中国移动 China Mobile | 研究院 CMRI

6G Service-Based RAN White Paper

（2022）

China Mobile Research Institute

Preface

The fifth generation (5G) network has initiated a new era of the Internet of everything that supports abundant mobile services and Internet of Things services. In order to meet the requirements of fast delivery of diversified services, 5G core network (CN) revolutionarily takes the service-based architecture (SBA) as infrastructure, and makes optimization on the decoupling of control plane and user plane. Control plane of the 5G CN is disassembled into different network functions (NFs) based on modularization. The disassembled network functions can be independently expanded, independently evolved and on-demand deployed. Each NF has multiple network function services (NFSs). Thus on-demand customization of CN NFs and flexibly supporting of different scenarios and requirements can be implemented.

At present, research on SBA mainly focuses on the CN control plane. Research on service-based RAN is still in its initial stage. For a long time, base stations have been developed in an integrated monolithic way to guarantee the ultimate performance of the "last mile". Mostly due to the performance, academia and industry are hesitating to develop service-based RAN. But from our perspective, performance concerns would not be an obstacle in exploring service-based RAN. On one hand, current performance indicators defined by 3GPP/ITU are for air interface, but in some scenarios, such indicators as high reliability and low latency are not the most important, but the end-to-end performance indicators are what we need to pay attention to. On the other hand, there is a "game" relationship between performance indicators. Although it seems that service-based RAN may reduce some performance, its performance loss can be compensated by improving system stability, availability and other capabilities. In addition, further development in cloud native technology and chip capabilities will also mitigate these performance losses.

Facing uncertain new business and new scenario requirements in the future [1], we should focus more on improving the adaptability of the network to all scenarios. At network function layer, necessary network functions are flexibly combined as required to provide customized network services. At infrastructure and resource layer, appropriate network resources (including computing, storage, spectrum, power, and deployment location, etc.) are allocated as required to maximize network resource efficiency. At application and service layer, service requirements should be accurately sensed, also network functions and network resources should be intelligently orchestrated and managed in a multi-dimensional manner to fully adapt to various scenarios. We believe that the end-to-end service architecture based on cloud native technology is a necessary technical means to improve network ability of scenario adaptation. In order to maximize the adaptability of network, research on service-based RAN is a top priority in future network architecture design [2-3].

Starting from the drivers of service-based RAN, this white paper presents the design principles, overall vision, key technologies, and development challenges of service-based RAN, as well as possible application scenarios and standardization prospect. We hope to work with industry partners to promote service-based RAN research.

Contents

1. Cloud RAN Leads Industrial Development and Paves the Way for Service-based RAN

To meet the requirements of diversified services on the flexibility and scalability of the 5G network, the softwarization, virtualization and cloudification of network have become an inevitable trend. Cloud RAN has also been widely favored by global operators.

Cloud RAN helps operators build low-cost, resource-efficient, green, fully automated radio networks. It has two core features that lay the foundation for successful deployment of cloud RAN:

- ***Common hardware platforms (including accelerators):*** The capabilities of traditional dedicated hardware base station (BS) are limited by the capabilities of hardware. A common approach to improve the signal processing capacity of BS is to supplement the baseband processing boards. Additional hardware boards will not only increase equipment costs, but also increase power consumption which is heavily affected by the number of boards and is less related to the work load. Different from traditional dedicated hardware BSs, RAN functional software of cloud RAN is carried on the commercial-off-the-shelf (COTS) platform in cloud RAN, which is facilitated for the sharing of hardware resources and the on-demand deployment of functional software. It is expected that the equipment cost will be reduced by 40% and operation cost by 30%, and the power consumption of each BS is expected to be saved by about 5% in typical scenarios.
- ***Cloud native technology***: In order to meet the user's needs, equipment manufacturers have to constantly upgrade the software version of BS. Currently, the upgrade is cumbersome, time-consuming, error-prone, and prone to introduce service interruptions. Differently, in cloud RAN, with cloud native technologies such as Kubernetes and the principle of DevOps, the RAN functions can be deployed in containers on bare servers, leading to a faster version upgrade and application online.

Equipment vendors are actively developing the cloud RAN products, and the performance differences of the products with dedicated hardware are gradually narrowing. As early as 2016, the world's first cloud RAN system developed by Nokia is commercialized in South Korea, which has been large-scale commercial deployed in mid-2019. The second generation of the system with virtualized centralized units (CUs) and virtualized distributed units (DUs) is in trial and will be commercialized in 2022. Ericsson has also launched a series of cloud RAN products and is leveraging Intel Xeon processors and other technologies to improve cloud RAN performance for 5G and beyond. During the same period, Samsung deployed the industry's first fully virtualized end-to-end (E2E) commercial 5G RAN, demonstrating that the virtualized RAN solutions are capable of processing massive amounts of data and its performance can achieve the same level as traditional dedicated hardware products.

Operators are actively carrying out large-scale deployment of building base band unit (BBU) -centralized RAN, and cloud RAN has become the general trend. According to IDC, 80 percent of enterprises will accelerate the applications on cloud by the end of 2021. Chinese mobile operators also put forward "the integration of cloud and network" as one of the main goals of future network construction. China Mobile is vigorously promoting the deployment of cloud network, and is expected to achieve 100% cloud core network by 2025. By July 2021, China Mobile has deployed more than 500,000 5G base stations, and the deployment proportion of centralized BBU will exceed 70% in 2021 and 75% in 2022. Centralized BBU will contribute to reducing power consumption and facilitate collaborative, virtualization, and cloud deployment. AT&T has also signed a five-year agreement with Ericsson to accelerate the deployment of its 5G centralized RAN network, paving the way for the evolution of Cloud RAN.

Virtualized RAN and cloud RAN are not the ultimate goal of future radio networks. In our understanding, the service-based RAN would be the next step of cloud RAN. Cloud RAN has provided a flexible and scalable platform for future radio networks, but its service capabilities remain limited: From the functional point of view, the minimum development granularity of BBU is CU or DU, which is relatively large and still cannot meet the requirements of rapid online and flexible deployment of specific new functions. From the perspective of interface, point-to-point dedicated interfaces are still used within BS, between BSs, and between BS and core network (CN). Whenever the BS or CN network functions (NFs) change, the corresponding modifications have to be made on related interfaces, resulting in high standardization workload and high complexity in operation administration and maintenance (OAM).

In order to respond nimbly to more diversified function requirements, quality of service (QoS) requirements, management strategy requirements, deployment requirements and exposure requirements of the future applications and scenarios, and make the network capable of forwarding compatibility, it is necessary to make efforts to improve the service capability of the next generation of radio access network so as to better leverage the platform advantages of cloud RAN.

- Through the definition of RAN function services, the upgrade of base station software version will be realized faster, and the application functional requirements will be met in time;
- Through the definition of service-based interfaces between RAN and CN services, a new way of interaction will be brought into the E2E procedures.
- Through the definition of service-based interface between RAN services and third party services, more timely and multidimensional radio network capability will be exposed.
- Through the cloud native infrastructure platform, hardware resources can be pooled and shared, reducing the cost and power consumption of the entire network, and quickly responding to service deployment requirements.
- Through the integrated orchestration and management of RAN and CN services, the complexity of OAM can be reduced and the adaptability of the network to new services can be improved.

2. Vision of Service-based RAN

2.1 Design Principles of Service-based RAN

Before innovating the radio access network architecture through micro-services, it is necessary to determine the basic design principles of service-based RAN to avoid the "Distributed Monolith" caused by unreasonable service division. The design principles of service-based RAN mainly include the following five aspects:

1. **The services provided by RAN need to be defined according to external requirements.** The network is used to deal with business requirements, so the first step in defining a service-based RAN design is to classify external requirements into key requests. For the access network, the requirements may come from the CN NFs, the access network nodes, the third-party applications or user equipment (UE).

2. **The definition of RAN services needs to meet the requirements of "loose coupling" and "high cohesion" [5-6].** Robert Martin has an exposition of the Single Responsibility Principle: "Gather together those things that change for the same reason, and separate those things that change for different reasons." That is, there should be only one reason to change a service, and each responsibility

carried by the service is the potential reason to modify it. Micro-services should be designed to be as small, cohesive, and contain only a single responsibility, which reduces the size of the service and improves its stability. But "small" is not the primary goal of micro-services.

"Loose coupling" is the core feature of micro-service architecture. A micro-service is an independent entity that can be modified independently, and the deployment of a service should not cause changes to the service consumers. For a service, it should be carefully considered what should be exposed and what should be hidden. If it is exposed too much, the service consumer will be coupled with the service provider to some extent, which will cause additional coordination work between the two services, thereby reducing the autonomy of the service.

3. Ensuring the private properties of data is one of the prerequisites for loose coupling. If data synchronization needs to be maintained between services, the modification of the service provider will greatly affect the service consumer.

4. The goal of service-based RAN is to define RAN services and access relationships between services. At present, even in the field of Information Technology (IT) where micro-services technologies are very mature, there is no specific algorithm that can assist in service splitting or service definition. Based on the previous research, we believe that the design of service-based RAN architecture can be realized through the following three steps: First, classify external requirements into key requests, namely system operations (there are two system operation modes in 5GC: request-response and subscribe-notify). Second, determine the RAN services. Third, appropriately allocate system operations to the RAN services.

5. The deep separation of the control plane and user plane functions to meet the lightweight, low-cost, and flexible deployment requirements of the user plane functionalities.

2.2 Overall Concept of Service-based RAN

Taking into account various aspects such as industry maturity and technology maturity, the development of service-based RAN may include the following five levels, and different levels may appear independently or simultaneously.

2.2.1 Level 1: Service-based Control Plane Interfaces, Enables the Direct Communication between RAN and CN NFs

In the traditional communication model, the base station and the CN NF communicate with each other using the pre-established point-to-point interface. Whenever a new function is introduced, the existing network functions need to be enhanced, and new point-to-point interfaces need to be defined between the new function and the existing network function with which it communicates.

With the evolution of 5G SBA-CN, the SBA will not be limited to the CN, but will be extended to the N2 interface between the CN and gNB CU-CP. At this level, the gNB CU-CP as a whole will operate as an RAN service and interact with the CN NFS.

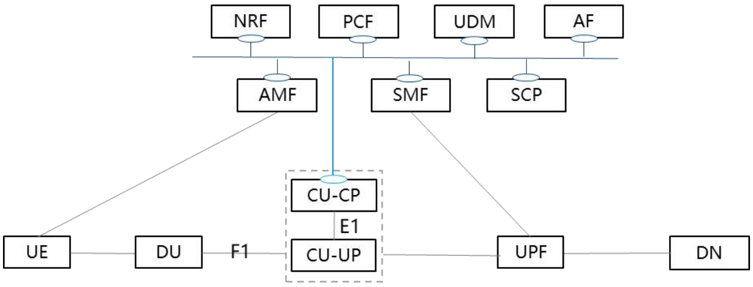

Figure 1: Level 1: Service-based control plane interfaces

2.2.2 Level 2: Service-based RAN Control Plane, Enables the Optimization of End-to-End Procedures

At this level, the control plane functions of RAN will be reconstructed into multiple RAN control plane services (CPS). The CPS can roughly include the following types: radio bearer management service (RBS), connection mobility management service (CMS), local location service (LLS), multicast broadcast service (MBS), data collection service (DCS), signaling transmission service (STS) and RAN exposure service (RES).

The service-based RAN control plane scheme can bring at least the following two technical advantages. First, RAN services can directly communicate with CN services, which will reduce unnecessary AMF forwarding in the network. Second, with the service-based RAN control plane, the interaction between the RAN CPS and other services (including CN NFs and other RAN CPSs) can be changed from serial interaction to parallel interaction, which can optimize the control plane procedures.

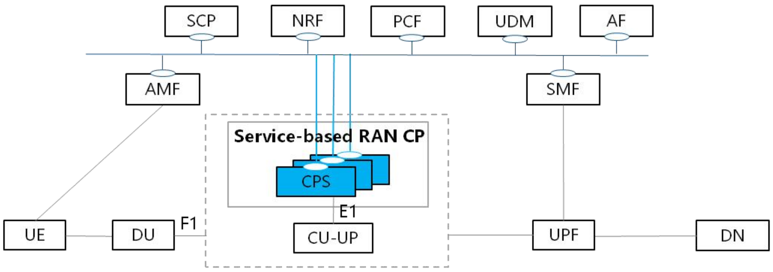

Figure 2: Level 2: Service-based RAN control plane

2.2.3 Level 3: Service-based RAN User Plane, Makes Cross-Layer Transmission Easier

Traditional mobile communication protocols follow the layered Open System Interconnection (OSI) Reference Model. Each layer receives specific services provided by its lower layer and is responsible for providing specific services to its upper layer. The interaction between the upper and lower layers follows the "interface" agreement, and the interaction between the same layer follows the "protocol" agreement. The problem with this layered design concept is that the protocol and service model are fixed, which can not realize flexible cross-layer signaling interaction and cross-layer function combination. With the help of micro-services, the traditional layered OSI reference model can be breached and the user plane function of RAN will be reconstructed into multiple RAN user plane services (UPS). With this change, the UPSs can be flexibly combined as needed to better meet various business requirements.

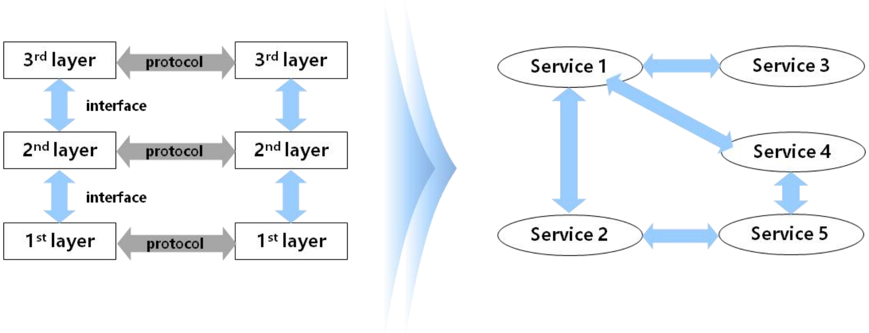

Figure 3: Service-based RAN user plane reference model

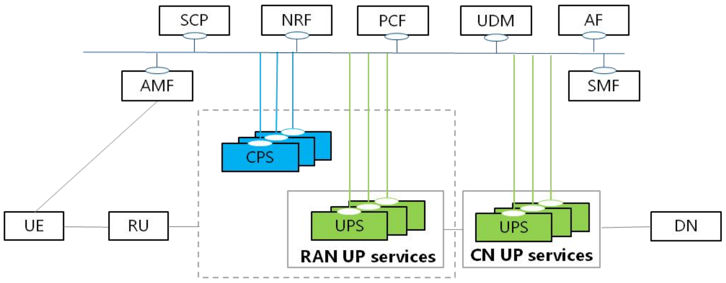

Figure 4: Level 3: Service-based RAN user plane

2.2.4 Level 4: DOICT driven Service-based RAN

With the in-depth integration and development of data, operation, information, and communication technologies (DOICT), native AI, native security, integration of sensing and communication, integration of computing and communication, integration of computing and storage have become the trends of future network development, and corresponding network service capabilities also need to be introduced into the network. For example, native AI services in the network may include AI task flow splitting services, policy generation services, data processing services, etc.

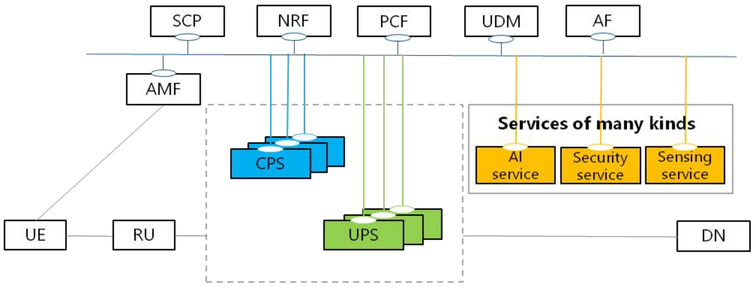

Figure 5: Level 4: DOICT driven service-based RAN

2.2.5 Level 5: Service-based UE Enables the Sharing of UE and Network Services

With the re-emergence of the cloud mobile phone market, UE is also capable of providing UE Services (UESs) such as computing power, measurement, and UE information to operator networks, third-party applications, and other UEs. UESs will be integrated with network services and flexible and direct communicate with each other through the service-based interfaces.

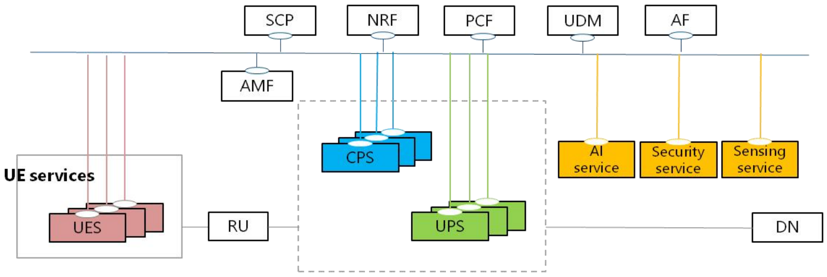

Figure 6: Level 5: Service-based UE

3. Key Technologies of Service-based RAN

The key technologies of service-based RAN cover three layers: infrastructure layer, network function layer, orchestration and management layer.

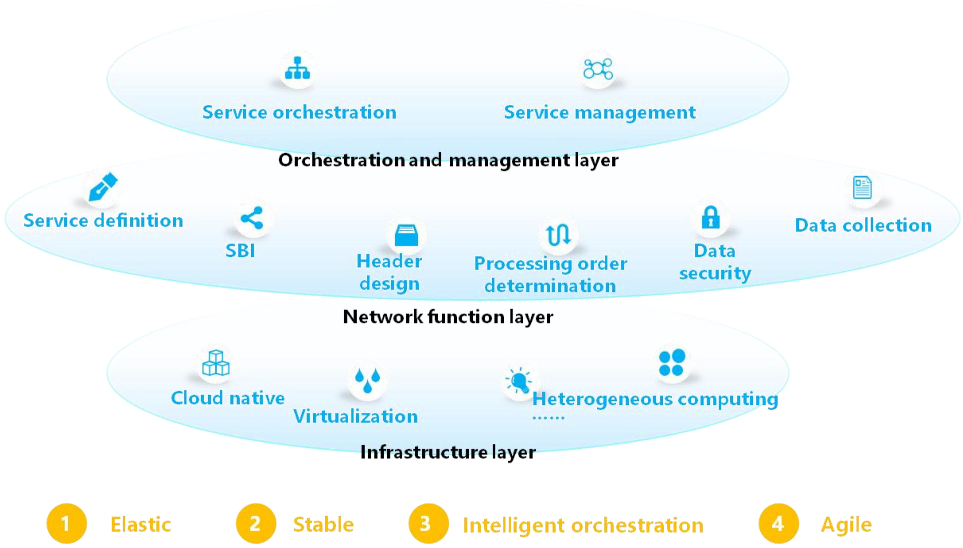

Figure 7: Service-based RAN Technical Capability Map

3.1 Key Technologies of Infrastructure Layer

3.1.1 Cloud Native

Cloud Native is a collection of ideas for a range of technologies, design patterns, and management methods, including DevOps, continuous delivery, micro-services, agile infrastructure, Conway's Law, and restructuring of the company's organizational structure. In the Top Strategic Technology Trends for 2022 Report, Gartner predicts that “cloud-native platforms will serve as the foundation for more than 95% of new digital initiatives by 2025 - up from less than 40% in 2021"[7]. Cloud native technologies have been widely used in the networks of telecommunication operators for CN elements and edge computing nodes. In the RANs, base station units based on cloud native are also being tested; some have even been put into commercial use on a small scale.

Typical cloud native technologies include container, micro-service, service mesh, immutable infrastructure, and declarative Application Programming Interface (API)[8].

1. Container

Container technology is an operating system-based virtualization technology that allows different applications to run in a separate sandbox environment without interacting with each other. Docker is one of the most popular container-based platforms. Docker container engine greatly reduces the complexity of container technology. Docker image decouples applications and operating environment, enabling applications to run consistently and reliably across different computing environments, thereby accelerating the spread of container technology. Container technology has now developed into many different forms, including full containers, edge containers, Serverless containers, bare metal containers and other forms.

2. Micro-service

Micro-services separate modules of different life cycles through a service-oriented architecture, and perform business iterations separately, thereby speeding up the overall progress and stability. Micro-services are based on containers, and each micro-service can scale and upgrade independently, which makes the deployment and iteration more efficient. Service-based architecture is API oriented: all the functions within the service are highly cohesive, and the degree of software reuse is increased through the extraction of common function modules.

3. Service Mesh

Service Mesh separates the service plane from the control plane. Functions such as service proxy, service discovery, and service governance are split out and put into a dedicated Mesh layer, and those functions in the mesh layer are transparent to the applications. After the separation, only light weighted agents (called sidecars) are kept in the application, and those agents are responsible for communication with the control plane in the Mesh layer. All the other control functions such as circuit-breaking, rate limiting, downgrade, retry, fallback, bulkhead isolation can be completed by the service mesh control plane, which can also ensure better security.

4. Serverless

Serverless is an architectural concept, the core idea is to abstract the infrastructure functions that provide different resources into different services. All the services provide interfaces to users through API, making it highly scalable and pay-per-use. This architecture eliminates the need for traditional mass continuous online server components, lowers the complexity of development and maintenance, reduces operating costs, and shortens the delivery cycle of business systems, allowing users to focus on higher value-density business function development.

5. Cloud Native in Telecommunication

In terms of standards, ETSI NFV ISG released its Research Report on Enhanced Network Functions Virtualization (NFV) architecture for Cloud Native Containers and Platform as a Service (PaaS) in October 2019, followed by a series of technical specifications for container layer north-facing interfaces and management and network orchestration (MANO) management containers. It also plans to carry out standard studies, including Container Cluster Management Technical Specification (IFA036), Container Network Research Report (IFA038) and Container Security Specification (SEC023), and to expand existing NFV MANO interface functions to support life cycle management and organization of containerized VNF. On the open source side, CNCF established Telecom User Group (TUG) to import telecommunication industry needs into upstream open source project design, and completed the Cloud Native Thinking for Telecommunications white paper. The Linux Foundation has set up a CNTT working group to study cloud native/container technology application scenarios in the telecommunications industry, mainly based on open source Kubernetes to define cloud native network infrastructure, analyze gaps in telecommunication business needs, and provide a reference implementation and test validation framework for the corresponding infrastructure [9].

3.1.2 Virtualization

Virtualization is a technology that logically redistributes physical resources. It logically divides "large blocks of resources" into "small blocks of resources with independent functions", which can not only maximize the utilization of resources, but also realize resources isolation for different users on the basis of shared resources. Virtualization technology is the foundation of cloud computing and is used everywhere in the cloud.

In computer science, virtualization includes several levels:

- Virtualization based on hardware abstraction: Provides hardware abstraction layer, including the abstraction of hardware resources such as processors, memory, I/O devices, interrupts, etc.
- Virtualization based on operating system: Provides multiple isolated user-state instances, i.e. containers, which have separate file systems, networks, system settings, library functions, etc.
- Virtualization based on programming language, such as Java Virtual Machine (JVM), is

process-level virtualization.

Based on hypervisor type, virtualization can be divided into the following categories:

- Full software virtualization: There is no need to modify the guest operating system. All guest operating systems are simulated by software, but the performance consumption is high, ranging from 50% to 90%.
- Para-virtualization: The guest operating system invokes hypercall provided by Hypervisor by modifying the kernel and driver, with a performance consumption of 10%-50%.
- Full Hardware Virtualization: Hardware supports virtualization with performance consumption ranging from 0.1% to 1.5%.

In addition, virtualization technologies include Central Processing Unit (CPU) virtualization, memory virtualization, IO device virtualization, storage virtualization, network virtualization, container virtualization, network function virtualization, as well as 5G network slicing.

3.1.3 Heterogeneous Computing

Heterogeneous computing is a hybrid computing system composed of CPU, coprocessor, on-chip system (SoC), Graphics Processing Unit (GPU), Application Specific Integrated Circuits (ASIC), Field Programmable Gate Array (FPGA), and other computing units with different types of instruction sets and different architecture. Heterogeneous computing appears as CPU+, which has good feasibility and versatility, and can greatly improve system performance and power consumption efficiency.

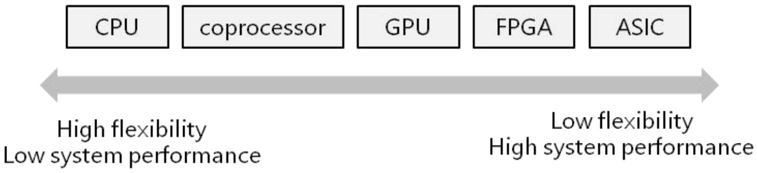

Figure 8: Characteristics of different heterogeneous hardware

Application heterogeneity acceleration requires an overall solution of hardware and software based on the acceleration platform to achieve better performance and cover more scenarios, including GPU-based, FPGA-as-a-service (FaaS) and Domain Specific Architecture (DSA)/ASIC-based acceleration solutions. For example, the GPU acceleration of NVIDIA is mainly achieved through the programming development framework of CUDA. FaaS relies on the hardware programmability provided by the FPGA, and requires users or third-party developers to complete the accelerated hardware and software mirroring development for specific application scenarios. The acceleration of DSA for specific application scenarios provides more flexibility on the basis of ASIC and is more efficient than GPU and FPGA [10].

3.2 Key Technologies of Network Function Layer

3.2.1 Service Definition

In the process of re-designing the RAN control plane, the inherent association of the radio resource management (RRM) modules should be considered, also the principles of micro-services, such as "high cohesion" and "loose coupling", should be considered. Services are defined according to external requirements. For RAN services, requirements may come from CN NFs, other RAN nodes, or UEs. The CP interface protocol between RAN and CN, more specifically, between the BS and AMF is Next Generation Application Protocol (NGAP). All the interactions between the BS and CN NFs need to go through this interface. Therefore, we can extract the services of RAN for CN requirements from the NGAP. Similarly, more RAN services can be defined based on the XnAP and Radio Resource Control (RRC) protocols.

3.2.2 Service-based Interface

In terms of functionalities, the design of service-based interface (SBI) not only needs to meet the functional interaction requirements between various microservices, but also needs a unified interface design.

In terms of performance, the current SBI is based on the Transmission Control Protocol (TCP), which overhead is intolerable to the user plane. Therefore, the current UP interface still uses the UDP (User Datagram Protocol) and the GTP-U (General Packet Radio Service Tunneling Protocol). In the future, it is necessary to consider the service interface design not only for the CP but also for the UP. We hope to explore the SBI that has less overhead and near real-time performance, and meet the requirements of both CP and UP.

3.2.3 Data Processing Order

Before processing a specific data stream, it is necessary to identify the related UP services or service instances combination as well as the order of data processing. There are two possible ways to determine the data processing order: The first way is that the service consumer dynamically selects the most suitable instance of the service provider, which is consistent with the existing 5G mechanism. The second way is to introduce a new service to determine the association of service instances related to the data processing. The association will be indicated to the relevant service instances by the new service for subsequent data processing.

3.2.4 Packet Format Definition

Different from the stylized packet headers in the traditional layered protocol, the service-based packet headers will be generated on demand. The content in the packet header is related to the services on the data processing chain, and may also be related to business requirements, measurement results, and so on. A possible implementation is to divide the data packet header into multiple parts, and a certain part is associated with only one or some service/service instance(s). Only the specific service/service instance can modify the associated part of the packet header.

3.2.5 Data Security

In the traditional layered protocol architecture, the headers and functions of a specific protocol are visible only to that layer. But the packet header information is visible to all “network function services” in service-based RAN, leading to the data security issues. On one hand, illegal network function services may obtain data and information, resulting in information leakage; On the other hand, approved “network function services” may illegally utilize data information, such as analyzing user privacy. Therefore, data access should be controlled so that only the authorized network function service is allowed to read/write the relevant packet and/or packet header. It may be necessary to introduce the security control function to ensure the secure interaction between services. The control function is responsible for issuing corresponding keys to the services, so that they can modify or read relevant data accordingly.

3.2.6 Data Collection Mechanism

In the traditional network, the collected data should be reported to the application server or centralized network function for analysis and processing. The process not only increases the data acquisition delay, but also brings signaling transmission overhead. Observability is one of the basic principles of micro-service design. Every service will natively support the data collection mechanism and flexibly adapt to a variety of data analysis frameworks. Therefore, in the SBA network, the data collection and analysis mode will undergo a fundamental change, which will occur at any elements of the E2E SBA network.

3.3 Key Technologies of Orchestration and Management Layer

With the advancement of service-based RAN and service-based CN, 6G will be an E2E service-based network. In the E2E service-based network, orchestration and management are a crucial part of ensuring a diverse service experience for users. Efficient orchestration and management can organically integrate services to ensure the user performance requirements, and the maximum utilization efficiency of various network resources (including wireless resources, computing resources, storage resources, deployment resources, etc.).

For orchestration and management, the Open Network Automation Platform (ONAP) has become a de facto standard in the fields of network automation, closed loops, and orchestrators. This can be used as a benchmark to further promote the orchestration and management framework and related technologies for 6G SBA networks.

It can be expected that in the future 6G networks will be deployed on a large scale across domains and suppliers in a wide geographical area. In order to achieve scalability, hierarchical cross-domain orchestration and management will be a feasible way. The "domains" can be divided according to the technical fields, such as RAN, TN, and CN. Also, it can be divided according to the deployment fields, such as edge data center domain, core data center domain, etc. In the hierarchical cross-domain orchestration and management framework, the service orchestrator and the service controller are the fundamental components.

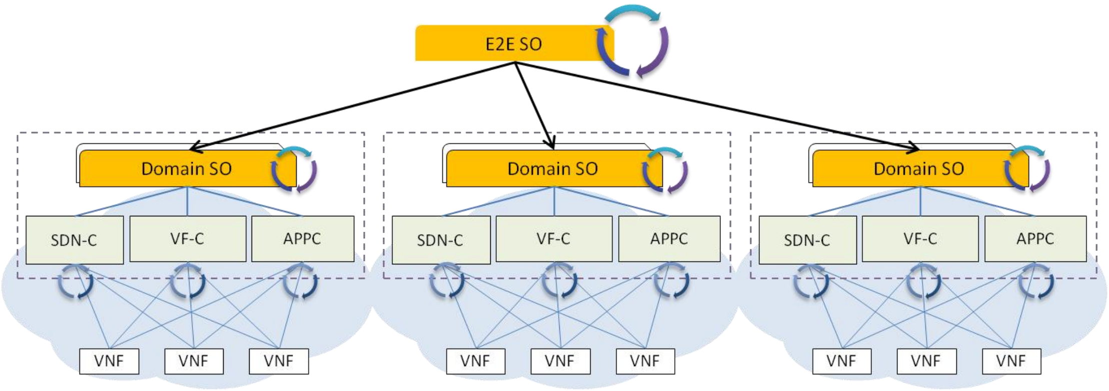

Figure 9: Hierarchical domain orchestration management framework

3.3.1 Service Orchestration

The Service Orchestration (SO) is responsible for the orchestration at the service level to realize the closed-loop automation of related processes, such as service instantiation, service updating, and globally optimized deployment. The E2E SO performs E2E service orchestration, connecting various SO domains with the southbound interface and connecting the operation and maintenance support system/business support system with the northbound interface.

3.3.2 Service Control

The service controller is responsible for controlling the closed-loop automation of resources, including the allocation, recycling, adjustment, expansion and contraction of local resources. Controllers in different domains is respectively responsible for closed-loop automation within the domain. Controllers may include application controllers, network controllers, virtual function controllers, infrastructure controllers, etc.

4. Challenges of service-based RAN

In the process of studying service-based RAN, the inherent limitations of traditional RAN must be considered, also the new problems caused by the introduction of new technologies must be solved. Therefore, the challenges are enormous.

4.1 Service definition

At present, there is no unified and feasible algorithm for service splitting in the industry. If the RAN functionalities are not decomposed properly, it is likely to develop a distributed monolithic application: a distributed system containing a large number of tightly coupled services that must be deployed together. This will combine the drawbacks of both monolithic architecture and microservice architecture. The BS has been developed as a monolithic application for decades, which makes the service decomposition more complicated. The difficulties can be observed from the tightly coupled RRM modules.

4.2 Network performance

The telecom equipment requires always-on operation without interruption, that is why BSs have always used dedicated hardware instead of general purpose processing platform (GPPP). The main advantage of GPPP is that it can accomplish more tasks, but at the cost of scattered computing power and reduced performance.

4.3 Test and operation

Different RAN services may be implemented by different vendors. Due to the different technical solutions of different manufacturers and their understanding of interface specifications may also be different, the interoperation testing will be full of challenges. At the same time, it may be difficult to distinguish the responsibilities of vendors at the stage of installation or maintenance, which would affect the recovery process. As the number of devices that need to be maintained increases, the complexity and cost of network management increase.

4.4 Energy efficiency

The service-based RAN will be implemented on the generic server. This kind of server applies general-purpose chips, which have the advantages of low cost and high flexibility and the disadvantages of low processing efficiency and high power consumption. According to industry research, the number of general-purpose chips used to implement 5G BS functions is 18 times that of dedicated chips, and the power consumption is about 30 times that of dedicated chips. Similarly, the high power consumption is also a tricky issue for service-based RAN and needs to be carefully considered.

4.5 Heterogeneous hardware

Service-based RAN seems to be a software reform, but in fact it depends heavily on the support of the underlying heterogeneous hardware platform. This involves the abstraction of hardware resources, virtualization, orchestration and management. The virtualization capabilities of heterogeneous hardware vary greatly due to different application scenarios. In addition, heterogeneous hardware solutions usually involve many hardware manufacturers, cloud manufacturers and network equipment manufacturers, and the technical challenges cannot be completely solved by any individual company, therefore, a close cooperation between the suppliers in different fields is required.

4.6 Network security

The accuracy and reliability of data directly affect the performance of a network or equipment. Therefore, it is important to ensure the security of data, especially the data exchanged between RAN services from different vendors. In addition, the risks of network attacks caused by network exposure, security risks introduced by cloudification and virtualization, and vulnerability from open source code also need to be considered simultaneously [11].

4.7 Initial cost

The cost advantage of service-based RAN is not significant at this initial stage. Vendors involved in BS development include software providers, hardware providers, radio remote unit providers and system integrators. Before the scale effect kicks in, the cost advantages are not obvious. But it can be expected that the cost advantage would be significant once scaled up.

4.8 The way of standardization

Currently, it may take two years for a software version to go from research, standardization to test and deployment. In addition, new versions need to be backward compatible with older versions, making the introduction of new technologies or features very complicated. The existing standardization organization structure and way of working cannot meet the rapid development and deployment requirements of service-based RAN, therefore adjustment is highly required.

5. Application Scenarios of Service-based RAN

5.1 On Demand Providing Necessary Services for Vertical Industries

The traditional BS has comprehensive functionalities, such as allowing the interoperability between 5G and 4G, supporting a variety of frequency bands, supporting Global Navigation Satellite System (GNSS) positioning, supporting IMS voice. This constant addition makes the BS omnipotent, but more complex and unstable. For industrial scenarios, the BS with lightweight, low cost and flexible deployment is more needed to adapt to the customized vertical requirements.

With service-based RAN, vertical industry partners are able to choose the services that they really want and deploy the services on different types of cloud platforms in different geographic locations as needed. For example, RAN services for ultra-low latency and ultra-high reliability applications may be deployed on edge cloud platforms to reduce access latency; RAN services for high-broadband applications may be deployed on convergent cloud platforms to reduce deployment costs. In addition, vertical industry partners can deploy RAN services with CN services on the same platform for better system performance.

5.2 Providing User-specific Service Capabilities for Individual Users

Traditional networks provide services for users based on the QoS requirements of the requested service. For the same service, the network delivers almost the same configuration to different users. However, in the future, different users may have different requirements even for the same service. The service-based RAN makes it possible. Based on user needs, customized network function service instances for UE can be generated and deployed on appropriate cloud platforms, for example, some UE-specific service instances related to AI model training can be deployed at the edge cloud platform to help UE timely complete computation-intensive tasks.

6. Preliminary Considerations on Service-based RAN Standardization

In the past decade, 5G CN has undergone network evolution from software-defined network to network function virtualization, and then to cloud native network. The evolution of network technology and the prosperity of open source communities have brought unprecedented changes to the network industry. It is the attempt of these technologies and the innovation of communities that make the new technologies can be fast applied in the networks. This successful historical experience is worth continuing and recalling.

The 3GPP standardization and industrialization are promoted in the granularity of Release, and the time period of each Release is about 15/18 months which is long, and accordingly, the development cycle of manufacturers and deployment cycle of operators are also long. Different from the traditional development and deployment mode in the CT field, service development in the IT field is usually carried out independently with granularity of Feature and adopts the "open source" cooperative development mode, therefore the development and deployment cycle of an individual service is relatively short, some of which can be implemented in two weeks.

Standardization and industrialization of service-based RAN can also be explored with reference to this mode, such as "standardization + open source". In the early stage of service-based RAN research, the standardization, development and deployment can be carried out only for basic services and interfaces. Later, features can be granularity, and necessary standardized design can be carried out for each "open source" service. Multiple features can be developed in parallel, thus promoting the overall process of standardization and industrialization.

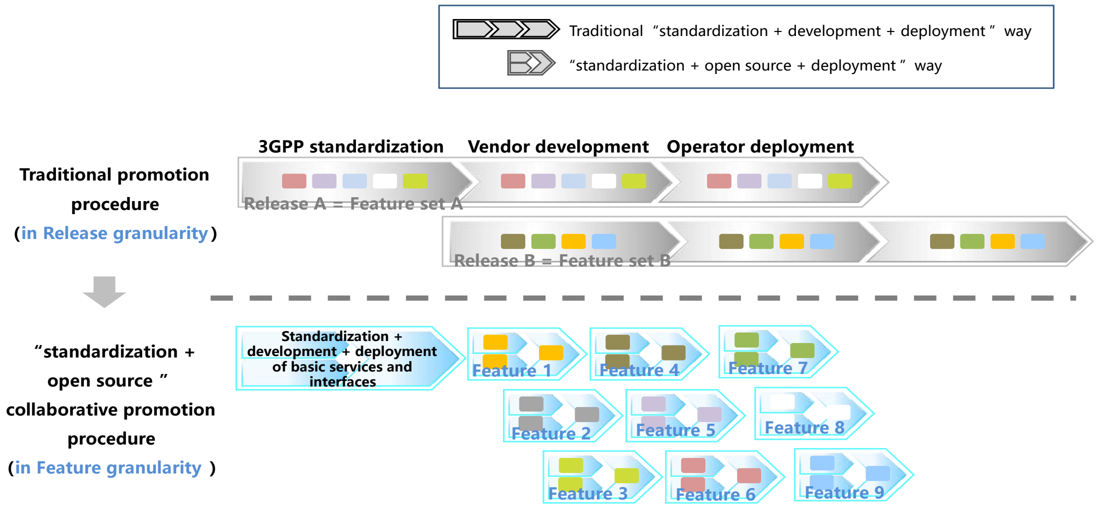
Traditional "standardization + development + deployment" way
"standardization + open source + deployment" way
3GPP standardization
Vendor development
Operator deployment
Traditional promotion procedure
(in Release granularity)
Release A = Feature set A
Release B = Feature set B
"standardization + open source" collaborative promotion procedure
(in Feature granularity)
Standardization + development + deployment of basic services and interfaces
Feature 1
Feature 4
Feature 7
Feature 2
Feature 5
Feature 8
Feature 3
Feature 6
Feature 9

Figure 10. Preliminary considerations on service-based RAN standardization

7. Development Prospect

With the rapid development of ICT convergence, the cloud and service-based RAN is a tendency. The "open mode" of "whitebox hardware + open source software" in mobile communication network has also been paid attention to by global operators and industry, and become a direction that cannot be ignored. At present, the head mobile infrastructure manufacturers have accumulated strong advantages in “integration mode”, but not in “open mode”. The two industrial modes are likely to coexist for a long time to come. Facing the unstable and uncertain global industry situation in the future, we hope to work together with industry partners to actively explore the direction of service-based RAN, promote the industry to mature, and jointly create the prosperity of the entire mobile communication industry.

References

[1] China Mobile, Vision and Requirements for 2030+(version 2.0), 2020.

[2] China Mobile, Network Architecture Prospect for 2030+, 2020.

[3] China Mobile, Technology Trends for 2030+, 2020.

[4] Rakuten Mobile, Inc. Rakuten Mobile Plans to Acquire Innoeye to Support Rakuten Communications Platform Launch [E]. http://global.rakuten.com/corp/news/press/2020/0513_02.html. 2020.

[5] C.Richardson. Microservices pattern[M]. Beijing: China Machine Press,2020.

[6] S.Newman. Building Microservices[M]. Beijing: Post&Telecom Press, 2021.

[7] Gartner. Top Strategic Technology Trends for 2022[E]. https://www.gartner.com/en/information-technology/insights/top-technology-trends. 2021.

[8] CNCF. Cloud Native Computing Foundation (CNCF) Cloud native definition [E]. https://www.cncf.io/about/who-we-are/. 2021.

[9] China Mobile, China Telecom, China Unicom, et al. Cloud Native Telecom Industry White Paper, 2020.

[10] Chaobo Huang. Software and Hardware Integration: A innovation road to Super-scale Cloud Computing architecture [M]. Beijing: Publishing House of Electronics Industry, 2021.

[11] Ericsson. Security considerations of Open RAN[E]. https://www.ericsson.com/en/security/security-considerations-of-open-ran

中国移动 China Mobile 研究院 CMRI

Digital Twin, Ubiquitous Intelligence