocated for logical connections in the logical topology, namely software-defined resource allocation (SDRA). SDT, SDP, and SDRA constitute a list of key functions required for Service Oriented Network Auto Creation (SONAC).

Conclusion: Cloud-Native Architecture is the Foundation of 5G Innovation

In existing networks, operators have gradually used SDN and NFV to implement ICT network hardware virtualization, but retain a conventional operational model and software architecture. 5G networks require continuous innovation through cloud adoption to customize network functions and enable on-demand network definition and implementation and automatic O&M.

Physical networks are constructed based on DCs to pool hardware resources (including part of RAN and core network devices), which maximizes resource utilization. In addition, E2E network slicing provides logically separated virtualized network slices for diversified services, which significantly simplifies network construction for dedicated services.

CloudRAN is built based on MCE. Multi-connectivity helps aggregate access capabilities of multiple RATs, frequency bands, and site types to maximize network efficiency. Flexible deployment of network functions helps customize networks for various differentiated services. CloudRAN allows operators to address challenges and proof themselves against potential prospective uncertainties.

Based on the control and user plane separation, 5G core networks using component-based control planes, programmable user planes, and unified database will simplify signaling interaction and allow for the deployment of distributed gateways. Customized network functions can allow operators to generate increasingly flexible additional network slices to better serve subscribers needs.

SONAC implements 5G automation using SDT, SDP, and SDRA to ensure automatic implementation of service deployment, resource scheduling, and fault recovery based strictly on a detailed and thorough network data analysis.

NO WARRANTY

THE CONTENTS OF THIS MANUAL ARE PROVIDED "AS IS". EXCEPT AS REQUIRED BY APPLICABLE LAWS, NO WARRANTIES OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE,ARE MADE IN RELATION TO THE ACCURACY, RELIABILITY OR CONTENTS OF THIS MANUAL.
TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, IN NO CASE SHALL HUAWEI TECHNOLOGIES CO., LTD BE LIABLE FOR ANY SPECIAL, INCIDENTAL, INDIRECT, OR CONSEQUENTIAL DAMAGES, OR LOST PROFITS, BUSINESS, REVENUE, DATA, GOODWILL OR ANTICIPATED SAVINGS ARISING OUT OF OR IN CONNECTION WITH THE USE OF THIS MANUAL.

HUAWEI TECHNOLOGIES CO., LTD.
Bantian, Longgang District
Shenzhen518129, P. R. China
Tel:+86-755-28780808

www.huawei.com